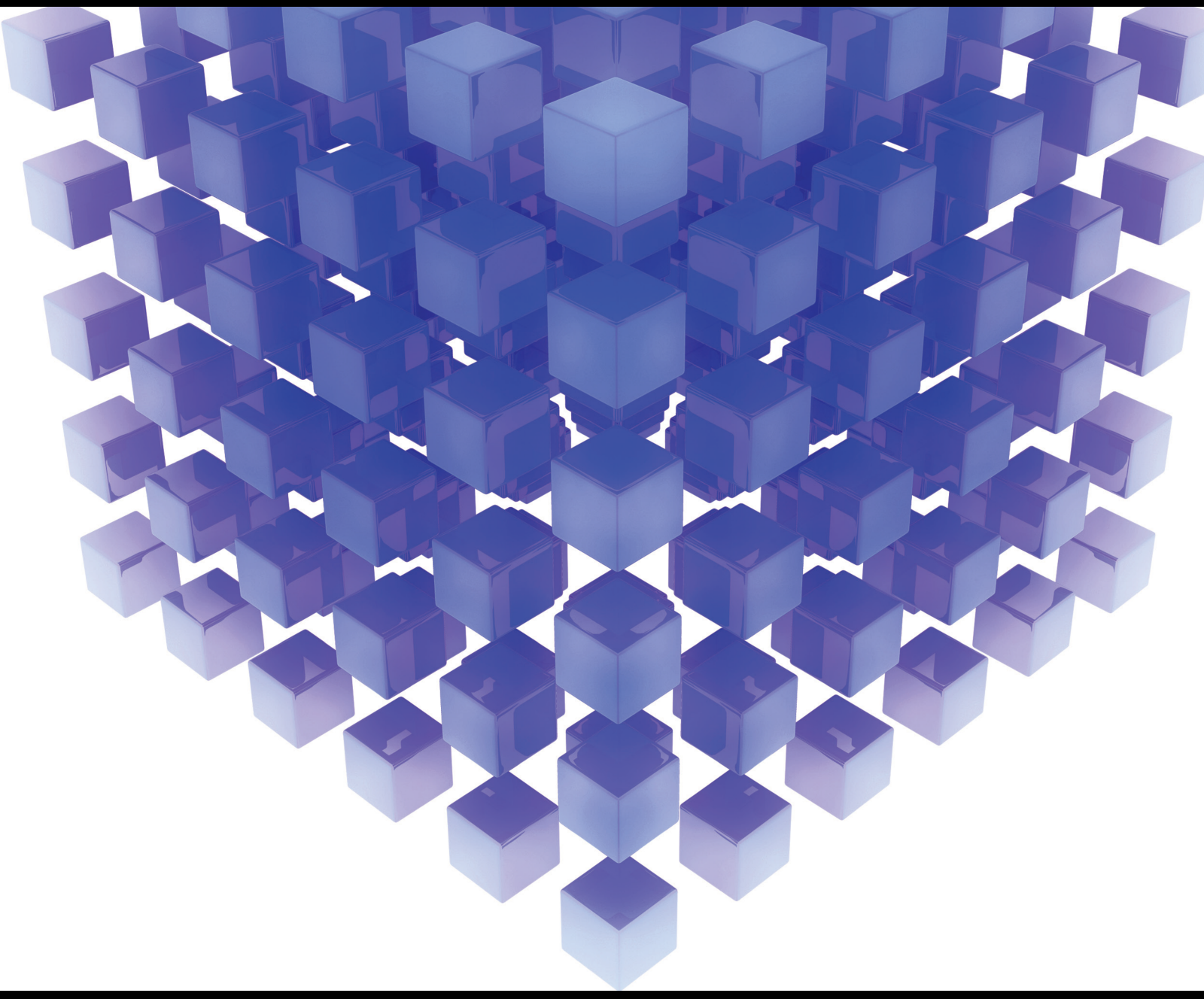
Mathematical Problems in Engineering

Safety Technologies and Fault Tolerant Methods for Engineering

Lead Guest Editor: Yong Chen
Guest Editors: Mahdi Tavakoli, Mohammed Abouheaf, and Esam Hafez Abdelhameed



Safety Technologies and Fault Tolerant Methods for Engineering

Mathematical Problems in Engineering

Safety Technologies and Fault Tolerant Methods for Engineering

Lead Guest Editor: Yong Chen
Guest Editors: Mahdi Tavakoli, Mohammed Abouheaf, and Esam Hafez Abdelhameed

Chief Editor

Academic Editors

Contents

A Classification Algorithm of Fault Modes-Integrated LSSVM and PSO with Parameters' Optimization of VMD
Yunqian Li, Darong Huang, and Zixia Qin
Research Article (12 pages), Article ID 6627367, Volume 2021 (2021)

A Parameter-Optimized Variational Mode Decomposition Investigation for Fault Feature Extraction of Rolling Element Bearings
Guoping An, Qingbin Tong, Yanan Zhang, Ruifang Liu, Weili Li, Junci Cao, Yuyi Lin, Qiang Wang, Ying Zhu, and Xiaowen Pu
Research Article (15 pages), Article ID 6629474, Volume 2021 (2021)

Adaptive Extraction Method Based on Time-Frequency Images for Fault Diagnosis in Rolling Bearings of Motor
Yunchao Ma, Chengdong Wang, Dongchen Yang, and Cheng Wang
Research Article (12 pages), Article ID 6687195, Volume 2021 (2021)

Shield Reliability Analysis-Based Transfer Impedance Optimization Model for Double Shielded Cable of Electric Vehicle
Xiaoshan Wu, Xiaohui Shi, Jin Jia, Heming Zhao, and Xu Li
Research Article (8 pages), Article ID 5373094, Volume 2021 (2021)

Adaptive Fuzzy Modified Fixed-Time Fault-Tolerant Control on SE(3) for Coupled Spacecraft
Yafei Mei, Ying Liao, Kejie Gong, and Da Luo
Research Article (21 pages), Article ID 6648578, Volume 2021 (2021)

Multiparty Homomorphic Machine Learning with Data Security and Model Preservation
Fengtian Kuang, Bo Mi, Yang Li, Yuan Weng, and Shijie Wu
Research Article (11 pages), Article ID 6615839, Volume 2021 (2021)

Equipment Operational Reliability Evaluation Method Based on RVM and PCA-Fused Features
Linbo Zhu, Dong Chen, and Pengfei Feng
Research Article (9 pages), Article ID 6687248, Volume 2021 (2021)

Interval Number-Based Safety Reasoning Method for Verification of Decentralized Power Systems in High-Speed Trains
Peng Wu, Ning Xiong, Jiqiang Liu, Liujia Huang, Zhuoya Ju, Yannan Ji, and Jinzhao Wu
Research Article (12 pages), Article ID 6624528, Volume 2021 (2021)

Formal Verification on the Safety of Internet of Vehicles Based on TPN and *Z*
Yang Liu, Liyuan Huang, and Jingwei Chen
Research Article (11 pages), Article ID 6618168, Volume 2020 (2020)

A Fault Diagnosis Method of Rolling Bearing Integrated with Cooperative Energy Feature Extraction and Improved Least-Squares Support Vector Machine
Zhang Xu, Darong Huang, Tang Min, and Yunhui Ou
Research Article (13 pages), Article ID 6643167, Volume 2020 (2020)

An Integrated Health Condition Detection Method for Rotating Machinery Using Refined Composite Multivariate Multiscale Amplitude-Aware Permutation Entropy
Fuming Zhou, Wuqiang Liu, Ke Feng, Jinxing Shen, and Peiping Gong
Research Article (23 pages), Article ID 5303658, Volume 2020 (2020)

The Extraction Method of Gearbox Compound Fault Features Based on EEMD and Cloud Model
Ling Zhao, Jiaxing Gong, and Hu Chong
Research Article (8 pages), Article ID 6661975, Volume 2020 (2020)

A Short-Term Traffic Flow Reliability Prediction Method considering Traffic Safety
Shaoqian Li, Zhenyuan Zhang, Yang Liu, and Zixia Qin
Research Article (9 pages), Article ID 6682216, Volume 2020 (2020)

Balancing Access Control and Privacy for Data Deduplication via Functional Encryption
Bo Mi, Ping Long, Yang Liu, and Fengtian Kuang
Research Article (11 pages), Article ID 6662662, Volume 2020 (2020)

Travel Time Reliability-Based Signal Timing Optimization for Urban Road Traffic Network Control
Zhengfeng Ma, Darong Huang, Changguang Li, and Jianhua Guo
Research Article (11 pages), Article ID 8898062, Volume 2020 (2020)

Analysis of Vibration and Noise for the Powertrain System of Electric Vehicles under Speed-Varying Operating Conditions
Chenghao Deng, Qingpeng Deng, Weiguo Liu, Cheng Yu, Jianjun Hu, and Xiaofeng Li
Research Article (9 pages), Article ID 6617291, Volume 2020 (2020)

A Novel Median-Point Mode Decomposition Algorithm for Motor Rolling Bearing Fault Recognition
Ganzhou Yao, Bishuang Fan, Wen Wang, and Haihang Ma
Research Article (10 pages), Article ID 9406479, Volume 2020 (2020)

Stability Coordinated Control of Distributed Drive Electric Vehicle Based on Condition Switching
Zhao Jingbo, Chen Jie, and Liu Chengye
Research Article (10 pages), Article ID 5648058, Volume 2020 (2020)

Rotor Temperature Safety Prediction Method of PMSM for Electric Vehicle on Real-Time Energy Equivalence
Anjian Zhou, Changhong Du, Zhiyuan Peng, Qianlei Peng, and Datong Qin
Research Article (10 pages), Article ID 3213052, Volume 2020 (2020)

Fault Detection of the Wind Turbine Variable Pitch System Based on Large Margin Distribution Machine Optimized by the State Transition Algorithm
Mingzhu Tang, Jiahao Hu, Zijie Kuang, Huawei Wu, Qi Zhao, and Shuhao Peng
Research Article (9 pages), Article ID 9718345, Volume 2020 (2020)

Contents

Traffic Flow Anomaly Detection Based on Robust Ridge Regression with Particle Swarm Optimization Algorithm
Mingzhu Tang, Xiangwan Fu, Huawei Wu, Qi Huang, and Qi Zhao
Research Article (10 pages), Article ID 3673085, Volume 2020 (2020)

An Integrated Method for Fire Risk Assessment in Residential Buildings
Hongfu Mi, Yaling Liu, Wenhe Wang, and Guoqing Xiao
Research Article (14 pages), Article ID 9392467, Volume 2020 (2020)

Application of Model-Based Deep Learning Algorithm in Fault Diagnosis of Coal Mills
Yifan Jian, Xianguo Qing, Yang Zhao, Liang He, and Xiao Qi
Research Article (14 pages), Article ID 3753274, Volume 2020 (2020)

Fault Detection of Wind Turbine Pitch System Based on Multiclass Optimal Margin Distribution Machine
Mingzhu Tang, Zijie Kuang, Qi Zhao, Huawei Wu, and Xu Yang
Research Article (10 pages), Article ID 2091382, Volume 2020 (2020)

Composite Compensation Control of Robotic System Subject to External Disturbance and Various Actuator Faults
Hao Sheng and Xia Liu
Research Article (11 pages), Article ID 1247079, Volume 2020 (2020)

Adaptive Cruise Control Strategy Design with Optimized Active Braking Control Algorithm
Wenguang Wu, Debiao Zou, Jian Ou, and Lin Hu
Research Article (10 pages), Article ID 8382734, Volume 2020 (2020)

Probability of Roadside Accidents for Curved Sections on Highways
Guozhu Cheng, Rui Cheng, Yulong Pei, and Liang Xu
Research Article (18 pages), Article ID 9656434, Volume 2020 (2020)

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6627367, 12 pages
https://doi.org/10.1155/2021/6627367

Research Article

A Classification Algorithm of Fault Modes-Integrated LSSVM and PSO with Parameters' Optimization of VMD

Yunqian Li,[1] **Darong Huang**,[1] **and Zixia Qin**[1,2]

[1]*Institute of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[2]*Institute of Design and Art, Beijing Institute of Technology, Beijing 100086, China*

Correspondence should be addressed to Darong Huang; drhuang@cqjtu.edu.cn

Received 1 November 2020; Revised 26 November 2020; Accepted 18 February 2021; Published 28 February 2021

Academic Editor: Emilio Insfran

To overcome the shortcomings that the early fault characteristics of rolling bearing are not easy to be extracted and the identification accuracy is not high enough, a novel collaborative diagnosis method is presented combined with VMD and LSSVM for incipient faults of rolling bearing. First, the basic concept of VMD was introduced in detail, and then, the adaptive selection principle of parameter K in VMD was constructed by instantaneous frequency mean. Furthermore, we used Lagrangian polynomial and Euclidean norm to verify the value of K accurately. Secondly, we proposed a classification algorithm based on PSO-optimized LSSVM. Meanwhile, the flowchart of the classification algorithm of fault modes may be also designed. Third, the experiment shows that the presented algorithm in this paper is effective by using the existing failure data provided by the laboratory of Guangdong Petrochemical Research Institute. Finally, some conclusions and application prospects were discussed.

1. Introduction

In recent years, the machinery has become more high-speed, intelligentized, and complicated with the development of the modern industrialization. As we all know, the rotating machinery is the cornerstone of transportation, power electronics, and manufacturing. So, how to guarantee the security of whole rotating machinery systems is very important in the industrial field. In the actual industrial scenario, the engineers and researchers have noticed that the safety of the bearings is often one of the critical joints, which ensures the global safety of whole rotating machinery [1]. Therefore, it is essential to detect and assess the performance of the running state of the bearings. The traditional fault diagnosis methods that judge and evaluate the running state of the bearing are operated or implemented by observing the frequency of the vibration signal. The skeleton of these methods consists of just three steps: signal processing, feature extraction, and fault pattern recognition. In most realistic scenarios, signal processing is often used as the preparing work for the feature extraction. Of course, the feature extraction is also used as the prepared work for the fault pattern recognition because the classification accuracy of the fault modes is the final objective in fault diagnosis of the bearing, so the signal processing and feature extraction are often integrated to analyze the vibration signals of the bearing. And, how to implement them becomes critical.

For the question raised above, the scholars have presented and constructed some models such as Empirical Mode Decomposition (EMD), Wavelet Transform (WT), Local Mean Value Decomposition (LMVD), and Variational Mode Decomposition (VMD) in references [2–4]. The experiment results show that these methods may acquire the most of the valuable information in the specified scenarios. Unfortunately, almost all these methods have some shortcomings. For example, the EMD and LMD have the phenomenon such as modal aliasing and endpoint effect. The WT needs to select the wavelet base and decomposition scale because the finite length may cause inaccurate decomposition of complete components. In the VMD, if the parameter K is wrongly selected, the phenomenon such as overdecomposition or underdecomposition will appear. To overcome these shortcomings, some improved algorithms have been presented such as Simplistic Geometric Mode

Decomposition (SGMD), Adaptive Chirped Mode Decomposition (ACMD), and New Spectral Analysis Methods (NSAM) in [5–7]. Especially, to address the shortcoming of the EMD, the study in [8] has constructed the EMD envelope correction method using B-spline interpolation and base spline. These methods may alleviate the modal aliasing problem of high-frequency signals. Meanwhile, to optimally select the parameter K of VMD, the genetic variation sample group, kurtosis criterion variational mode decomposition, and self-organizing mapping (SOM) neural network have been adopted to adaptively determine the optimal value of the parameter K in [9–12]. To verify the effectiveness of these new methods, some experiment examples have been used to simulate in [13–17]. The simulated results showed that these improved models may solve the shortcomings to a certain extent. For practical application, the constant improvement of the existing methods is the goal of the engineers and scholars. Thus, we will treat the problem in this paper.

On the contrary, in the view of fault diagnosis, to get the accurate classification of fault modes is the other main objective of the bearing fault diagnosis. In fact, an excellent pattern recognition method of the fault modes has an important influence for the final diagnosis accuracy. Based on this objective, support vector machine (SVM), least-squares' support vector machine (LSSVM), BP neural network (BPNN), fuzzy logic (FM), and other methods have been successfully applied in [18–23]. And, then, these fault pattern recognition methods have been widely used in different industrial environments. Further, some improved methods of the fault pattern recognition were studied in [24, 25]. For example, the double support-vector machine and smooth iterative online-support tensor algorithm are proposed to improve the performance of the traditional support vector machine in [26, 27]. The least-squares' ground projection method of the double support-vector machine reduces the diagnostic error in [28]. Meanwhile, to optimize the penalty factor C and kernel parameter of LSSVM, some new algorithms such as the Moth-flame Optimization (MFO), the von Neumann Topology Whale Optimization Algorithm (VNWOA), Quantum Particle Swarm (QPS), and Chaotic Antlion Algorithm (CAA) were introduced to implement the optimization operation for enhancing the precision of fault diagnosis in [29–34]. The experiments have verified the performance of these presented algorithms. However, the global searching ability of these algorithms is weak in the real industrial environment. So, searching the improved pattern recognition method to enhance the global searching ability and improve the classification accuracy of fault modes is another concern in our paper.

Based on the two points mentioned above, an improved fault diagnosis of the bearing will be presented combined with the VMD algorithm based on instantaneous frequency optimization and particle swarm optimization least-squares' support vector machine in our paper. The rest of this paper is arranged as follows. In Section 2, the adaptive selection principle of K value in the VMD algorithm is given in detail. In Section 3, the least-squares' support vector machine classification model for particle swarm optimization (PSO) is established, and the concrete flowchart of the fault diagnosis process is designed and analyzed. In Section 4, some simulated examples were used to verify the effectiveness of our algorithm through the existing failure data provided by the laboratory of Guangdong Petrochemical Research Institute. Finally, some conclusions are summarized in Section 5.

2. Adaptive Selection Principle of Parameter K in the VMD Algorithm

2.1. The Basic Concept of the VMD Decomposition Principle. The intrinsic mode function (IMF) is defined as an FM and AM signal by VMD decomposition and is expressed as follows:

$$u_k(t) = A_k(t)\cos\left[\phi_k(t)\right], \quad k = 1, 2, \ldots, K, \tag{1}$$

where $A_k(t)$ expresses the instantaneous amplitude, $w_k = \phi_k'(t)$ is the instantaneous frequency, and K represents the number of signal components after decomposition.

Suppose the original signal f is a multicomponent signal, which is composed of the K IMF component with limited bandwidth, and the central frequency of each IMF is w_k. To determine the bandwidth of each mode, the following steps are used to obtain it:

(1) Analytic signals of modal functions are obtained, and Hilbert transformation is performed for each modal function $u_k(t)$:

$$\left[\sigma(t) + \frac{j}{\pi t}\right] u_k(t). \tag{2}$$

(2) Mix the estimated center frequency $e^{-jw_k t}$ of each modal analytic signal. The spectrum of each modal is modulated to the corresponding baseband as follows:

$$\left[\left[\sigma(t) + \frac{j}{\pi t}\right] \times u_k(t)\right] e^{-jw_k t}. \tag{3}$$

(3) Calculate the square L^2 norm of the gradient of the above demodulation signal, and estimate the bandwidth of each modal component. The constraint variational model is established as follows:

$$\begin{aligned} &\min_{\{u_k\},\{w_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\sigma(t) + \frac{j}{\pi t} \right) \times u_k(t) \right] e^{-iw_k t} \right\|_2^2 \right\} \\ &\text{s.t} \sum_k u_k(t) = f, \end{aligned} \tag{4}$$

where $u_k = \{u_1, u_2, \ldots, u_k\}$ represents the K IMF components obtained by decomposition and $w_k = \{w_1, w_2, \ldots, w_k\}$ represents the center frequency of each component.

In order to solve the above constraint variational model, the quadratic penalty factor α and Lagrangian multiplication operator $\lambda(t)$ are introduced, where the quadratic penalty factor can guarantee the reconstruction accuracy of the

signal in the presence of Gaussian noise and $\lambda(t)$ keeps the constraint conditions strict. The expanded Lagrangian expression is as follows:

$$L(\{u_k(t)\},\{\omega_k(t)\},\lambda) = \alpha\sum_k \left\|\partial_t\left[\left(\sigma(t)+\frac{j}{\pi t}\right)\times u_k(t)\right]e^{-j\omega_k t}\right\|_2^2 + \left\|f(t)-\sum_k u_k(t)\right\|_2^2 + \langle\lambda(t), f(t)-\sum_k u_k(t)\rangle. \tag{5}$$

The multipliers' alternating direction algorithm is used to update the IMF and its center frequency, and the saddle point of formula (4) is the optimal solution of the original problem. All IMF in the frequency domain can be obtained by the following formula:

$$\hat{u}_k^{n+1}(w) = \frac{\hat{f}(w)-\sum_{i\neq k}\hat{u}_i(w)+\hat{\lambda}(w)/2}{1+2\alpha(w-w_k)^2}, \tag{6}$$

where $\hat{u}_k^{n+1}(w)$ is the current residual quantity and $\hat{f}(w) - \sum i\neq k\hat{u}_i(w)$ is the result of Wiener filtering. The new IMF power-spectrum centers in the algorithm are as follows:

$$\hat{w}_k^{n+1} = \frac{\int_0^\infty w\left|\hat{u}_k(w)\right|^2 dw}{\int_0^\infty \left|\hat{u}_k(w)\right|^2 dw}, \tag{7}$$

where w_k^{n+1} is the power spectrum center.

The above process is the adaptive decomposition process of VMD. From the decomposition principle, it can be known that VMD can well avoid the endpoint effect and modal confusion. But, from the perspective of the actual decomposition process, the VMD algorithm loses the ability to decompose signals independently, which needs to preset the value of K. And, the reasonableness of the K value determines the signal decomposition accuracy of VMD. If the K value is estimated according to the existing observation method, that is, observing the center frequency differentiation of the signal component, the better the center frequency differentiation is, the better the selection of the K value is, and there is no overdecomposition and underdecomposition. However, there is a large error in this method, which makes it difficult to guarantee the decomposition accuracy of the signal and also affects the final classification accuracy. Therefore, this paper proposes a method to optimize the K value of VMD by using instantaneous frequency, which can make use of the difference of instantaneous frequency between signal components to measure the advantage of the K value.

2.2. K Value Estimation of the VMD Component Based on Instantaneous Frequency. If the K value is set too high, the decomposition number will be too large, and then, the component will be fragmenting, especially at high frequency, and the average instantaneous frequency will decrease. If the K value is set too low, the signal will not be completely decomposed, and the superiority of the signal component cannot be reflected. Therefore, the original signal may be decomposed by VMD once the K value was traversed from 2 to 10. And then, the mean values of instantaneous frequencies may be calculated under different K values, and the line graph may be also drawn. Lagrange polynomials were used to fit the discrete points, and the polynomial coefficients under different K values were extracted to construct the coefficient vector, and then, the Euclidean norm of the vector was calculated. The smaller the norm was, the smoother the fitting instantaneous frequency curve was and the better the value was.

The definition of instantaneous frequency is as follows:

$$f_i(t) = \frac{1}{2\pi}\frac{d\varphi(t)}{dt}, \tag{8}$$

where $\varphi(t)$ is a single-valued function of time t, that is, a single-component signal on frequency. The analytic signal of instantaneous frequency is defined as follows:

$$z(t) = x(t) + j\hat{x}(t) = a(t)e^{j\varphi(t)}, \tag{9}$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$, $z(t)$ is the analytic signal of $x(t)$, $a(t)$ is the module of the signal $a(t) = \sqrt{x^2(t)+\hat{x}^2(t)}$, and $\varphi(t)$ is the phase of the signal, which is expressed as $\varphi(t) = \arctan(\hat{x}(t)/x(t))$. The instantaneous frequency multiplying the integral of the density function over the entire time axis is the average frequency of the signal. Through the Fourier transform of the analytic signal $z(t)$ in formula (9), we can get the following formula:

$$Z(f) = \int_{-\infty}^{\infty} a(t)^{j[\varphi(t)-2\pi f_i t]}dt. \tag{10}$$

According to the principle of the stationary phase, the integral of equation (10) has a maximum value at the frequency f_i, which needs to meet the condition $d/dt[\varphi(t) - 2\pi f_i(t)] = 0$, namely, $f_i(t) = (1/2\pi)(d\varphi(t)/dt)$. This conclusion indicates that the energy of nonstationary signals is mainly concentrated at the instantaneous frequency. This conclusion indicates that the instantaneous frequency plays a very important role in the recognition, detection, estimation, and modeling of signals, and it can also be used as the evaluation index of VMD decomposition signals. Therefore, on the basis of the original VMD, the original signal is decomposed into different signal components, and the decomposed number is the K value. Then, the average instantaneous frequency under different mode numbers of K from 2 to 10 is calculated to judge the trend of the corresponding line graph. The flatter the trend is, the better the corresponding K value is, so as to realize the optimization of the parameters of the VMD algorithm. However, this method may be misjudged to some extent. In order to measure the instantaneous frequency change more accurately and select the optimal K value, corresponding methods should be adopted to obtain numerical results.

2.3. The Superiority Distinction of K Values Based on Lagrange Polynomials. After calculating the average instantaneous frequency of each component, we need to adopt an index to measure the variation trend of the instantaneous average frequency, which can avoid the error caused by subjective judgment. By fitting the mean instantaneous frequency, the Lagrangian polynomials can be calculated, the

vector norm of their coefficients can be compared, and the merits and disadvantages of the K value can be evaluated.

For any point $x_k (k = 0, 1, \ldots, n)$ in the interpolation node $x_0, x_1, \ldots, x_n$, make a polynomial $l_k(x)$ of degree n, which satisfies the following formula:

$$l_k(x) = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases} \tag{11}$$

The basic function of Lagrange interpolation is as $l_k(x)$, and the node is presented as $x_i (i = 0, 1, \ldots, k-1, k, k+1, \ldots, n)$. So, $l_k(x)$ is a polynomial with n null points. Therefore,

$$l_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{(x - x_i)}{(x_k - x_i)} = \frac{(x - x_0) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)}{(x_k - x_0) \ldots (x_k - x_{k-1})(x_k - x_{k+1}) \ldots (x_k - x_n)}, \tag{12}$$

where $l_k(x) (k = 0, 1, \ldots, n)$ is the n-order basic interpolation polynomial or n-order Lagrangian interpolation basis function on $n+1$ interpolation nodes. Using the n-order basic interpolation polynomial, the n-order Lagrange polynomial satisfying the interpolation condition $P_n(x_i) = f(x_i) = y_i (i = 0, 1, 2, \ldots, n)$ can be written as follows:

$$\begin{aligned} p_n(x_i) = L_n(x) = \sum_{k=0}^{n} y_k l_k(x) = y_0 l_0(x) \\ + y_1 l_1(x) + \cdots + y_n l_n(x). \end{aligned} \tag{13}$$

The average instantaneous frequency of different components is taken as the discrete point of calculating Lagrangian polynomials. After obtaining the simplest form of the Lagrangian polynomial by calculation, the coefficients of the polynomial are extracted and constructed into a vector, and the Euclidean distance of the vector with different K values is calculated. For the coefficient vector $v = (v_1, v_2, \ldots, v_3)$, the Euclidean distance of the vector is $\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2}$.

3. Classification Algorithm-Based Least-Squares' Support Vector Machine with Particle Swarm Optimization

3.1. Basic Concept of LSSVM. The LSSVM is an improved algorithm of the support vector machine; however, as a binary classifier, its core idea remains unchanged, that is, to find a hyperplane that optimizes classification and maximizes the gap between classifications, so as to improve the credibility of classification. The difference between LSSVM and SVM is that the construction of the objective function of LSSVM is through the binomials for the error factor, and at the same time, constraints are equally constraint, and in terms of solving the optimization problem, because the LSSVM is the constraint equation form, the solution is the system of linear equations; to a certain extent, it reduced the difficulty of the algorithm and raised the solving speed, and these advantages make it different from other improvement on the SVM algorithm. The basic principle of this method is described as follows.

The sample of training data can be expressed as $\{x_i, y_i\}_{i=1}^{l}$, $x_i \in R^n$ is the input vector of the ith sample, $y_i \in R$ is the target value of the ith sample, and l is the number of training samples. In special space, the LSSVM model can be expressed as

$$y(x) = w^T \varphi(x) + b, \tag{14}$$

where $\varphi(x)$ is the mapping function of nonlinear transformation, which maps the input sample data to the high-dimensional feature space. W is the weight vector, and B is the offset. The objective function of least-squares' support vector machines is described as

$$\min J\left(w^T \xi\right) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{l} \xi^2, \quad i = 1, 2, \ldots, l. \tag{15}$$

Type ξ is the error variable, and $\Upsilon > 0$ is the penalty coefficient. For the simplicity of analyzing, the Lagrangian function is designed as follows:

$$L(w, B, \xi, a) = J(w, \xi) - \sum_{i=1}^{l} a_i \left[w^T \varphi(x_i) + b + \xi - y_i \right], \tag{16}$$

where a_i is the Lagrange multiplier. In the real operation, the KKT optimal condition is used to calculate $\partial L/\partial w = 0$, $\partial L/\partial b = 0$, $\partial L/\partial \xi = 0$, and $\partial L/\partial a_i = 0$. So, the following system of linear equations should be obtained:

$$\begin{bmatrix} 0 & q^T \\ q & pp^T + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ l \end{bmatrix}, \tag{17}$$

$$\begin{aligned} p &= \left[\varphi(x_i)^T y_i, \varphi(x_2)^T y_2, ..., \varphi(x_l)^T y_l \right], \\ l &= [1, 1, \ldots, 1]^T, \\ q &= [y_1, y_2, \ldots, y_l]^T, \\ a &= [a_1, a_2, \ldots, a_l]^T. \end{aligned} \tag{18}$$

In equation (17), I is the identity matrix. According to the Mercer condition, the kernel function can be written as

$$k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j). \tag{19}$$

After a and b can be obtained from equations (18) and (19), the nonlinear function of LSSVM can be obtained as follows:

$$y(x) = \sum_{i=1}^{l} a_i k(x_i, x_j) + b. \tag{20}$$

3.2. PSO Parameter Optimization for LSSVM. LSSVM requires two parameters to be tuned: gam and sig2, where gam is the regularization parameter, which determines the minimization and smoothness of the adaptation error, and sig2 is the parameter of the RBF function. PSO optimizes two parameters of LSSVM, gam and sig2, to find the optimal combination of parameters, so as to improve the classification accuracy. The general optimization steps are as follows:

(1) Initializing the various parameters of the PSO algorithm, such as population size, learning factor, the maximum number of iterations, initial position, and the velocity of particles.

(2) Respectively, in the LSSVM predictive learning sample of each particle vector, get the prediction error of the current position value of the particle, which is used as the fitness value for each particle. Then, the current fitness value of each particle is compared with the best fitness value of the particle itself. If there are many, the current position of the particle is taken as the best position of the particle.

(3) The adaptive value of the optimal position of each particle was compared with the adaptive value of the optimal position of the population. If it is better, the optimal position of the particle is regarded as the optimal position of the population.

(4) Use formulas (21) and (22) to update the particle velocity and position:

$$\begin{aligned} v = w \times v + C_1 \times \text{Rand} \times (p_{\text{best}} - x) \\ + C_2 \times \text{Rand} \times (g_{\text{best}} - x), \end{aligned} \tag{21}$$

$$x = x + v, \tag{22}$$

where V is the particle speed, X is the position of the current particle, Rand () is a random number between (0, 1), and $C1$ and $C2$ are the learning factors, usually $C1{=}C2 = 1.5$.

(5) Check the result of optimization (maximum number of iterations or expected accuracy) is met or not. If so, the optimization is completed and the optimal solution is found. Otherwise, go to Step (2) and continue the search.

3.3. Rolling Bearing Fault Diagnosis Steps. The acceleration sensor is used to collect four state signals of the rolling bearing, which are normal, bearing external crack, bearing internal crack, and bearing wear. 10 groups of data of each state signal are collected. Take the normal state as an example, they were normal 1, normal 2, . . ., normal 10. Set the signal period to 1024, which means that, in a file such as "normal 1" with 10240 pieces of data, it is divided into 1024∗10 groups. Based on the above analysis, in order to ensure the realization of fault diagnosis and classification, the classification algorithm flow chart can be designed as follows:

(1) Traverse the K value of VMD (K value is the number of original signals decomposed into different components), input the first group of data of each state of the original signal collected into the VMD algorithm, and get K components under different K values ($K = 2, 3, ..., 10$).

(2) Calculate the average instantaneous frequency corresponding to different components of K, draw a line chart, and estimate the K value through the trend of the line chart.

(3) In order to further verify the pros and cons of K, the different components of the average instantaneous frequency may be used as computing Lagrange polynomial of discrete points, and then the most simplified forms of Lagrange polynomial can be computed. So, the extraction of polynomial coefficients may be constructed as a vector. In fact, the vector may be demonstrated and calculated under different K values, coefficient of Euclidean distance, and judge norm size to determine the optimal values of K.

(4) Set the optimal K value as the mode number that VMD needs to decompose. Decompose the 10 groups of data of each state and extract the time-domain features to form the feature set.

(5) The parameters of gamand sig2 of the LSSVM algorithm were optimized by the PSO algorithm.

(6) The obtained data set is input into the LSSVM classification algorithm, which is divided into training data and test data. The parameters of the model are updated with the training data, and the test data is input into the trained model to obtain the diagnosis results of fault pattern recognition;

The corresponding flowchart is shown in Figure 1.

4. Classification Experiment

To test the validity and rationality of the algorithm, some test data of the bearing provided by Guangdong Key Laboratory of Petrochemical Equipment Fault Diagnosis was used to

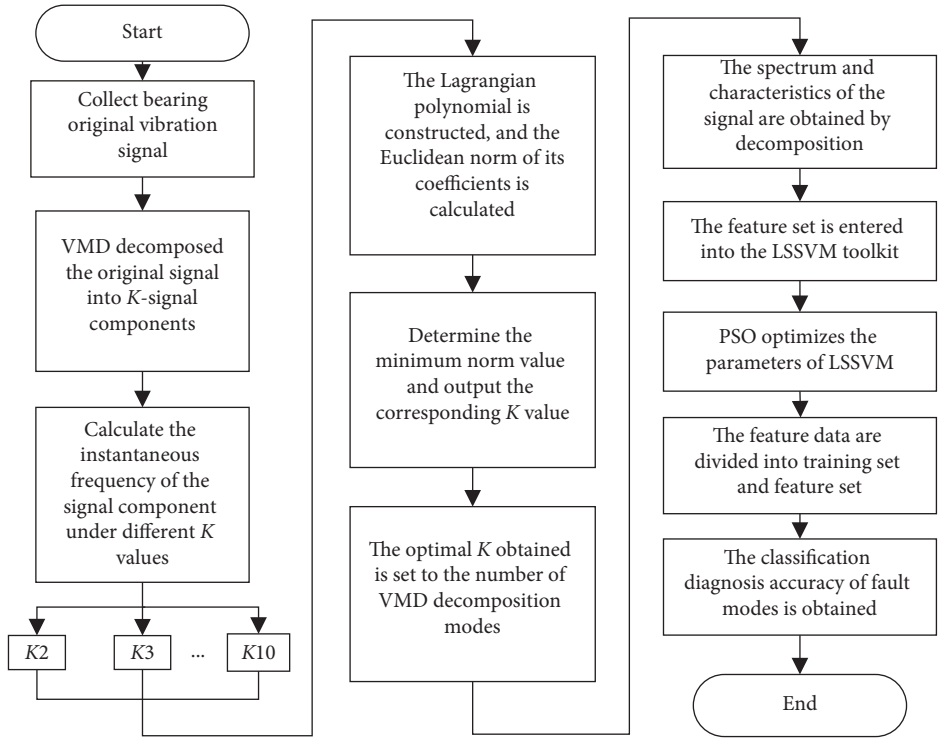


Figure 1: Fault diagnosis flow chart based on optimized VMD.

simulate and experiment. This data set included the acceleration changes with four different states: normal, bearing internal crack, bearing external crack, and bearing wear. The bearing damage and data acquisition platform are shown in Figure 2.

Figures 2(a)–2(c), respectively, represent bearing internal crack, bearing external crack, and bearing fault data acquisition platform. The acceleration sensor was used for data collection, with the collection period $T = 1024$. The collected fault data was divided into 10 groups according to the period, and the four different bearing states were divided into 40 groups.

Since these data are the most original vibration signal data, it is difficult to extract subsequent features without processing, so VMD is used to preprocess vibration signals. In order to select the optimal decomposed mode number K, first select 1 group from the 10 groups of data of each bearing state to input VMD, traverse K values from 2 to 10, calculate the instantaneous frequency mean, and get the corresponding broken line chart, as shown in Figure 3.

From Figure 3, a rough estimate of K may be obtained. Noticing that four kinds of condition is the most gentle the most ideal when $K = 2$, there is no high frequency under the intermittent and suddenly curved because the original signal only is decomposed into two components. Because the result do not conform to the actual, $K = 2$ is not as the objects of choice. Thus, it can be estimated that the optimal value K in the normal state is 3, the optimal value K in the bearing wear state is 6, the optimal value K in the bearing internal crack state is 4, and the optimal value K in the bearing external crack state is 5. However, such estimation may lead to wrong choices when the difference between the broken lines is not large. Therefore, it is necessary to choose an index to accurately judge the advantages and disadvantages of the K value. In this paper, Lagrange polynomials are proposed to be established, and instantaneous frequencies under each K value are used as discrete points to calculate Lagrange polynomials. After obtaining the simplest polynomials, coefficients are extracted and corresponding coefficient vectors are calculated. The smaller the Euclidean norm is, the better the K value is. The normal state data are selected here for experimental verification.

Table 1 shows the average instantaneous frequency corresponding to different K values in the normal state of the bearing, and Table 2 shows the corresponding Euclidian norm. From Table 2, it can be seen that the norm is the smallest when $K = 2$, but because it is not consistent with the actual situation, the value of K is excluded as 2. Therefore, it can be known that when the K value is 3, the corresponding norm is the smallest, and the optimal modal component number in the normal state of the bearing is 3, which is also consistent with the estimated result of the line graph of the K value above.

Setting the optimal $K = 3$ as the number of modes that VMD needed to decompose, 10 groups of data in the normal state were decomposed in a cycle to obtain the spectrum diagram and time-domain feature set of the modal components after VMD decomposition.

Figure 4 represents the spectrum diagram of VMD decomposition when $K = 3$ under the normal state. From the

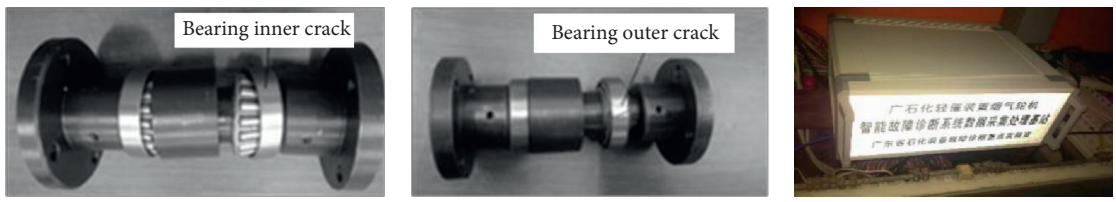


Figure 2: Bearing damage and fault signal acquisition platform. (a) Internal crack. (b) External crack. (c) Data acquisition platform.

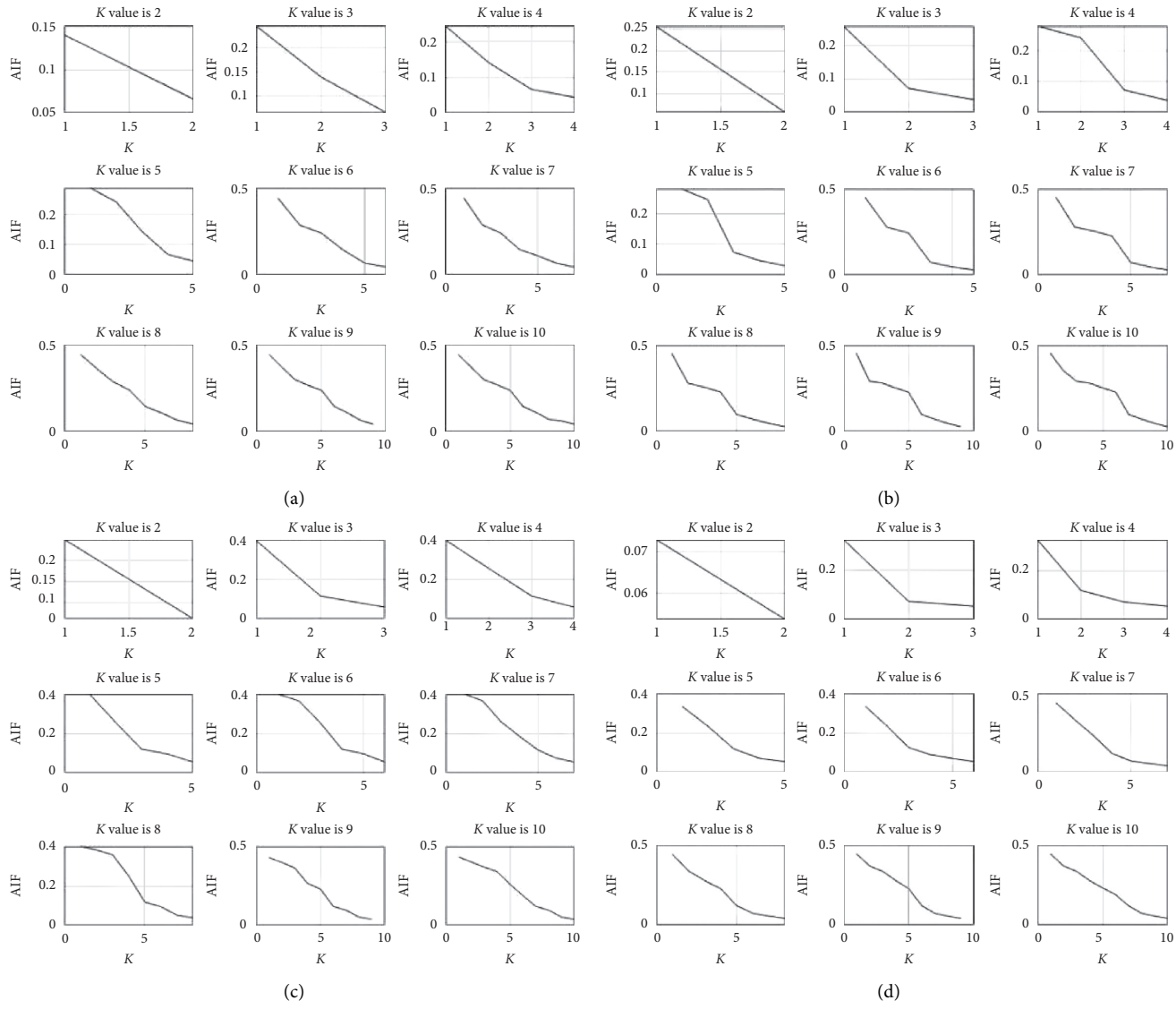


Figure 3: Broken line diagrams of instantaneous frequencies with different K values under various bearing conditions. (a) The changing curve of the K value under the normal state. (b) The changing curve of the K value with the worn state. (c) The changing curve of the K value with the internal crack. (d) The changing curve of the K value with the external state.

spectrum diagram obtained, VMD avoids the defects of modal aliasing and endpoint effect of decomposition methods such as EMD. Figure 5 shows the characteristic signals extracted from the signal components under the normal state. In this paper, 16 time-domain indexes are used to reflect the features. The three states of wear, internal crack, and external crack are also obtained through these steps, and then, these feature data are put into an Excel sheet to form a feature set. The feature set is divided into training data and test data and input into the LSSVM toolbox for fault pattern recognition.

This paper made three contrast figures of fault diagnosis precision. They are, respectively, as follows: VMD was optimized

Table 1: Mean instantaneous frequency corresponding to different K values under the normal condition of bearing.

K	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
2	0.1396	0.0656								
3	0.2451	0.1391	0.0656							
4	0.2450	0.1420	0.0661	0.0440						
5	0.2859	0.2396	0.1420	0.0661	0.0440					
6	0.4418	0.2859	0.2396	0.1420	0.0661	0.0440				
7	0.4418	0.2869	0.2397	0.1455	0.1081	0.0661	0.0430			
8	0.4418	0.3611	0.2859	0.2397	0.1455	0.1081	0.0661	0.0430		
9	0.4418	0.3670	0.2971	0.2655	0.2377	0.1450	0.1081	0.0661	0.0430	
10	0.4418	0.3690	0.2976	0.2684	0.2382	0.1455	0.1110	0.0712	0.0615	0.0430

Table 2: Euclidean norm for different values of K.

K value	2	3	4	5	6	7	8	9	10
Norm	0.3525	0.7622	1.3034	2.7008	2.7934	1.6146	16.5533	3.6468	20.5225

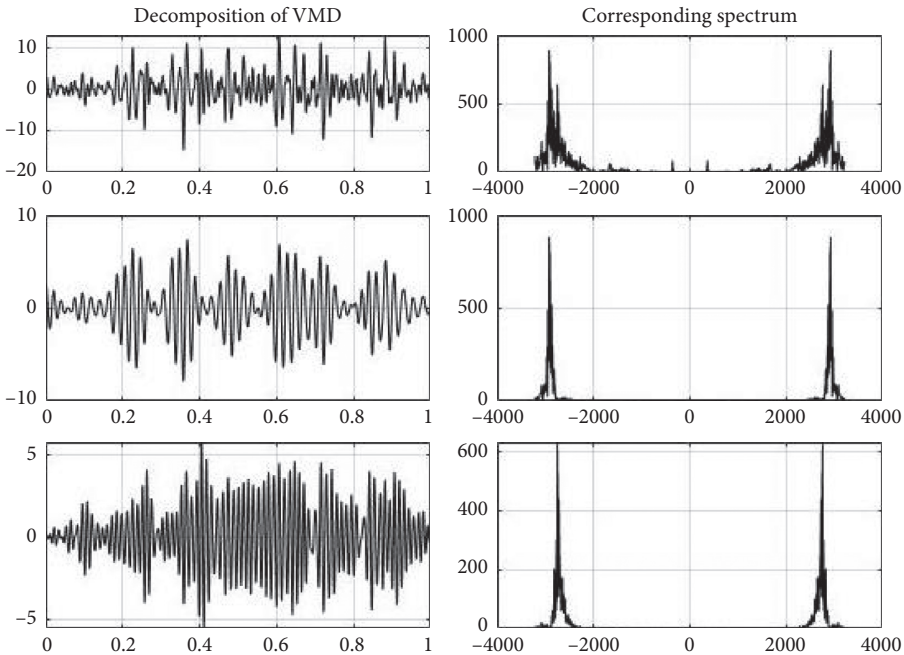


Figure 4: VMD decomposition spectrum at $K = 3$ in the normal state.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
########	1.88543	1.341274	1.599889	-1.02457	18.9554	3.554846	2.20633	-2.39983	4.606164	1.178475	1.1702	1.379052	1.644951	-0.15287	1.5
1.85E-17	0.80753	0.651561	0.716668	0.190782	0.637862	0.652105	1.075002	-0.87132	1.946323	1.126784	1.331222	1.5	1.649888	0.362294	1.5
3.70E-17	0.835586	0.764699	0.787485	0.411057	0.731235	0.698205	1.181227	-0.61948	1.800711	1.061082	1.413651	1.5	1.544697	0.704575	1.5
7.40E-17	3.334475	2.703836	2.965488	-13.8281	185.4391	11.11873	3.579848	-4.44823	8.028081	1.124427	1.073587	1.20717	1.323989	-0.37298	1.5
1.48E-16	2.874759	2.397531	2.588275	-10.3923	102.4465	8.26424	2.98596	-3.88241	6.868374	1.110685	1.038682	1.153649	1.245431	-0.43743	1.5
########	2.092366	1.873741	1.949671	-5.80732	28.75028	4.377996	1.852689	-2.92451	4.777196	1.073189	0.885452	0.950257	0.988765	-0.63396	1.5
########	1.055992	0.762716	0.900142	0.203066	1.865236	1.115119	1.350213	-1.22771	2.577925	1.17314	1.27862	1.5	1.770268	0.172447	1.5
4.63E-17	0.898326	0.666452	0.772358	0.15261	0.976847	0.806989	1.158538	-1.03074	2.189278	1.163094	1.289663	1.5	1.738366	0.210514	1.5
9.25E-18	0.615115	0.490548	0.543273	0.078661	0.214742	0.378366	0.814909	-0.67107	1.485977	1.13224	1.324808	1.5	1.661221	0.337979	1.5
7.40E-17	2.946918	2.221058	2.547458	6.018305	113.1262	8.684324	3.821187	-3.35121	7.1724	1.156807	1.296672	1.5	1.720436	0.235164	1.5
1.85E-17	2.562754	1.661417	2.12669	1.035979	64.70221	6.567709	3.190034	-3.08476	6.274791	1.205044	1.244768	1.5	1.920069	0.06155	1.5
0	1.925602	1.565912	1.714613	-2.70786	20.62326	3.707943	2.061096	-2.57192	4.633016	1.123053	1.070364	1.202076	1.316227	-0.37925	1.5
########	0.120137	0.096728	0.106525	-0.00062	0.000312	0.014433	0.129896	-0.15979	0.289684	1.127781	1.081238	1.219399	1.342899	-0.35781	1.5
########	0.197003	0.163988	0.17722	-0.00331	0.002259	0.03881	0.205139	-0.26583	0.470969	1.111633	1.0413	1.157543	1.250941	-0.43286	1.5
########	0.254554	0.200665	0.223777	-0.00519	0.006298	0.064798	0.280497	-0.33567	0.616163	1.137534	1.101913	1.253463	1.397839	-0.31491	1.5
########	2.661134	2.039942	2.314493	-4.96642	75.22431	7.081634	2.993931	-3.47174	6.465671	1.14977	1.125058	1.293558	1.467655	-0.26354	1.5
9.25E-17	2.281643	1.718056	1.971721	-2.77554	40.652	5.205894	2.596097	-2.95758	5.553679	1.157183	1.137819	1.316666	1.511067	-0.23367	1.5
########	1.625899	1.233083	1.408616	-1.05461	10.48251	2.643547	1.841946	-2.11292	3.95487	1.154253	1.132879	1.307629	1.493773	-0.24537	1.5
1.85E-17	0.402741	0.36874	0.379648	0.046126	0.039463	0.1622	0.569471	-0.29353	0.863002	1.060829	1.413989	1.5	1.544372	0.706095	1.5
0	0.324073	0.295647	0.304907	0.02362	0.016545	0.105023	0.457361	-0.25418	0.711544	1.062856	1.411291	1.5	1.546983	0.693993	1.5
1.85E-17	0.217502	0.198922	0.204911	0.007227	0.003357	0.047307	0.307366	-0.16393	0.471295	1.061446	1.413167	1.5	1.545163	0.702403	1.5
########	0.661091	0.600157	0.620391	0.195822	0.286507	0.437041	0.930586	-0.54319	1.473779	1.065604	1.407653	1.5	1.550571	0.677766	1.5
########	0.555364	0.492108	0.514713	0.10303	0.142693	0.308429	0.772069	-0.51084	1.282905	1.078978	1.390204	1.5	1.568902	0.601496	1.5
########	0.377365	0.277079	0.323344	0.010493	0.030418	0.142404	0.485016	-0.43532	0.920334	1.167068	1.285272	1.5	1.750461	0.195265	1.5
########	0.688011	0.574113	0.619601	-0.14289	0.336104	0.473359	0.714097	-0.9294	1.643499	1.11041	1.037915	1.152511	1.243827	-0.43876	1.5

Figure 5: Feature data extracted at $K = 3$ in the normal state.

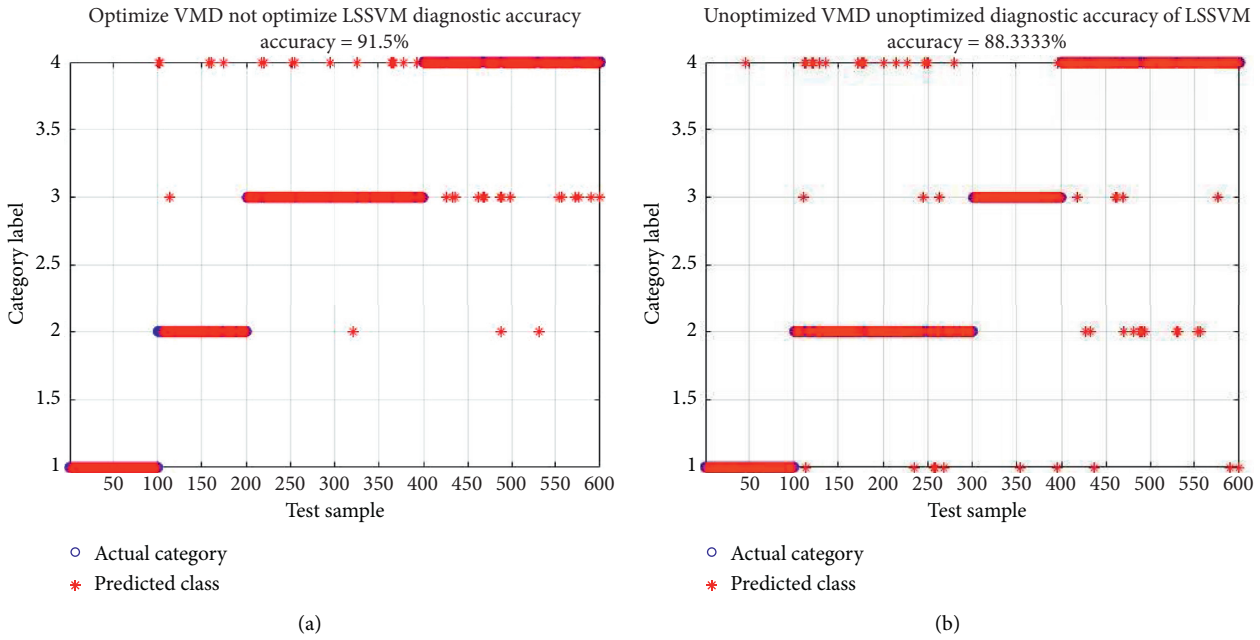


Figure 6: Comparison of diagnosis accuracy of optimized VMD, unoptimized LSSVM, and unoptimized. (a) Result of optimized VMD and unoptimized LSSVM. (b) Diagnosis accuracy of nonoptimized.

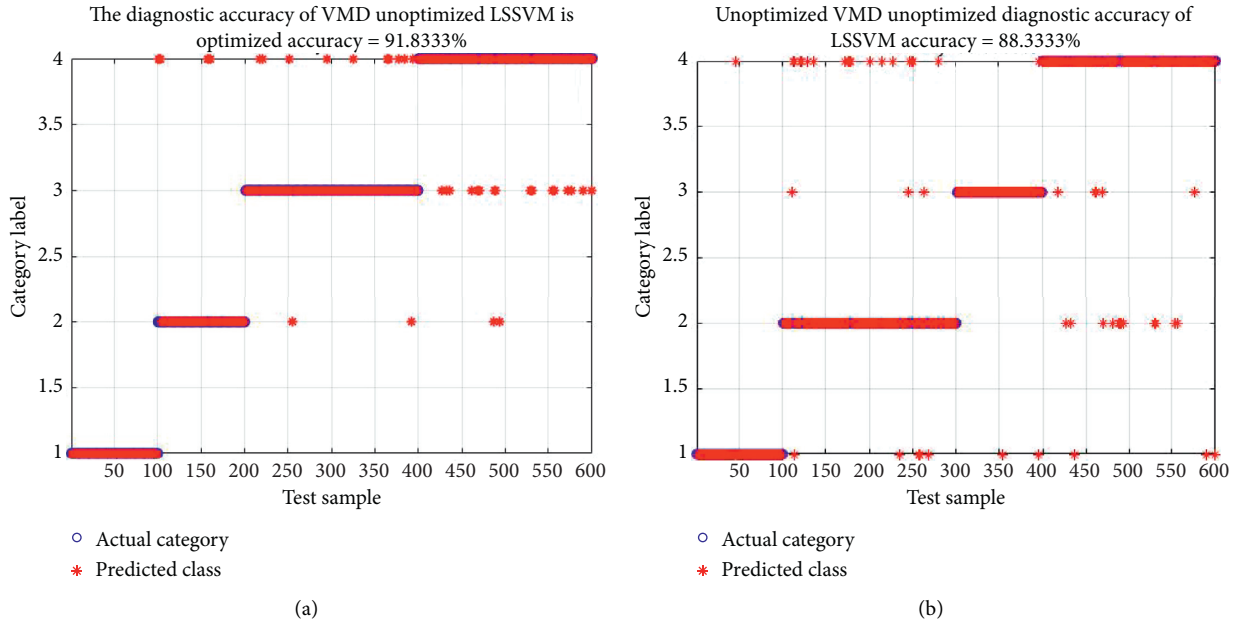


Figure 7: Comparison of diagnosis accuracy of optimized LSSVM, unoptimized VMD, and unoptimized. (a) Result of optimized LSSVM and unoptimized VMD. (b) Diagnosis accuracy of nonoptimized.

and unoptimized, LSSVM was optimized and unoptimized, as well as the condition of the VMD and LSSVM was optimized and unoptimized. Figure 6(a) shows that the fault diagnosis accuracy of optimized VMD is 91.5%. Figure 6(b) shows that the diagnostic accuracy of unoptimized VMD is 88.3333%, and the contrast figure from this group that can validate the proposed VMD optimization method is effective; Figure 7(a) shows that the fault diagnosis accuracy of optimized LSSVM is 91.8333%. Compared with the result of 88.3333% without optimization, the optimization of LSSVM also has the effect of improving the accuracy. When VMD and LSSVM were optimized, it further improved the accuracy of fault diagnosis, as shown in Figure 8(a), as the accuracy was 92%. Let the optimized VMD be abbreviated as P-VMD, and the optimized LSSVM is abbreviated as P-LSSVM. Table 3 lists and illustrates the fault diagnosis accuracy of different algorithms and our algorithm.

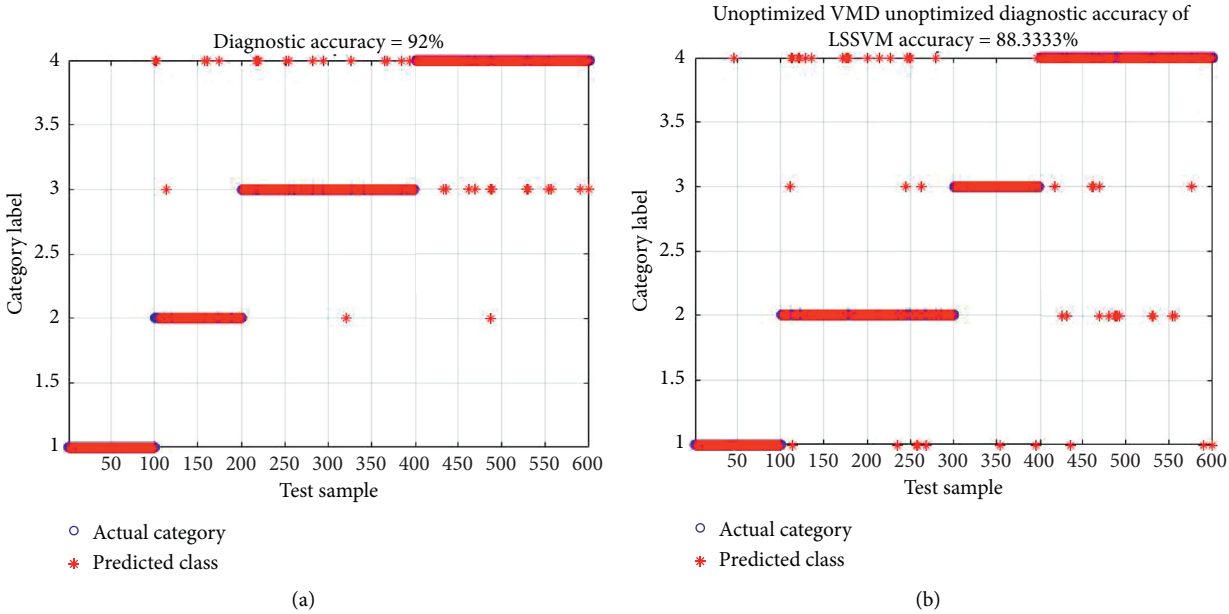


Figure 8: Comparison of diagnosis accuracy of both PSO and LSSVM optimized and nonoptimized. (a) Result of both PSO and LSSVM optimized. (b) Diagnosis accuracy of nonoptimized.

Table 3: Comparison results of fault diagnosis accuracy.

Methods	VMD + LSSVM	P-VMD + LSSVM	VMD + P-LSSVM	P-VMD + P-LSSVM
Accuracy	88.3333%	91.5%	91.833%	92%

From the comparison of three sets of results, we can clearly see that the proposed method in this paper based on instantaneous frequency optimization of the VMD fault diagnosis method is effective.

5. Conclusions

In this paper, the K value optimization problem of the variational modal decomposition algorithm is studied. Considering that the mode number K of VMD needs to be selected according to prior knowledge, improper selection will lead to overdecomposition or underdecomposition so that useful characteristic data cannot be extracted, ultimately leading to the problem of low accuracy of fault diagnosis. In this paper, the instantaneous frequency is used to find the optimal K value of VMD decomposition. Finally, the LSSVM model optimized by particle swarm optimization is combined to carry out fault pattern recognition. The results show the following:

(1) The advantage of measuring the value of K by the change of the instantaneous frequency of the signal component after VMD decomposition is more accurate and simple than the previous observation method to judge the value of K, which can avoid overdecomposition and underdecomposition.

(2) The optimized VMD decomposition algorithm can better reflect the characteristic parameters of vibration signals, which make subsequent feature extraction easier and helps to improve the diagnostic accuracy. As shown in the final experimental results, the accuracy of the optimized VMD is nearly 4% higher than that of the unoptimized results, indicating the effectiveness of this method.

(3) The use of the PSO-LSSVM classification model for fault diagnosis can further improve the accuracy of the final diagnosis. This conclusion can be verified by Figures 7 and 8 in the final experimental results.

It can be seen that a joint fault diagnosis method based on optimized VMD and LSSVM proposed in this paper improves the accuracy of fault diagnosis.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the Guangdong Institute of Petrochemical Technology for providing the data set of the rolling bearing in this paper. This work was supported in

part by the National Natural Science Foundation of P.R. China under Grant 61663008 and Chongqing Technology Innovation and Application Special Key Project under Grant cstc2019jscx-mbdxX0015.

References

[1] D. Bostjan, B. Pavle, and J. Dani, "Distributed bearing fault diagnosis based on vibration analysis," *Mechanical Systems and Signal Processing*, vol. 66-67, pp. 521–532, 2016.

[2] S. Alejandro, Z. Alejandro, G. Jacobo Manuel Machuca et al., "Early fault detection of single-point rub in gas turbines with accelerometers on the casing based on continuous wavelet transform," *Journal of Sound and Vibration*, vol. 487, 2020.

[3] A. Joshuva, R. Sathish, S. Sivakumar et al., "An insight on VMD for diagnosing wind turbine blade faults using C4.5 as feature selection and discriminating through multilayer perceptron," *Alexandria Engineering Journal*, vol. 59, no. 5, pp. 3863–3879, 2020.

[4] Q. Yang, JY. Zhang, L. Chen et al., "Fault diagnosis of motor bearing based on improved convolution neural network based on VMD," in *Proceedings of the 2019 31st Chinese Control and Decision Conference (CCDC 2019)*, pp. 405–409, Nanchang, China, June 2019.

[5] H. Pan, Y. Yang, X. Li, J. Zheng, and J. Cheng, "Symplectic geometry mode decomposition and its application to rotating machinery compound fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 114, pp. 189–211, 2019.

[6] Q. Yang, J. Ruan, and Z. Zhuang, "Fault diagnosis for circuit-breakers using adaptive chirp mode decomposition and attractor's morphological characteristics," *Mechanical System and Signal Processing*, vol. 145, 2020.

[7] Y. J. Mao, M. P. Jia, and X. A. Yan, "A new bearing weak fault diagnosis method based on improved singular spectrum decomposition and frequency-weighted energy slice bispectrum," *Measurement*, vol. 166, 2020.

[8] H. Li, C. Wang, and D. Zhao, "Filter bank properties of envelope modified EMD methods," *IET Signal Processing*, vol. 12, no. 7, pp. 844–851, 2018.

[9] H. Yang, S. Liu, and H. Zhang, "Adaptive estimation of VMD modes number based on cross correlation coefficient," *Journal of Vibroengineering*, vol. 19, no. 2, pp. 1185–1196, 2017.

[10] X. B. Bi, J. S. Lin, D. J. Tang et al., "VMD-KFCM algorithm for the fault diagnosis of diesel engine vibration signals," *Energies*, vol. 13, no. 1, 2020.

[11] J. Ding, D. Xiao, and X. Li, "Gear fault diagnosis based on genetic mutation particle swarm optimization VMD and probabilistic neural network algorithm," *IEEE Access*, vol. 8, pp. 18456–18474, 2020.

[12] D. M. Xiao, J. K. Ding, X. J. Li et al., "Gear fault diagnosis based on kurtosis criterion VMD and SOM neural network," *Applied Sciences-Basel*, vol. 9, no. 24, 2019.

[13] J. K. Ding, L. P. Huang, D. M. Xiao et al., "GMPSO-VMD algorithm and its application to rolling bearing fault feature extraction," *Sensors*, vol. 20, no. 7, 2020.

[14] H. Jin, J. H. Lin, and X. Q. Chen, "VMD entropy method and its application in early fault diagnosis of bearing," in *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning (SPML 2018)*, pp. 128–134, Shanghai, China, November 2018.

[15] H. Luo, S. Yin, T. Liu, and A. Q. Khan, "A data-driven realization of the control-performance-oriented process monitoring system," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 1, pp. 521–530, 2020.

[16] H. Luo, X. Yang, M. Krueger, S. X. Ding, and K. Peng, "A plug-and-play monitoring and control architecture for disturbance compensation in rolling mills," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 200–210, 2018.

[17] H. Luo, K. Li, O. Kaynak, S. Yin, M. Huo, and H. Zhao, "A robust data-driven fault detection approach for rolling mills with unknown roll eccentricity," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2641–2648, 2020.

[18] X. L. Zhang, W. Chen, B. J. Wang, and X. F. Chen, "Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization," *Neurocomputing*, vol. 167, pp. 260–279, 2015.

[19] Y. Wang, "Research on bearing fault diagnosis of large machinery based on mathematical morphology," *Advances in Materials, Machinery, Electronics II*, vol. 1955, 2018.

[20] A.-b. Ji, J.-h. Pang, and H.-j. Qiu, "Support vector machine for classification based on fuzzy training data☆," *Expert Systems With Applications*, vol. 37, no. 4, pp. 3495–3498, 2010.

[21] RL. Lang, ZP. Xu, F. Gao et al., "A knowledge acquisition method for fault diagnosis of airborne equipments based on support vector regression machine," *Chinese Journal of Electronics*, vol. 22, no. 2, pp. 277–281, 2013.

[22] R. F. Zhang and Y. X. Liu, "Research on development and application of support vector machine-transformer fault diagnosis," in *Proceedings of the International Symposium on Big Data and Artificial Intelligence (ISBDAI'18)*, pp. 262–268, Hong Kong, China, December 2018.

[23] H. Y. Xiao, Y. J. Sun, Y. B. Jin et al., "Optimization about fault prediction and diagnosis of wind turbine based on support vector machine," in *Proceeding of the 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 139–142, Changsha, China, February 2018.

[24] F. Chen, B. Tang, and R. Chen, "A novel fault diagnosis model for gearbox based on wavelet support vector machine with immune genetic algorithm," *Measurement*, vol. 46, no. 1, pp. 220–232, 2013.

[25] Y. P. Du, W. J. Zhang, Y. Zhang et al., "Fault diagnosis of rotating machines for rail vehicles based on local mean decomposition-energy moment-directed acyclic graph support vector machine," *Advances in Mechanical Engineering*, vol. 8, no. 1, 2016.

[26] X. Xie and S. Sun, "Multitask centroid twin support vector machines," *Neurocomputing*, vol. 149, pp. 1085–1091, 2015.

[27] X. W. Xu, N. Zhang, Y. B. Yan et al., "Smooth iteration online support tension machine algorithm and application in fault diagnosis of electric vehicle extended range," *Advances in Mechanical Engineering*, vol. 10, no. 12, 2018.

[28] S. Ma, B. Cheng, Z. Shang, and G. Liu, "Scattering transform and LSPTSVM based fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 104, pp. 155–170, 2018.

[29] C. Li, S. Li, and Y. Liu, "A least squares support vector machine model optimized by moth-flame optimization algorithm for annual power load forecasting," *Applied Intelligence*, vol. 45, no. 4, pp. 1166–1178, 2016.

[30] T. Wu, C. C. Liu, and C. He, "Fault diagnosis of bearings based on KJADE and VNWOA-LSSVM algorithm," *Mathematical Problems in Engineering*, vol. 2019, Article ID 8784154, 19 pages, 2019.

[31] A. Tharwat and A. E. Hassanien, "Quantum-behaved particle swarm optimization for parameter optimization of support

vector machine," *Journal of Classification*, vol. 36, no. 3, pp. 576–598, 2019.

[32] A. Tharwat and A. E. Hassanien, "Chaotic antlion algorithm for parameter optimization of support vector machine," *Applied Intelligence*, vol. 48, no. 3, pp. 670–686, 2018.

[33] X. S. Zhang and F. Pan, "Parameters optimization and application to glutamate fermentation model using SVM," *Mathematical Problems in Engineering*, vol. 2015, Article ID 320130, 7 pages, 2015.

[34] G. Gintautas and D. Paulius, "Particle swarm optimization for linear support vector machines based classifier selection," *Nonlinear Analysis-Modelling and Control*, vol. 19, no. 1, pp. 26–42, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6629474, 15 pages
https://doi.org/10.1155/2021/6629474

Research Article

A Parameter-Optimized Variational Mode Decomposition Investigation for Fault Feature Extraction of Rolling Element Bearings

Guoping An,[1] Qingbin Tong,[1] Yanan Zhang,[2] Ruifang Liu,[1] Weili Li,[1] Junci Cao,[1] Yuyi Lin,[3] Qiang Wang,[1] Ying Zhu,[1] and Xiaowen Pu[1]

[1]*School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China*
[2]*State Grid JIBEI Electric Power Co., Ltd. Maintenance Branch State Grid, Beijing 102488, China*
[3]*Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia MO 65211, USA*

Correspondence should be addressed to Qingbin Tong; qbtong@bjtu.edu.cn

Received 21 October 2020; Revised 30 January 2021; Accepted 8 February 2021; Published 20 February 2021

Academic Editor: Yong Chen

Reliable fault diagnosis of the rolling element bearings highly relies on the correct extraction of fault-related features from vibration signals in time-frequency analysis. However, considering the nonlinear, nonstationary characteristics of vibration signals, the extraction of fault features hidden in the heavy noise has become a challenging task. Variable mode decomposition (VMD) is an adaptive, completely nonrecursive method of mode variation and signal processing. This paper analyzes the advantages of VMD compared with EMD in robustness of against noise, overcoming the end effect and mode aliasing. The signal decomposition performance of VMD algorithm largely depends on the selection of mode number k and bandwidth control parameter α. To realize the adaptability of influence parameters and the improvement of decomposition accuracy, a parameter-optimized VMD method is presented. The random frog leaping algorithm (SFLA) is used to search the optimal combination of influence parameters, and the mode number and bandwidth control parameters are set according to the search results. A multiobjective evaluation function is constructed to select the optimal mode component. The envelope spectrum technique is used to analyze the optimal mode component. The proposed method is evaluated by simulation and practical bearing vibration signals under different conditions. The results show that the proposed method can improve the decomposition accuracy of the signal and the adaptability of the influence parameters and realize the effective extraction of the bearing vibration signal.

1. Introduction

Rolling element bearings, as a very important component of rotating machinery, has been widely used in modern industry such as engineering machinery and aerospace [1, 2]. The working state of rolling element bearings is directly related to the safety of the rotating machinery. Rolling element bearings are easily damaged under the long-term operation of harsh environment with high speed, heavy load, strong impact, and high temperature. The developed mechanical faults may cause the deterioration of machine operating conditions, resulting in serious economic losses and casualties [3–5]. The vibration signal detected by the sensor is always related to the important physical information that a series of shock pulses will occur when the rolling element bearing is subjected to a local fault [6, 7]. However, the defect-induced impulses in practice are too weak to distinguish well from vibration signal corrupted by a large amount of background noise. Therefore, it is critical to remove noise and extract intrinsic fault features from the measured original signal for the fault diagnosis of rolling element bearing.

Many vibration analysis methods have been proposed in the literature for bearing fault detection in the time domain, the frequency domain, and the time-frequency domain, respectively [8, 9]. However, the vibration signal of rolling

element bearings is the nonstationary and nonlinearity signal. It is very difficult to identify the fault characteristics of the rolling element bearing only in the analysis of time domain or the frequency domain. To effectively analyze the fault features from the vibration signals, some traditional time-frequency analysis methods have been widely used, such as short-time Fourier transform (STFT) [10], Wigner–Ville distribution (WVD) [11], and wavelet transform (WT) [12]. However, due to the limitation of Heisenberg's uncertainty principle, STFT method cannot get high resolution in the time domain and frequency domain simultaneously when dealing with the nonstationary signals. The disadvantage of the WVD method is that it cannot guarantee nonnegativity and produce serious cross-term interference for multicomponent signals or signals with complex modulation laws. The WT method decomposes the signal by performing scaling and translation operations on the wavelet basis and can effectively obtain time-frequency information from the measured signal. It has good localization properties in the time domain and frequency domain and has multiresolution analysis features [13, 14]. However, the WT cannot accurately split the high-frequency band where the modulation information of machine fault always exists.

Compared with the traditional analysis methods, empirical mode decomposition (EMD) offers a different analysis approach to signal processing in the time-frequency domain. The EMD provides more realistic signal representations without artifacts imposed by the nonadaptive limitations of both Fourier and wavelet transform-based time-frequency analysis methods and is suitable for the analysis of the nonlinear and nonstationary signals [15–17]. It is based on the local characteristic time scales of a signal and can self-adaptively decompose the complicated signal into a limited number of intrinsic mode functions (IMFs) through automatically performing a series of recursive calculations. The IMFs represent the fundamental oscillatory modes embedded in the signal, from which the instantaneous time-frequency features of interest are deemed to be observed. This enables the EMD-based methods to have potential as promising tools for dealing with the engineering problems associated with the analysis of nonstationary signals [18]. Therefore, the EMD and its extension forms (such as Ensemble Empirical Mode Decomposition (EEMD)) have attracted the attention of many researchers and are widely applied in the fault diagnosis and recognition of rolling element bearings [19–23]. In practical applications, although the EMD and its improved method have advantages in the processing of the nonstationary signals, the method itself still has the following inherent defects:

Weak Robustness of against Noise. The EMD-based methods are sensitive to the complex noise in the vibration signal. A little change in the signal-to-noise ratio (SNR) can lead to the different signal decomposition results [18].

Mode Aliasing. The local mean is defined by the upper and lower envelopes of the signal in the EMD. Based on this definition, different modal components can be distinguished through the characteristic scale of the signal. The IMF is no longer limited to the narrowband signal, and it can also show amplitude modulation and frequency modulation at the same time. However, when there is a jump change in the time scale of the signal, the direct screening process will produce mode aliasing issues. Intuitively, it is impossible to effectively separate the different modal components according to the characteristic scale, which makes the existing IMFs contain the different time-scale components and cannot clearly reflect the intrinsic properties of the signal.

End Effect. The upper and lower envelopes of the signal are interpolated by the cubic spline interpolation in the EMD. The cubic spline interpolation needs two adjacent points. As a result, the divergence occurs at both ends of the data, and the divergent results gradually "pollute" the whole data sequence during the data decomposition process, which leads to the serious distortion and energy leakage.

Variational mode decomposition (VMD) method has been proposed and developed recently, which is an alternative nonrecursive signal decomposition method that can adaptively determine the relevant frequency bands and the corresponding mode simultaneously [24–26]. The VMD method decomposes a signal into a series of band-limited modes. These modes can be continuously updated with Wiener filtering, and the central frequency of each mode can be gradually demodulated to the corresponding baseband. The nonrecursive signal decomposition of VMD is more efficient than the EMD and its extension forms in computation. At the same time, the application of Wiener filtering makes the VMD method robust to the background noise. Due to the application of Wiener filters, the narrowbanded function of VMD resultant modes not only reduces the mode mixing issues existing in the EMD but also helps to accurately extract the fault characteristics of the signal through the Hilbert transform. However, the decomposition accuracy of the VMD method is usually affected by the number of modes k and the bandwidth control parameter α. The original VMD method used the default values to implement the signal analysis, which largely limits its decomposition precision and the capability of feature extraction to a certain extent.

In this paper, we firstly analyze the advantages of VMD compared with EMD in robustness of against noise, overcoming the end effect and mode aliasing. To realize the adaptability of influence parameters and the improvement of decomposition accuracy, a parameter-optimized VMD method is presented. The random frog leaping algorithm (SFLA) is used to search the optimal combination of influence parameters, and the mode number and bandwidth control parameters are set according to the search results. A multiobjective evaluation function is constructed to select the optimal mode component. The envelope spectrum technique is used to analyze the optimal mode component. The proposed method is evaluated by simulation and practical bearing vibration signals under different conditions. The remaining section of the paper is organized as

follows: Section 2 introduces the fundamental theory of the VMD. The superiorities of the VMD over the EMD are analyzed in Section 3. The parameter-optimized VMD algorithm is presented in Section 4. The fault feature extraction based on the parameter-optimized VMD is given in Section 5. Section 6 will present the experimental results and analysis. Finally, the conclusion is drawn in Section 7.

2. Brief Introduction to VMD

The VMD algorithm is an adaptive, quasiorthogonal, and completely nonrecursive signal processing method. It decomposes the input signals composed of multicomponents into several inherent modes with limited bandwidth, and most of these modes are closely around their corresponding central frequencies [24]. By solving the optimal solution of constrained variational problem, the central frequency and band limit of each mode can be decided. An input signal $f(t)$ can be expressed as follows:

$$f(t) = \sum_{k=1}^{K} u_k(t), \tag{1}$$

where the number of modes k is defined in advance and $u_k(t)$is the narrowband mode function. It can be written as

$$u_k(t) = A_k(t)\cos\left(\phi_k(t)\right), \tag{2}$$

where $A_k(t)$ is the instantaneous amplitude of $u_k(t)$, $\phi_k(t)$ is the instantaneous phase, and $\phi_k(t)$ is the reduction function that instantaneous frequency $\omega_k(t) = d\phi_k(t)/dt \geq 0$. Compared to $\phi_k(t)$, the variation in $A_k(t)$ and $\omega_k(t)$ is more gradual that can be regarded as a harmonic signal of constant amplitude and frequency in a smaller time horizon.

The VMD decomposes the input signal into a certain number of modes, and the decomposed modes have specific sparsity property while reproducing the input signal. It is assumed each mode is closely integrated around the center frequency. To assess the bandwidth of a mode, the following scheme is needed: the VMD method decomposes the input signal into a certain number of modes, which make them reappear the input signal and have specific sparsity properties. It is assumed each mode is closely integrated around the center frequency. To assess the bandwidth of a mode, the following scheme is needed: (1) compute the analyzed signal by means of the Hilbert transform to get a one-sided frequency spectrum for the mode; (2) transform the frequency spectrum of each mode to the baseband by mixing with an exponential tuned to the estimated center frequency; (3) estimate the bandwidth through the H^1 Gaussian smoothness of the demodulated signal, that is, the squared L^2-norm of the gradient. The constrained variational problem would be expressed as follows:

$$\begin{aligned} &\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_k \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t_2^2} \right\}, \\ &\text{s.t.} \sum_k u_k(t) = f(t), \end{aligned} \tag{3}$$

where $\{u_k\}(k = 0, 1, 2, \ldots, K)$ represents the k-th mode component obtained by decomposition and $\{\omega_k\}$ represents the corresponding central frequencies of the k-th mode component.

To solve the constrained variational problem, the augmented Lagrange is introduced and the unconstrained variational problem is gotten by

$$L\left(\{u_k\}, \{\omega_k\}, \lambda\right) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle, \tag{4}$$

where α represents the quadratic penalty factor, which can guarantee the accuracy of signal reconstruction in the presence of Gauss noise, and λ represents the Lagrange operator, which can be used to maintain the strictness of constraints. The saddle point of the augmented Lagrange L is the optimal solution of original minimization problem, which can be solved using alternate direction method of multipliers (ADMM). All the modes can be obtained from (5) in the frequency domain through updating each mode:

$$\widehat{u}_k^{n+1}(\omega) = \frac{\widehat{f}(\omega) - \sum_{i \neq k} \widehat{u}_i(\omega) + \widehat{\lambda}(\omega)/2}{1 + 2\alpha\left(\omega - \omega_k\right)^2}, \tag{5}$$

where $\widehat{u}_k^{n+1}(\omega)$ can be equivalent to the Wiener filter of the current residual signal and the full spectrum of the real mode can be obtained by conjugate symmetry. Thus, the real part $\{u_k(t)\}$ can be achieved through utilizing the inverse Fourier transform of $\{\widehat{u}_k^{n+1}(\omega)\}$.

Similarly, to obtain the minimum value of ω_k^{n+1}, the central frequency updating problem can be transformed into the corresponding frequency domain, and the solutions of the central frequencies can be given as follows:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left| \widehat{u}_k(\omega) \right|^2 d\omega}{\int_0^\infty \left| \widehat{u}_k(\omega) \right|^2 d\omega}. \tag{6}$$

Therefore, the new value of ω_k can be set to the center of gravity of the corresponding modal power spectrum.

To update the Lagrange operator λ, the following expression is given:

$$\widehat{\lambda}^{n+1}(\omega) = \widehat{\lambda}^n(\omega) + \tau \left(\widehat{f}(\omega) - \sum_k \widehat{u}_k^{n+1}(\omega) \right). \tag{7}$$

According to the above theoretical description, the detailed procedures of VMD algorithm are given as follows:

(1) Define the number of modes k;

(2) Initialize: $\{\widehat{u}_k^1\}\{\widehat{\omega}_k^1\}$, $\widehat{\lambda}^1$ and $n = 0$;

(3) Update u_k and ω_k according to Equations (5) and (6);

(4) Update λ according to Equation (7);

(5) Set the error $\varepsilon > 0$, if the inequality $(\sum_k \|\widehat{u}_k^{n+1} - \widehat{u}_{k^2}^{n^2}\| / \|\widehat{u}_{k^2}^{n^2} < \varepsilon\|)$ holds, then the iteration stops, or else go back to step 2.

3. Superiorities of VMD over EMD

The VMD algorithm transfers the signal decomposition process to the variational framework and achieves adaptive signal decomposition by searching the optimal solution of the constrained variational model. Compared with the EMD method, the VMD method has more advantages in the noise robustness, the mode aliasing, and the end effect. The superiorities of the VMD over the EMD will be investigated in this section.

3.1. Noise Robustness and Elimination of Mode Aliasing. To verify the advantages of VMD in the noise robustness, the mode aliasing, and the end effect, a simulated signal $x(t)$ is designed in Figure 1, which is the sum of a harmonic signal and white noise η. The signal is modulated by 30 Hz and 56 Hz. The purpose of this simulation is to simulate the phenomenon that fault vibration signal is a multicomponent modulation signal. The noise existing in the original signal often appears as white noise in the practical applications, which covers the entire frequency domain. White noise with a signal-to-noise ratio of 2 db is added to the signal. The simulated signal is expressed as follows:

$$x(t) = \sin(2\pi \times 30t)) + \frac{1}{4}\sin(2\pi \times 56t) + \eta. \tag{8}$$

The time waveform of the simulated signal collected by using a sampling frequency of 1000 Hz is shown in Figure 1. It can be clearly seen from Figure 1 that the harmonic signal has been seriously distorted by noise. Then, the EMD method is applied to process the simulated signal, and the corresponding results are shown in Figure 2.

From the spectrum diagram of the EMD decomposition shown in Figure 2, it can be seen that the extracted IMF4 mainly includes 30 Hz frequency components, but not 56 Hz frequency components. For the multicomponent simulation signal, the decomposed results also show that the decomposition effect of the EMD algorithm is not ideal, and there is mode aliasing the first three-order intrinsic modal function (IMFs). This is because some useful weak signals are submerged in the decomposed signal. The cubic spline fitting in the decomposition process of the EMD method leads to the deviation of decomposition. In addition, the first three-order IMFs also contain other components of the mode function, and the decomposition of EMD has pseudocomponents.

The VMD method is more effective for the decomposition of simulation signal; not only it can effectively remove pseudocomponents but also each IMF component shows a certain scale of modalities, and there is no mode aliasing between each other. VMD can realize the multiscale representation of the simulated signal. Compared with Figures 2 and 3, the VMD method has stronger ability in the noise filtering than EMD method. This method successfully suppresses the noise distributed in different frequency bands, and its decomposition effect is better than the EMD decomposition method. EMD cannot effectively remove the noise, especially in the high frequency band. This means that noise still exists in the IMFs generated by EMD.

3.2. Suppression of End Effect. The decomposition results of EMD and VMD are orthogonal to the signal. That is, the energy sum of the decomposed mode functions is equal to the signal energy before the decomposition. However, if the decomposition result has an end effect, it will affect the decomposition accuracy of the signal and produce false components, which will result in a change in the energy sum of the modal function after decomposition. By analyzing the changes in the energy value before and after decomposition, we can understand the inhibitory effect on the signal end effect when the two methods are used for the signal analysis.

The energy expression of the mode function generated after the signal decomposition by the EMD and VMD can be expressed as follows:

$$E = \sqrt{\frac{\sum_{i=1}^{n} x^2(i)}{n}}, \tag{9}$$

where E is the energy of the original signal or the energy of the mode functions after the decomposition, $x(i)$ is the signal sequence, and n is the number of sample points for the signal.

Comparing the deviation between the energy of all mode functions and the energy of the original signal, the evaluation index ξ can be defined as

$$\xi = \frac{\left|\sqrt{\sum_{k=1}^{m} E_k^2 - E_i}\right|}{E_i}, \tag{10}$$

where E_i is the energy of the original signal, E_k is the energy of the k-th modal function, and m is the total number of modal functions.

From the definition of the evaluation index depicted above, it can be seen that when ξ is the larger, the energy of the decomposed mode functions will be smaller. That is, the energy leakage after the signal decomposition becomes larger, and the end effect will become stronger. The simulated signals $X_2(t)$ and $X_3(t)$ are constructed as follows:

$$\begin{aligned} X_2(t) = {} & \sin(2\pi \times 3t) + 0.8\ \sin(2\pi \times 15t) \\ & + 0.4\ \sin(2\pi \times 45t))0.6\ \sin(2\pi \times 63t) + \sin(2\pi \times 90t), \end{aligned} \tag{11}$$

$$\begin{aligned} X_3(t) = {} & [1 + 0.5\ \sin(2\pi \times 3t)] \\ & \times \sin[2\pi \times 5t + \sin(2\pi \times 50t)] \\ & + [1 + \sin(2\pi \times 6t)] \\ & \times \sin[2\pi \times 8t + (0.6 \times \sin(2\pi \times 5t))]. \end{aligned} \tag{12}$$

The equations (8), (11), and (12) of the simulated signal are used to calculate the energy E_i, respectively. The energy E_k of

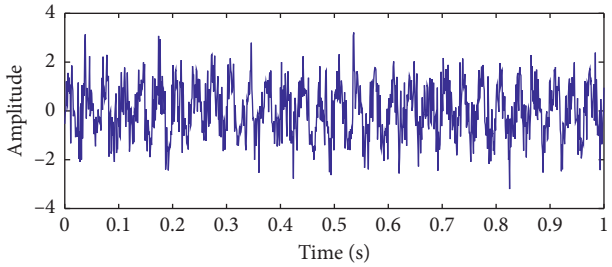


Figure 1: The time domain plot of the simulated signal.

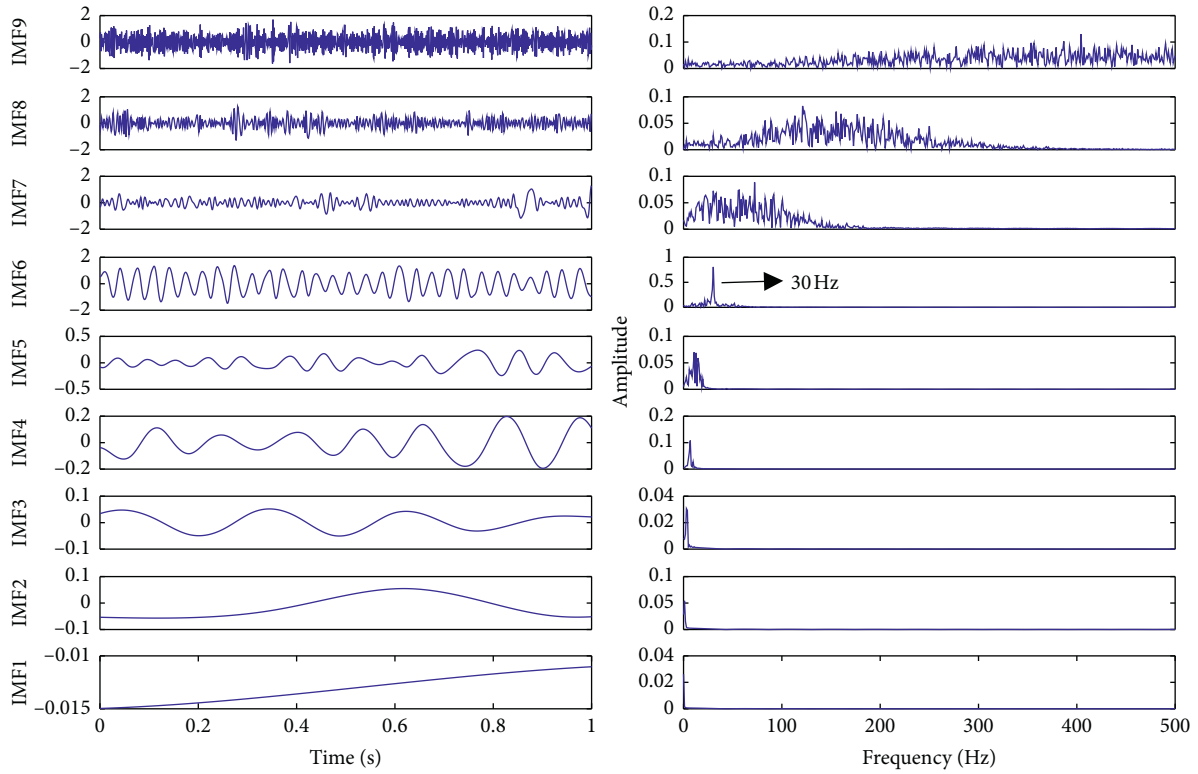


Figure 2: The decomposition results and corresponding frequency spectrum of IMFs with EMD.

each modal function and the evaluation parameter ξ of energy deviation after the decomposition of EMD and VMD are also calculated. The calculation results of the evaluation parameter ξ are shown in Table 1. It can be seen from Table 1 that the values of the evaluation parameter calculated after VMD decomposition are small, which indicates that compared with the EMD method, the energy leakage calculated by the VMD decomposition is smaller and the end effect is not obvious.

4. The Proposed Parameter-Optimized VMD Algorithm

The number of modes k and bandwidth control parameter α affect the accuracy of decomposition in the VMD algorithm. A large number of modes will lead to the redundant information in the result of signal decomposition, while a small number of modes will result in the phenomenon of mode mixing. On the contrary, a wider filter bandwidth will introduce more noise and interference items into the decomposition result. The narrow filter bandwidth will cause important information missing in the signal decomposition. Therefore, how to choose the optimal parameter combination is the key to eliminate the noise and mode aliasing and extract the feature information accurately in the VMD algorithm. In this section, shuffled frog leaping algorithm (SFLA) is introduced into the algorithm to achieve the combination of optimal influence parameters [27–30]. A multiobjective evaluation function is constructed to select the optimal mode component in the VMD algorithm.

4.1. Shuffled Frog Leaping Algorithm. The SFLA simulates the thought transfer process of frogs in searching for the food according to their population. It combines global

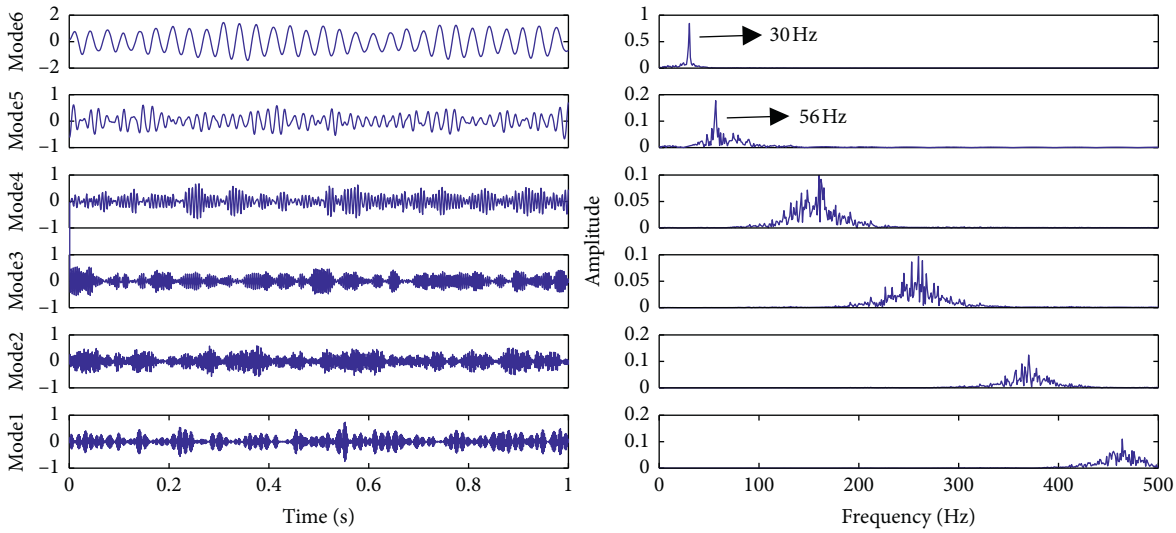


Figure 3: The decomposition results and corresponding frequency spectrum of modes with VMD.

Table 1: Energy leakage evaluation parameter.

Simulated signal	ξ	
	EMD	VMD
$X_1(t)$	0.2964	0.1620
$X_2(t)$	0.3648	0.0618
$X_3(t)$	0.3643	0.0425

information exchange and local deep search. Local search enables thoughts to be transmitted between local individuals, and hybrid strategies enable the exchange of local thoughts. Through this global information exchange and local depth exploration, the algorithm can jump out of the local extreme points and move towards the global optimum.

An initial population of P frogs is randomly generated within the S-dimensional space. The i-th frog is represented by S variables as $X_i = (x_{i1}, x_{i2}, \ldots, x_{iS})$. In each evolutionary iteration process, all frogs are arranged in a descending order according to the fitness value of the frogs. The population is divided into m subsets. The subset is referred to as memeplexes, and each contains n frogs. The method of allocation is the first frog enters the first memeplex, the second frog goes to the second memeplex, the m frog goes to the m-th memeplex, and the $m+1$ frog goes back to the first memeplex and so forth. Assuming that M^k is a set of frogs for the k-th memeplex, the allocation process can be described as follows:

$$M^k = \left\{X_{k+m(l-1)} \in P \mid 1 \le l \le n\right\}, \quad 1 \le l \le m. \tag{13}$$

Within each memeplex, the frogs with the best and the worst fitness are identified as X_b and X_ω, respectively. Also, the frog with the global best fitness is identified as X_g. Then, an evolution process is applied to improve only the frog with the worst fitness (i.e., not all frogs) in each cycle. Accordingly, the position of the frog with the worst fitness is adjusted as follows:

$$D_i = \text{Rand} \times (X_b - X_\omega), \tag{14}$$

$$X'_\omega = X_\omega + D_i, D \le D_{\max}, \tag{15}$$

where $D_{\max}$ is the maximum allowed change for the position of the frog.

If the evolution process produces a better frog (solution), it replaces the worst frog. Otherwise, the calculations in equations (14) and (15) are repeated with respect to the global best frog (i.e., X_g replaces X_b). There is no improvement in this situation, and a new solution will be randomly generated, that is, to replace the worst frog with another frog with any fitness. The calculation will continue for a specific number of evolutionary iterations in each memeplex. Therefore, SFLA uses a process similar to the PSO algorithm to simultaneously perform independent local searches in each memeplex.

A predetermined number of memetic evolution steps are performed in each memeplex, and the solution of the evolved memeplexes $\{X_1, X_2, \ldots, X_P\}$ is replaced with a new population, which is called the shuffling process. The shuffling process facilitated the global exchange of information among frogs. Then, the population is sorted in the descending order of fitness value, the position X_g of the best frog of the population is updated, and the frog group is redivided into the memeplexes and evolved in each memeplex until the conversion criterion is met. Generally, the convergence criterion can be defined as follows:

The relative change in the fitness of the global frog within a number of consecutive shuffling iterations is less than a prespecified tolerance.

The maximum predefined number of shuffling iterations has been obtained.

4.2. Parameter Optimization by Using SFLA. The SFLA is a metaheuristic intelligent optimization algorithm that has good capabilities of global optimization and fast convergence speed. The SFLA combines the advantages of gene-based memetic algorithm (MA) and the social behavior-based particle swarm optimization (PSO) algorithm. Therefore, the SFLA is used to optimize the influencing parameters of VMD, can avoid the intervention of subjective factors, and automatically screen out the best combination of influencing parameters.

Suppose that the population composed of N_{pop} frogs is X in D dimension space, and N_{pop} frogs are divided into N_m subgroups through the descending order. The best individual p_b and the worst individual p_w in the subgroup can be calculated. Group optimal solution s_1 in the maximum number of iterations M can be expressed as follows:

$$s_1 = A \times (p_b - p_w). \tag{16}$$

When using SFLA to optimize the mode number k and bandwidth control parameter α, the fitness function needs to be determined. Each update of the frog is achieved by comparing the fitness values.

Shannon entropy is a good indicator for evaluating signal sparsity. The size of entropy reflects the uniformity of probability distribution. The most uncertain probability distribution (equal probability distribution) has the largest entropy value. In order to reflect the sparseness of the measured signal, the concept of envelope entropy is proposed. The demodulated envelope signal is processed into a probability distribution sequence. The calculated entropy value reflects the sparsity of the original measurement signal [31]. The envelope entropy of the signal can be expressed as follows:

$$\left.\begin{aligned} E_p &= -\sum_{i=1}^{N} p_i \lg p_i \\ p_i &= a(i) / \sum_{i=1}^{N} a(i) \end{aligned}\right\}. \tag{17}$$

In order to search the global optimal component, that is, to extract the mode component with the most abundant feature information from the bearing fault signal, the multiobjective evaluation function is constructed for the selection of the optimal mode component and the calculation of fitness value, which is based on the envelope entropy, the kurtosis, and the correlation coefficients. When the i-th frog is located in the position j (corresponding to a set of parameters α_j and k_j), the kurtosis, the correlation coefficient, and the envelope entropies of all mode components obtained by VMD processing are all calculated. The components with the largest kurtosis value, the highest correlation, and the smallest envelope entropy are selected and reconstructed as the fitness value in the optimization processing. The optimization method of influencing parameters is briefly described below:

(1) Initialize the parameters: total number of frogs N_{pop}, number of subgroups N_m, number of each group frogs N_f, maximal number of iterations M, random initialization of frog individuals, and initialize the population.

(2) Implement VMD and obtain a set of IMFs.

(3) Construct the global fitness function based on the envelope entropy, the kurtosis, and the correlation coefficients.

(4) Calculate the fitness value of each frog.

(5) Rank the frogs according to their fitness values.

(6) Divided the sorted frogs N_{pop} into N_m subgroups according to the descending order of the objective function. The first frog goes to the first memeplex, the second frog goes to the second memeplex, frog m goes to the m-th memeplex, and frog $m + 1$goes to the first memeplex.

(7) Determine the best individual of the subgroup p_b, the worst individual p_w, and the optimal solutions in the population S_1; the worst solution is improved by equation (16) in evolutional iteration M.

(8) Update the worst individual and descend the order to the individual to form a new group.

(9) Judge whether the algorithm satisfies the terminating condition and outputs the optimum solution when the algorithm satisfies the termination condition and otherwise moves on to step 6.

5. The Fault Feature Extraction by Parameter-Optimized VMD

The periodic impact energy caused by the failure of the rolling element bearing is weak, and it is relatively difficult to extract fault features due to the effects of noise and signal attenuation. When there is a fault for the rolling bearing, the useful characteristic components usually have very little energy, and it is submerged by the background noise. It is difficult to extract useful fault features. In order to extract the fault feature effectively and realize the fault diagnosis, the parameter-optimized VMD is presented to extract the useful fault features for the fault diagnosis of the rolling element bearing. The vibration signal is decomposed into a series of intrinsic mode functions by the parameter-optimized VMD algorithm. The envelope spectrum technique is utilized to analyze the best signal component. The fault features of the rolling bearing would be easily detected and extracted. The fault features extraction procedure of the parameter-optimized VMD method is briefly described as follows:

(1) Initialize population and parameters: the numbers of subgroup N_m, the numbers of each group frogs N_f, the numbers of iteration within a group N_e, and the numbers of evolutional iteration M.

(2) Optimize VMD parameters by applying SFLA and obtain global optimal parameters k and α.

(3) Decompose the original vibration signal into a set of the IMFs by the improved VMD.

(4) Calculate the envelope entropy, kurtosis, and correlation coefficients of all IMF components.

(5) Select the reconstructed IMF component with the largest kurtosis value, the highest correlation, and the smallest envelope entropy as the optimal component.

(6) Implement the spectrum analysis and compare the fault feature frequency in the envelope spectrum with the theoretical value of the bearing fault and determine the fault.

6. Experimental Results and Analysis

6.1. Simulation Analysis Using the Parameter-Optimized VMD. To quantitatively evaluate the effectiveness of the parameters-optimized VMD method, the simulation signal of rolling element bearings is constructed because the faults of rolling element bearings produce a series of shocks. Therefore, the simulation signal is mainly composed of the impact signal and noise signal generated by a bearing fault. The signal is sampled at 12 kHz. The simulated fault frequency f_i is set to 80 Hz. The resonance frequency f_n is set to 3 kHz. The rotating frequency f_r is 20 Hz. The simulated signal is expressed as follows:

$$\left.\begin{aligned} x(t) &= s(t) + n(t) = \sum_i A_i h(t - iT - \tau_i) + n(t) \\ A_i &= 1 + A_0 \sin(2\pi f_r t) \\ h(t) &= e^{-Ct} \sin(2\pi f_n t) \end{aligned}\right\}, \quad (18)$$

where $h(t)$ is the generated waveform of a single impact; A_i is the amplitude of the i-th impact force and considers possible periodic modulations, and $n(t)$ is the white Gaussian white noise; signal-to-noise ratio R_{SNR} is -1 dB; T is the mean spacing among impacts; the attenuation coefficient C is 700; and τ_i is an independent and identically distributed random variable.

The time domain plot and the envelope spectrum of the simulated signal are shown in Figure 4. It can be seen form Figure 4 that the impact signal is submerged in the strong background noise. The resonance frequency band in the spectrum is also not obvious. The frequency and period of the signal cannot be found, and the characteristic frequency of the fault signal cannot be accurately found from the envelope spectrum.

The simulated signal is decomposed by EEMD, and the corresponding frequency spectrum is shown in Figure 5. Obviously, the useful frequency components could not be distinguished from the decomposed IMFs, and they are contaminated with noise. Many parts of IMF 1 are replaced by the intermittent pulse signal. The replaced parts of IMF 1 are shifted to IMF 2 resulting in the phenomenon of mode mixing in the second and the following IMFs. In addition, as noted in Figure 4, the first three IMFs provided more information than the other IMFs, and the rest of the IMFs contain many redundant low-frequency components. In other words, the first three IMFs could be regarded as valid components of the signal, while the other IMFs were the low-frequency pseudocomponents that can mislead the analysis of the signal. The optimal mode component corresponding to the EEMD is IMF1. The envelope spectrum of the optimal mode component decomposed by the EEMD method is shown in Figure 6. It can be seen from Figure 6 that the impact characteristics associated with faults could not be identified, and the feature frequency of fault signals could not be extracted.

The VMD method is used to decompose the simulated signal, and the decomposed simulated signal has 5 mode components. The waveform and the corresponding frequency spectrum are shown in Figure 7. From the decomposition results, the VMD method can realize the adaptive segmentation of each component in the frequency domain, effectively overcome the mode aliasing phenomenon in EEMD, and has stronger noise robustness and weaker end effect than EEMD. The mode component corresponding to the minimum envelope entropy is mode 4, which is selected as the best component, and the envelope analysis is further done. The envelope spectrum of the signal is shown in Figure 8. It can be seen that the characteristic frequency of the fault signal cannot be accurately extracted by the original VMD method.

The parameter-optimized VMD method is implemented to analyze the simulation signal. The decomposition results and corresponding frequency spectrum of modes are shown in Figure 9. According to the decomposition results, the mode component corresponding to the smallest envelope entropy is IMF2, the mode component corresponding to the largest kurtosis is IMF5, and the mode component corresponding to the largest correlation is IMF5. The three mode components are reconstructed and used as the optimal component. The envelope spectrum of the reconstructed signal is shown in Figure 10. It can be seen that the spectral amplitude is prominent at the characteristic frequency 80 Hz, and the corresponding frequency doubling can also be obtained, which means that the parameter-optimized VMD can effectively decompose the fault signal and accurately extract the characteristic frequency of the fault signal.

6.2. Actual Vibration Signal Analysis. To further verify the effectiveness of the proposed parameter-optimized VMD method, the fault feature extraction of the actual experiment is implemented. The vibration data of rolling bearings are provided by Case Western Reserve University bearing data center [32]. The test stand consists of a 2 hp, three-phase induction motor, a torque transducer/encoder, and a dynamometer. The test bearings support the motor shaft at the drive end. Single point faults were introduced to the test bearings. The deep groove ball bearing with the type of 6205-2RS JEM SKF was used in the test. The locations of fault cover inner raceway, outer raceway, and rolling element. The tests are carried out under the four different motor loads with the motor speed. The vibration data were acquired at the sampling frequency of 12 kHz by using the

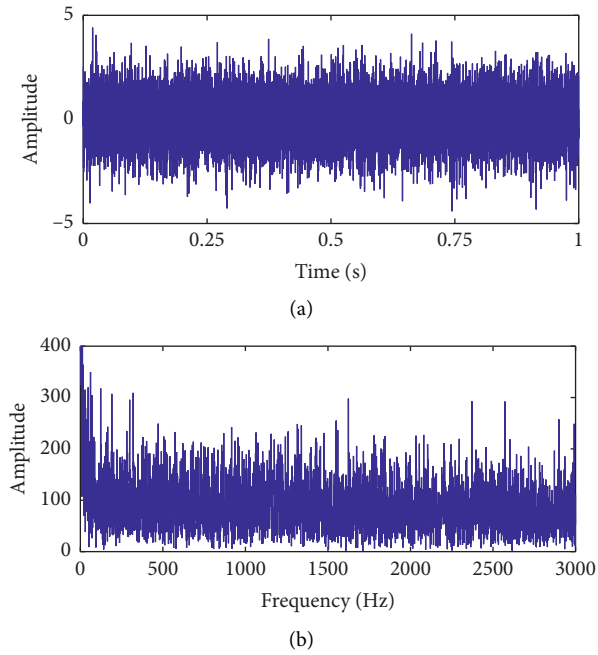


Figure 4: The time domain plot of the simulated signal and envelope spectrum.

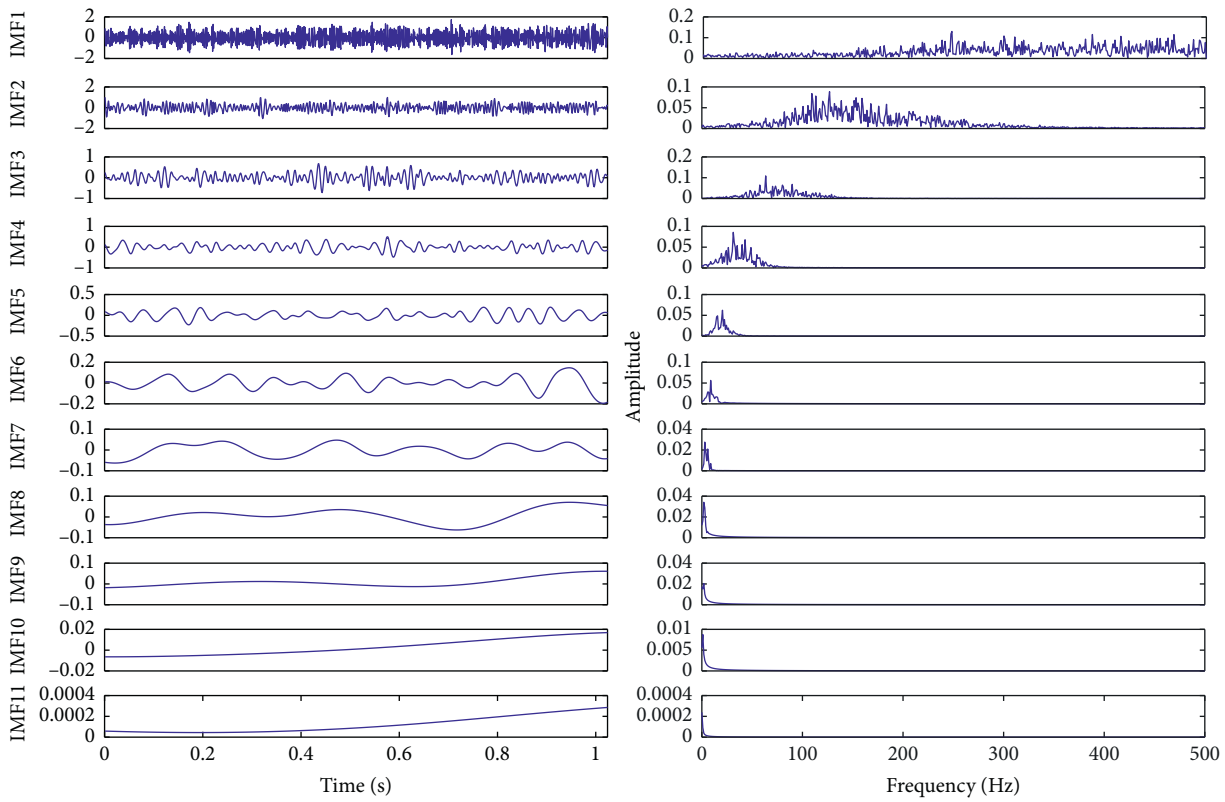


Figure 5: The decomposition results and corresponding frequency spectrum of IMFs with EEMD.

accelerometers, which are mounted at the drive end of the motor. The vibration signals of outer race defect with the motor load 0 hp and the fault diameters 7 mills are chosen to extract the fault feature. The characteristic frequency of the outer race defect signal is calculated to be at 107.37 Hz.

The time domain plot of the fault signal with outer race is shown in Figure 11, and Figure 12 shows the decomposition results of EEMD and corresponding the demodulated spectrum of IMFs. The first IMF component decomposed by EEMD contains abundant fault feature information and is

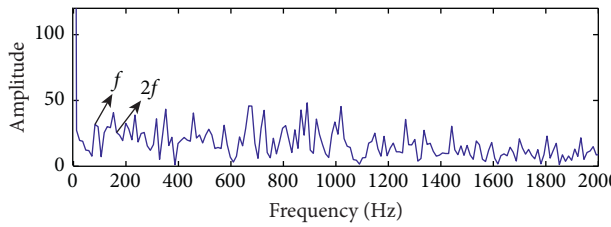


Figure 6: The envelope spectrum of optimal component by using EEMD.

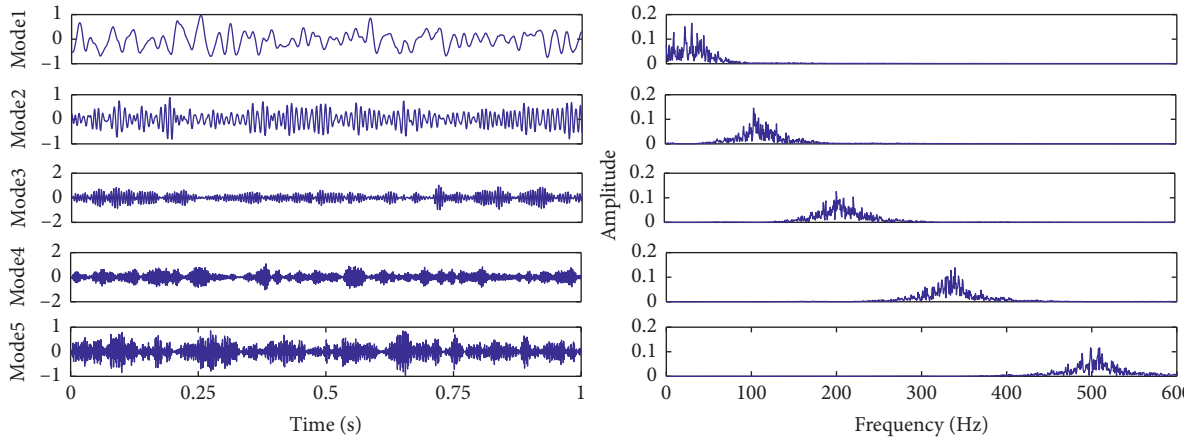


Figure 7: The decomposition results and corresponding frequency spectrum of modes with VMD.

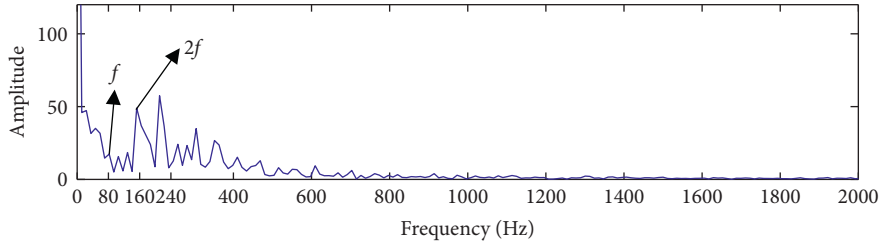


Figure 8: The envelope spectrum of optimal component by using VMD.

selected as the optimal feature component. The envelope spectrum of optimal component is shown in Figure 13. It can be seen from Figure 13 that the fault-rated impact features can be perceived from the time-frequency maps of the signals, and the characteristic frequency of the fault signal can be extracted. However, there are still many redundant components, and considerable background noise is also present in the figures, which smears the fault features and consequently leads to the risk of either false alarm or the failure of fault detection.

The VMD method is used to decompose the outer ring defect signal. The decomposition results and the corresponding frequency spectrum of IMFs are shown in Figure 14. According to the calculation results of decomposition, the mode component corresponding to the smallest envelope entropy is IMF6, the mode component corresponding to the largest kurtosis is IMF6, and the mode component corresponding to the largest correlation is IMF3. The three mode components are reconstructed and used as the optimal component. The envelope spectrum of the signal is shown in Figure 15. It can be seen from Figure 15 that when the default values of the mode number and bandwidth control parameter were adopted, the fault-rated impact features can be perceived from the time-frequency maps of the signals, and the characteristic frequency of the fault signal can be extracted. Compared with the EEMD method, the traditional VMD method has more superiorities than the EEMD method in the noise robustness and the elimination of mode aliasing. However, compared with EEMD, the fault feature extracted from the optimal mode component reconstructed by decomposing the signal with the default value of the influencing parameters is not good.

The proposed parameter-optimized VMD method is utilized to analyze the practical bearing vibration signal. The decomposition results and corresponding frequency spectrum of modes are shown in Figure 16. According to the calculation results of decomposition, the mode

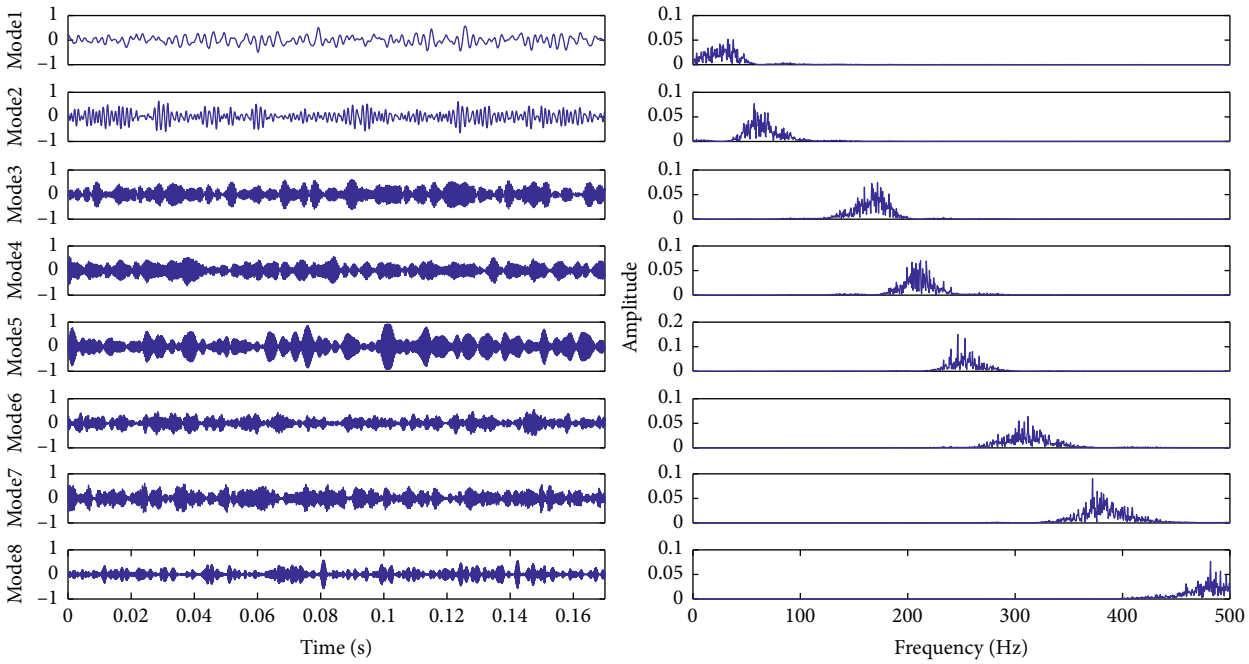


Figure 9: The decomposition results of parameter-optimized VMD and corresponding frequency spectrum of modes.

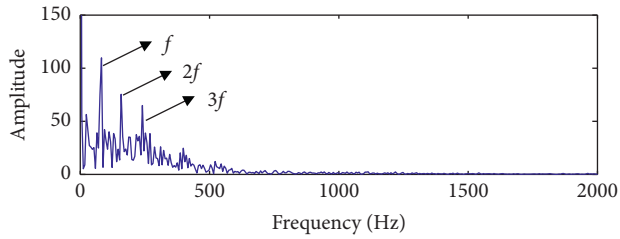


Figure 10: The envelope spectrum of optimal component by using the parameter-optimized VMD.

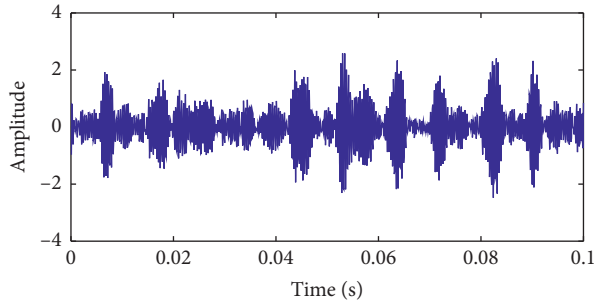


Figure 11: The time domain plot of the signal with outer race defect.

component corresponding to the smallest envelope entropy is IMF4, the mode component corresponding to the largest kurtosis is IMF5, and the mode component corresponding to the largest correlation is IMF3. The envelope spectrum of the reconstruction signal is shown in Figure 17. It can be seen that the spectral amplitude is prominent at the characteristic frequency 107.37 Hz, which means that the parameter-optimized VMD can correctly decompose the fault signal and accurately extract the characteristic frequency of the fault signal. Compared with the VMD method without the optimization and the EEMD method, the fault frequency extracted by the proposed method is more prominent and the noise is also suppressed.

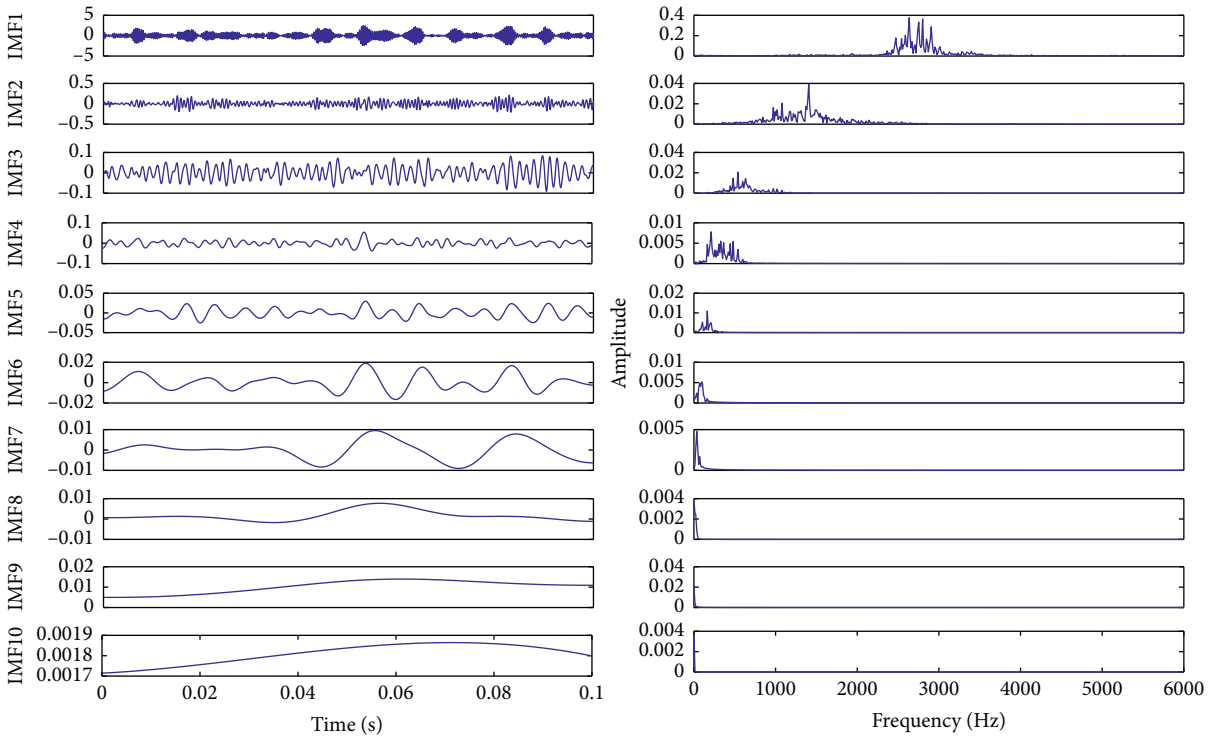


Figure 12: The decomposition results of EEMD and corresponding frequency spectrum of IMFs for outer ring defect.

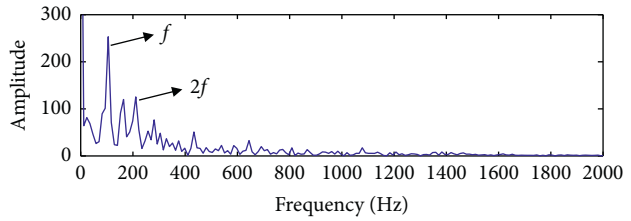


Figure 13: The envelope spectrum of the optimal component for outer ring defect by using EEMD.

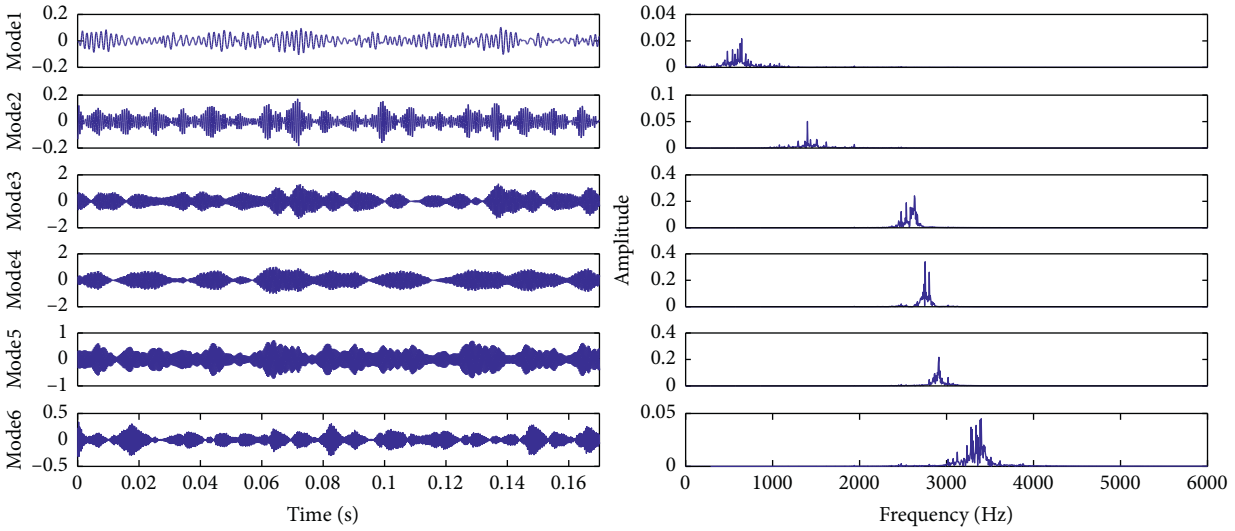


Figure 14: The decomposition results and corresponding frequency spectrum of modes for outer ring defect with VMD.

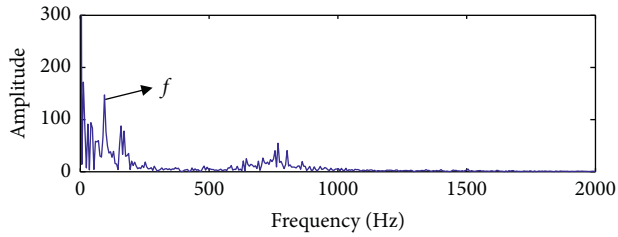


Figure 15: The envelope spectrum of optimal component for outer ring defect by using VMD.

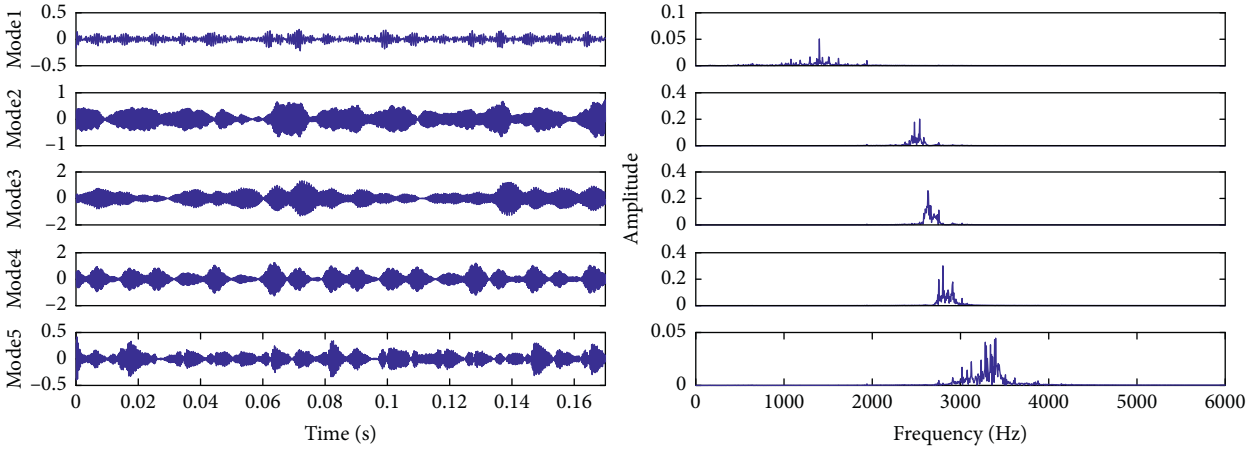


Figure 16: The decomposition results and corresponding frequency spectrum of modes by using the parameter-optimized VMD.

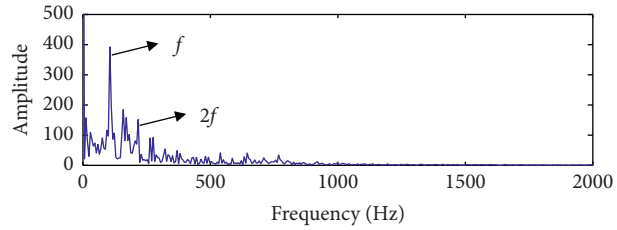


Figure 17: The envelope spectrum of the optimal component by using the parameter-optimized VMD.

7. Conclusion

The completely nonrecursive signal modal variational nature of the VMD method makes it more advantageous than EMD in terms of robustness against noise, overcoming end effects, and mode aliasing. This paper analyzes these three aspects. The decomposition accuracy of VMD method is affected by the choice of mode number k and bandwidth control parameter α. The parameter-optimized variational mode decomposition is developed to achieve the accurate decomposition of fault signal and adaptive control of influence parameters. Shuffled frog leaping algorithm is used to implement the optimization the influence parameters. The multiobjective evaluation function is constructed to select the optimal mode component. The envelope spectrum technique is used to analyze the optimal mode component. According to the characteristics of the vibration signal, we build the simulation signal to verify the feasibility and effectiveness of the signal and also use the vibration data of Western Reserve University to verify the proposed method. The experimental results show that the proposed parameter-optimized VMD method can correctly decompose the fault signal and accurately extract the characteristic frequency of the fault signal. Compared with the VMD method without the optimization and the EEMD method, the fault frequency extracted by the proposed method is more prominent and the noise is also suppressed. The proposed method also provides a new way to solve the problem for the analysis of vibration signal.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors wish to extend their sincere thanks for the support from the Beijing Municipal Science & Technology Commission of China. This work was supported by the Natural Science Foundation of Beijing Municipality, China (Grant no. 3212032).

References

[1] H. Zhao, S. Zuo, M. Hou et al., "A novel adaptive signal processing method based on enhanced empirical wavelet transform technology," *Sensors*, vol. 18, no. 10, Article ID 3323, 2020.

[2] H. Zhao, H. Liu, J. Xu, and W. Deng, "Performance prediction using high-order differential mathematical morphology gradient spectrum entropy and extreme learning machine," *Institute of Electrical and Electronics Engineers Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4165–4172, 2019.

[3] X. Y. Zhang and J. Z. Zhou, "Multi-fault diagnosis for rolling element bearings based on ensemble empirical mode decomposition and optimized support vector machines," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 127–140, 2013.

[4] J. Qu, Z. Zhang, and T. Gong, "A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion," *Neurocomputing*, vol. 171, pp. 837–853, 2016.

[5] Q. Tong, Z. Sun, Z. Nie, Y. Lin, and J. Cao, "Sparse decomposition based on ADMM dictionary learning for fault feature extraction of rolling element bearing," *Journal of Vibroengineering*, vol. 18, no. 8, pp. 5204–5216, 2016.

[6] F. Jiang, Z. Zhu, W. Li, G. Chen, and G. Zhou, "Robust condition monitoring and fault diagnosis of rolling element bearings using improved EEMD and statistical features," *Measurement Science & Technology*, vol. 25, no. 2, pp. 1–14, 2014.

[7] Y. Lei, J. Lin, Z. He, and Y. Zi, "Application of an improved kurtogram method for fault diagnosis of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 25, no. 5, pp. 1738–1749, 2011.

[8] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *Institute of Electrical and Electronics Engineers Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7319–7327, 2020.

[9] W. Deng, J. Xu, Y. Song, and H. Zhao, "Differential evolution algorithm with wavelet basis function and optimal mutation strategy for complex optimization problem," *Applied Soft Computing*, vol. 10, Article ID 106724, 2020.

[10] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics-A tutorial," *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, 2011.

[11] G. Dong and J. Chen, "Noise resistant time frequency analysis and application in fault diagnosis of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 33, pp. 212–236, 2012.

[12] S. Abbas, A. Rafsanjani, A. Farshidianfar, and N. Irani, "Rolling element bearings multi-fault classification based on the wavelet denoising and support vector machine," *Mechanical Systems and Signal Processing*, vol. 21, no. 7, pp. 2933–2945, 2007.

[13] W. J. Wang and P. D. Mcfadden, "Application of wavelets to gearbox vibration signals for fault detection," *Journal of Sound and Vibration*, vol. 192, no. 5, pp. 927–939, 1996.

[14] Z. K. Peng and F. L. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography," *Mechanical Systems and Signal Processing*, vol. 18, no. 2, pp. 199–221, 2004.

[15] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A-Mathematical Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[16] J. Cheng, D. Yu, J. Tang, and Y. Yang, "Application of SVM and SVD technique based on EMD to the fault diagnosis of the rotating machinery," *Shock and Vibration*, vol. 16, no. 1, pp. 89–98, 2009.

[17] V. K. Rai and A. R. Mohanty, "Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert-Huang transform," *Mechanical Systems and Signal Processing*, vol. 21, no. 6, pp. 2607–2615, 2007.

[18] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 01, no. 01, pp. 1–41, 2009.

[19] J. S. Smith, "The local mean decomposition and its application to EEG perception data," *Journal of the Royal Society Interface*, vol. 2, no. 5, pp. 443–454, 2005.

[20] B. J. Chen, Z. J. He, X. F. Chen et al., "A demodulating approach based on local mean decomposition and its applications in mechanical fault diagnosis," *Measurement Science and Technology*, vol. 22, no. 5, Article ID 055704, 2011.

[21] C. Park, D. Looney, M. M. Van Hulle, and D. P. Mandic, "The complex local mean decomposition," *Neurocomputing*, vol. 74, no. 6, pp. 867–875, 2011.

[22] W. Y. Liu, W. H. Zhang, J. G. Han, and G. F. Wang, "A new wind turbine fault diagnosis method based on the local mean decomposition," *Renewable Energy*, vol. 48, pp. 411–415, 2012.

[23] M. G. Frei and I. Osorio, "Intrinsic time-scale decomposition: time-frequency-energy analysis and real-time filtering of non-stationary signals," *Proceedings of the Royal Society A-Mathematical Physical and Engineering Sciences*, vol. 463, no. 2078, pp. 321–342, 2007.

[24] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *Institute of Electrical and Electronics Engineers Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[25] Y. Wang, R. Markert, J. Xiang, and W. Zheng, "Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system," *Mechanical Systems and Signal Processing*, vol. 60-61, pp. 243–251, 2015.

[26] Y. Liu, G. Yang, M. Li, and H. Yin, "Variational mode decomposition denoising combined the detrended fluctuation analysis," *Signal Processing*, vol. 125, pp. 349–364, 2016.

[27] M. Eusuff, K. Lansey, and F. Pasha, "Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization," *Engineering Optimization*, vol. 38, no. 2, pp. 129–154, 2006.

[28] M. M. Eusuff and K. E. Lansey, "Optimization of water distribution network design using the Shuffled Frog Leaping Algorithm," *Journal of Water Resources Planning and Management*, vol. 129, no. 3, pp. 210–225, 2003.

[29] R. Chen, S.-K. Guo, X.-Z. Wang, and T.-L. Zhang, "Fusion of multi-RSMOTE with fuzzy integral to classify bug reports with an imbalanced distribution," *Institute of Electrical and*

Electronics Engineers Transactions on Fuzzy Systems, vol. 27, no. 12, pp. 2406–2420, 2019.

[30] B. Amiri, M. Fathian, and A. Maroosi, "Application of shuffled frog-leaping algorithm on clustering," *The International Journal of Advanced Manufacturing Technology*, vol. 45, no. 1-2, pp. 199–209, 2009.

[31] J. Sun, Q. Xiao, J. Wen, and F. Wang, "Natural gas pipeline small leakage feature extraction and recognition based on LMD envelope spectrum entropy and SVM," *Measurement*, vol. 55, pp. 434–443, 2014.

[32] Seeded Fault Test Data from Bearing Data Center of Case Western Reserve University, 2016, http://csegroups.case.edu/bearingdatacenter.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6687195, 12 pages
https://doi.org/10.1155/2021/6687195

Research Article

Adaptive Extraction Method Based on Time-Frequency Images for Fault Diagnosis in Rolling Bearings of Motor

Yunchao Ma,[1] **Chengdong Wang**,[1,2] **Dongchen Yang**,[1] **and Cheng Wang**[1]

[1]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*Institute of Electric Vehicle Driving System and Safety Technology, University of Electronic Science and Technology of China, Chengdu 611731, China*

Correspondence should be addressed to Chengdong Wang; wangchengdong@uestc.edu.cn

Received 6 November 2020; Revised 7 December 2020; Accepted 28 January 2021; Published 15 February 2021

Academic Editor: Fazal Mahomed

In order to diagnose the faults of rolling bearings in motors via time-frequency analysis of bearing vibration signals quickly, this paper puts forward a method of extracting the main components from time-frequency images. A threshold is adaptively determined based on the gray histogram feature of the time-frequency images obtained from the vibration signals of the motor rolling bearings. Then, a mask template is generated by the threshold and a binarization processing. Based on a multiplication operation between the mask template and the original time-frequency image, the signal component with low energy in the time-frequency image is filtered out, and only the main components with high energy is remained for fault diagnosis, which is convenient for the subsequent identification of the faults for motor rolling bearings. The main components in the time-frequency images can be retained adaptively with the thresholds determined by the time-frequency images themselves.

1. Introduction

Condition monitoring and fault diagnosis for equipment can monitor the health status of equipment in real time and determine the fault location and severity by the changes of some signals, which can not only avoid the occurrence of major accidents but also greatly save maintenance costs. While a motor is working, factors such as overload impact, assembly error, poor lubrication, or impurity doping will lead to the failure of the bearing. The vibration signals of a motor will show the unsteady characteristic, and then, the nonstationary signals have the characteristics of limited duration and timely variation. The traditional signal processing methods are mostly based on the assumption of a stable state, which can only analyze the statistical characteristics of the signal in the time domain or frequency domain, but are unable to reveal the instantaneous characteristics in the joint time-frequency domain. The time-frequency representation of a signal can describe the energy distribution and time-varying characteristics in the time-frequency domain, which is the most complete expression method for unstable signals. Along with the development of image recognition, some mechanical fault identification methods are put forward based on the time-frequency image texture, shape, and other visual feature extraction. These methods can not only help us to understand the images but also are good to improve the recognition accuracy.

Many scholars have studied this problem. Hongkun et al. [1] make an investigation of the rolling bearing faults' diagnosis by a time-frequency image processing technology, and the experiment results showed that the Hough transform of time-frequency images can effectively classify the faults of rolling bearings. Isobe et al. [2] combined the local wave time-frequency spectrum with image processing to extract the features from vibration signals of reciprocating machines. Cai et al. [3] calculated the Wigner–Ville distributions of acceleration signals by time-frequency analysis, obtained a series of time-frequency gray images from the above distributions by image processing, and then obtained a group of fractal texture characteristic parameters from these gray images to identify the abnormal status of a diesel engine valve gap. Wei and Zhan-Sheng [4] studied a

diagnosis method that is based on gray level-gradient co-occurrence matrix, by extracting the information of image texture characteristic to conduct the fault diagnosis of a rotating machine. Cai et al. [5] proposed a new fault diagnosis method based on the time-frequency image recognition of EMD-WVD vibration spectrums by SVM. Through extracting the moment invariant feature of the images, the diagnosis eigenvectors were achieved, and their modes were recognized by an improved binary tree classifier. Verstraete at al. [6] proposed a deep learning enabled featureless method, where the images generated by time-frequency representations of the raw data were fed into a deep convolutional neural network (CNN) architecture for classification and fault diagnosis, and the results are good.

In time-frequency images, important information is expressed through time-frequency components with high energy. Therefore, when the distribution law of frequency components in time-frequency images is studied, the time-frequency components with low energy can be regarded as noise and be filtered out, which will help us to pay attention to the time-frequency components with high energy. n is the time-frequency images, and the energy of the time-frequency component is reflected with the gray value of image, so the classification of images can be achieved based on the important components. A noise removal method for time-frequency images is studied in this paper. A binarization processing is applied to the time-frequency images to get a mask template with which the original images are overlapped to highlight the components with concentrated energy. Then, the fault diagnosis can be carried out according to the remained signal components.

The remaining sections of this paper are arranged as follows. Firstly, the method of extracting the main components of time-frequency images is introduced in Section 2. Then, the OTSU method, the KSW-Entropy method, and our improved method based on OTSU and KSW-entropy methods are introduced in Section 3. The comparison of the results and the analysis of the experimental data are described in Section 4. The summary of our results is given in Section 5.

2. The Method of Extracting the Main Components from Time-Frequency Images

In fault diagnosis for mechanical equipment, especially in the treatment of nonstationary signals, we mainly focus on the changes of the main signal components which will greatly affect or even determine the characteristic of the whole signal. We often want to know how the frequency of a signal component is changed with time and how the energy of a signal component is changed with time. By the methods of time-frequency analysis, we can see the changes of signal components. There are many methods of time-frequency analysis, such as Wigner–Ville distributions (WVD) [7–9], short-time Fourier transform (STFT) [10–12], wavelet transform [13–15], and Hilbert–Huang transform [16–18]. Among these methods, the short-time Fourier transform is simple and can be worked out quickly, while giving the main information of how the signal component is changing with time. Although the time-frequency resolution of STFT is not as high as that of WVD, STFT is widely used because of its free of crossterms, which limits the application of WVD largely. In order to show the results of time-frequency analysis visually, images are usually used, where the time is expressed in horizontal coordinate and the frequency is expressed in a vertical coordinate. In this paper, STFT is used to get the time-frequency images of motor bearings.

2.1. Short-Time Fourier Transform (STFT) and Time-Frequency Images. The STFT is a popular method for analyzing nonstationary signals, which is a transform of traditional Fourier transform [19]. The basic idea of STFT is as follows.

When a short-time window function is applied to an original signal, the original nonstationary signal can be viewed as a stationary signal during the very short interval of the window. The window function $\omega(t)$ is then moved so that $x(\tau)\omega_{t,f}(\tau - t)$ can be always considered as a stationary signal for a continuous finite time length. Then, the power spectrum of the signal at different time periods can be calculated. The STFT of the signal $x(t)$ is defined as

$$F_x^w(t, f) = \int x(\tau)\omega_{t,f}(\tau - t)e^{-j2\pi f\tau}\mathrm{d}\tau, \tag{1}$$

where $x(\tau)$ is the signal to be analyzed, $\omega(\tau)$ is the sliding window function, and $F_x^w(t, f)$ is the spectral distribution of signal x at time t.

The discrete STFT is defined as

$$F_x^\omega(m, n) = \sum_{k=ms}^{ms+N-1} x(k)\omega(k - ms)e^{-j2\pi kn/N}, \quad (m = 0, 1, 2 \ldots M - 1;\ n = 0, 1, 2 \ldots N - 1), \tag{2}$$

where $\omega(k)$ is the window function with the length of N, the sliding step of the window function is s sampling time interval, m is the location of the window, corresponding to the time parameter of STFT, and n is the frequency parameter. Suppose the sampling frequency of the original signal $x(k)$ is f_s; then, the sampling time interval is $T_s = 1/f_s$. $F_x^\omega(m, n)$ is the spectrum of the signal at the time of msT_s, where the frequency parameter of n corresponds to nf_s/N.

By using STFT, we can get the power spectrum of the signal at different time. Then, we show the results of STFT in time-frequency images with the horizontal axis as time and the vertical axis as frequency and the amplitude of the STFT as the gray value. In order to observe the energy distribution in time-frequency images, this paper inverts the gray scale of time-frequency images, that is, at a certain moment and a certain frequency, the larger the energy is, the smaller the gray value will be.

2.2. Extraction of the Main Components from Time-Frequency Images. A time-frequency image can be regarded as an ordinary two-dimensional image, where the time is expressed in horizontal coordinate and the frequency is expressed in vertical coordinate. And, the energy of every time-frequency component is reflected with the gray value.

In the process of bearing faults' diagnosis, the classification of important feature components can be achieved based on the classification of the gray value of the image. That is to say, the features of faults are largely contained in the main components whose energy is expressed with large gray values in the time-frequency image. So, our attention can only focus on the parts with large gray values in the time-frequency image.

In this paper, an adaptive method of extracting the main components of time-frequency images is presented. Firstly, STFT is used to get the time-frequency image of vibration signals. Then, a suitable threshold is calculated according to the time-frequency image based on the methods of OTSU and KSW-entropy. Then, a mask template is generated according to the threshold with the same size as the original image. The value of each pixel is 0 or 1, where 1 means the pixels will be kept and 0 means the pixels will be removed. Then, the time-frequency image which only retains the main components is obtained by a multiplication between the mask template and the original time-frequency image. Finally, the fault diagnosis is carried out based on the time-frequency image with only main components. The recognition computation of the time-frequency image with only the important fault feature information retained will be much smaller than that of the original time-frequency image.

The process of the main components extraction method is shown in Figure 1.

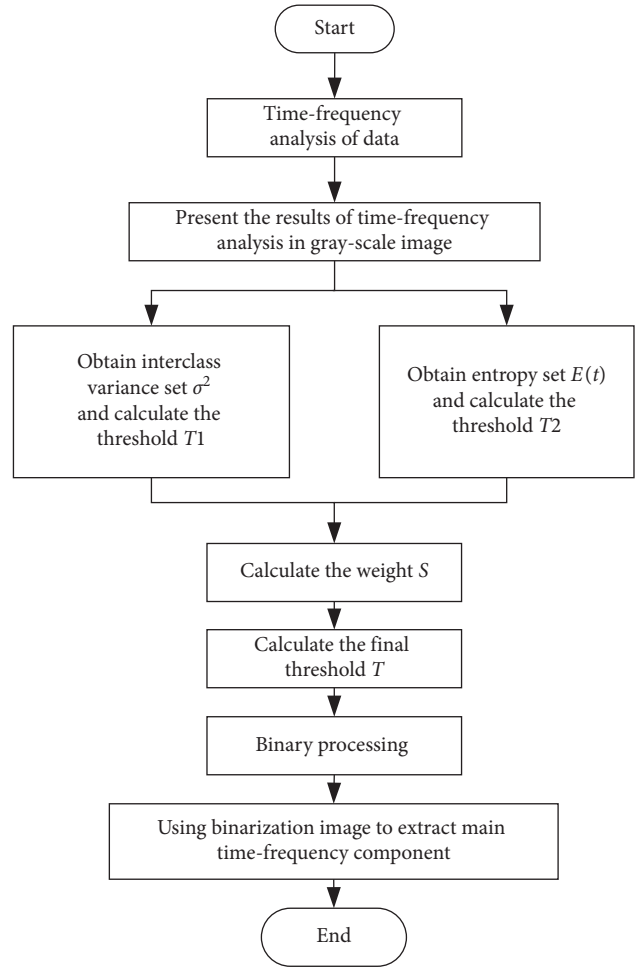


Figure 1: Algorithm flowchart of this paper.

3. The Adaptive Methods of Threshold Selection

The key of our method is the threshold selection of image binarization, which also means the selection of the energy threshold. An appropriate energy threshold can extract the main characteristics components of a time-frequency image and filter out other weak signals or irrelevant features. Therefore, an improved adaptive threshold selection method is proposed based on the KSW-entropy algorithm and OTSU threshold segmentation algorithm.

3.1. Threshold Based on OTSU. Among all the algorithms related to image threshold, OTSU algorithm [20], proposed by OTSU, a Japanese scholar, is considered as the best algorithm for threshold selection in image segmentation. It divides the image into background and foreground according to its gray scale. As variance is a measure of gray distribution uniformity, the greater the interclass variance between the background and foreground, the greater the difference between the two parts of the image. If part of the foreground is misclassified into background or part of the background is misclassified into foreground, the difference between the two parts will decrease. Therefore, the segmentation that maximizes the variance between classes means that the probability of misclassification is minimized.

The principle of OTSU is as follows.

If a threshold value is set as t, then the image pixel can be divided into two categories of $C1$ (whose gray value lesser than t) and $C2$ (whose gray value greater than t). Assuming that the mean gray values of the two classes of pixel grayscale are μ_1 and the average gray value of the whole image is μ, the percentage of $C1$ to total pixels is ω_1, the percentage of $C2$ to total pixels is ω_2, the total number of pixels is $N \times M$, and the interclass variance is σ^2. Then, the formulas can be expressed as follows:

$$\omega_1 = \frac{C1}{M \times N}, \tag{3}$$

$$\omega_2 = \frac{C2}{M \times N}, \tag{4}$$

$$\omega_1 + \omega_2 = 1, \tag{5}$$

$$\mu = \omega_1 \times \mu_1 + \omega_2 \times \mu_2, \tag{6}$$

$$\sigma^2 = \omega_1 \times (\mu_1 - \mu)^2 + \omega_2 \times (\mu_2 - \mu)^2. \tag{7}$$

According to formulas (6) and (7), the final expression of interclass variance is

$$\sigma^2 = \omega_1 \times \omega_2 \times (\mu_1 - \mu_2)^2. \tag{8}$$

If the maximal image gray is L, by trying every gray value and calculating the interclass variance of $C1$ and $C2$ pixels of

the image, the best threshold T can be found with the biggest interclass variance:

$$\sigma^2(T) = \max\left\{\sigma^2(t) | 0 \le t \le L-1\right\}. \quad (9)$$

3.2. Threshold Based on KSW-Entropy. In 1985, Kapur, Shaoo, and Wong proposed a method to select threshold automatically based on optimal entropy, which was abbreviated as KSW-entropy algorithm [21]. The method applies the entropy of image information to image segmentation. For an image, a threshold value is found to divide the histogram into two categories, and the information entropy of the two categories is calculated, respectively. Based on the threshold, the entropy is maximum. Entropy is used in information theory to describe uncertain factors. The more ordered a system is, the lower its entropy is. In the image, the boundary distribution of the target is the most uncertain, so the boundary between the image target and the background has the maximum entropy. The KSW-entropy algorithm is good for image segmentation with fuzzy boundaries between the target and background.

For an image with a gray scale of L, assuming that $p_0, p_1, p_2, \ldots, p_{L-1}$ are the probability distribution of each gray level in the image. Image pixels are divided into two categories by the threshold t. The pixels whose gray values are in the range of $[0, t]$ are divided into $C1$ category and the pixels whose gray values are in the range of $[t+1, L-1]$ are divided into $C2$ category. Let $P_{C1} = \sum_{i=0}^{t} p_i$ be the sum of the probability of pixels in C1 and $P_{C2} = \sum_{i=t+1}^{L-1} p_i$ be the sum of the probability of pixels in $C2$, and $P_{C1} = 1 - P_{C2}$. The probability distribution of each pixel in $C1$ is p_0/P_{C1}, $p_1/P_{C1}, p_2/P_{C1}, \ldots, p_t/P_{C1}$, and the probability distribution of each pixel in $C2$ is $p_{t+1}/P_{C2}, p_{t+2}/P_{C2}, p_{t+3}/P_{C2}, \ldots, p_{L-1}/P_{C2}$. Then, the information entropy $E(C1)$ of $C1$ and entropy $E(C2)$ of $C2$ are calculated as follows:

$$E(C1) = -\sum_{i=0}^{t} \frac{p_i}{P_{C1}} \ln \frac{p_i}{P_{C1}}, \quad (10)$$

$$E(C2) = -\sum_{i=t+1}^{L-1} \frac{p_i}{P_{C2}} \ln \frac{p_i}{P_{C2}}. \quad (11)$$

The total information entropy is

$$E(t) = E(C1) + E(C2). \quad (12)$$

After traversing the whole gray levels of L, the threshold T that maximizes entropy E is the optimal segmentation threshold:

$$E(T) = \max\{E(t) | 0 \le t \le L-1\}. \quad (13)$$

3.3. Threshold Based on Combined OTSU and KSW-Entropy. The segmentation result of OTSU is not good for the image with blurred edges, which is mainly reflected in the misclassification of image edges and the sensitivity to noise. However, the edge part of images is processed better with KSW-entropy than with OTSU, but in the background part, where a wrong segmentation may be classified. So, we combine the methods of OTSU and KSW-entropy to propose an adaptive threshold segmentation method.

In order to satisfy formulas (9) and (13) simultaneously as far as possible, considering the theory of multiobjective programming, the linear weighting method in the evaluation function is used to reconstruct a function of threshold selection. Suppose the weight of interclass variance is S, $E_{\min}$ is the minimum entropy in the calculation process of the calculating, $E_{\max}$ is the maximum entropy, and $\text{norm}(\sigma^2)$ is to normalize the interclass variance of all calculated thresholds into $[E_{\min}, E_{\max}]$. Then, the mathematical model of our method can be expressed as follows:

$$E(T) = \max\left\{S \times \text{norm}\left(\sigma^2(t)\right) + (1-S) \times E(t) | 0 \le t \le L-1\right\}. \quad (14)$$

The weight S is calculated by the threshold $T1$ and the threshold $T2$ which are determined by OTSU and KSW-entropy. Considering OTSU's missing edge and KSW's excessive background, the best threshold should be positioned between the thresholds determined by the two methods. So, when the threshold value of image is decided, both the variance and entropy should be taken into consideration. At the same time, due to the effect of both methods, the value of the variance should be moved towards the direction of the maximum entropy, and the entropy value should be moved towards the direction of maximum variance, to achieve a balance of the effect of two methods. Therefore, the definition of S can be expressed as the following formula:

$$S = \frac{T2}{T1 + T2}, \quad (15)$$

with the weight S, the threshold of an image can be selected dynamically and adjustable. The classification between the edge and the background of a time-frequency image can be achieved by taking the maximum intercategory variance and the maximum entropy into consideration as far as possible.

4. Experimental Results and Analysis

4.1. Introduction to the Bearing Data. The experimental data we used were obtained from the Bearing Datasets of Case Western Reserve University (CWRU) [22–24]. The test rig consisted of a 2 horsepower (hp) motor driving a shaft mounted with a torque transducer and encoder. The torque is applied to the shaft by a dynamometer and a control system. The acceleration data of vibration was measured near to the motor bearings. The faults of the motor bearings were artificially seeded using electro-discharge machining (EDM). Faults ranging from 0.007 inches (or 7 mil) to 0.040 inches in diameter were introduced separately at the inner raceway, rolling element (i.e., ball), and outer raceway. Faulted bearings were reinstalled into the test motor and the vibration data was recorded for motor loads of 0 to 3

horsepower (the motor speeds ranged from 1720 rpm to 1779 rpm).

A vibration data of a faulty bearing we analyzed came from the dataset, where the fault size is 7 mil with zero loading, and the shaft rotation speed is 1797 rpm, and the sampling frequency is 12 KHz. In the process of STFT, a hamming window with the length of 63 is used, and the sliding step of the window is 1. Firstly, the results of normal bearing in the same situation are shown in Figures 2 and 3. Figure 2 is the time domain and frequency domain waveforms, and Figure 3 is the joint time-frequency distribution image. The waveforms of time and frequency are also shown in Figure 3, where the upper waveform is for time domain and the left waveform is for frequency domain. The joint time-frequency distribution image of STFT is shown in the right-bottom corner in Figure 3. Figure 4 shows the waveforms of a faulty bearing, respectively, in time and frequency domains where the inner ring is faulty in size of 7 mil. The joint time-frequency distribution image of the faulty bearing is shown in Figure 5 as the same manner in Figure 4. In the following parts, we only show the time-frequency images of STFT.

By comparing the time domain waveforms, the frequency domain waveforms, and time-frequency images of the normal bearing and the fault bearing, it can be seen that the waveforms are quite different if a bearing has fault or not. From Figure 2, we can see that the frequency of vibration signals of normal bearings is mainly concentrated near 160 Hz, 360 Hz, 1050 Hz, and 2100 Hz, among which the component near 1050 Hz has the largest energy. The signal component at 160 Hz has the second largest energy. We can only obtain this information from the spectrum diagram. However, it can be seen from the time-frequency image that the components near 1050 Hz do not always exist; these components appear at about 0.009 s, 0.046 s, and 0.079 s, respectively, and the duration is less than 0.01 s, as shown in Figure 3.

In the vibration signal of the faulty bearing, as shown in Figure 4, the signal components are particularly rich, mainly concentrating in the frequency band range between 2600 Hz and 2900 Hz and around 3900 Hz. From the time-frequency image as shown in Figure 5, we can see that even within these two frequency bands. The signal components appear intermittently and the durations of each component are slightly different. At the same time, we can also see that, in addition to these main components, there are also many components of weak energy distributed randomly in the time-frequency domain, which tend to disturb our attention due to their weak energy and random distribution. We hope to filter out these disturbances and then we can concentrate on finding the components that reflect the characteristics of the bearing failure.

4.2. Comparison of the Extraction Effects. The original time-frequency image of the faulty bearing data is shown in Figure 6. The mask template and the extracted main components by the threshold of OTSU are shown in Figures 7 and 8. The mask template and the extracted main components using KSW-entropy are also shown in Figures 9 and 10. The mask template and main components extracted by our method are shown in Figures 11 and 12. And the main components of the normal bearing extracted by our method are shown in Figure 13.

The threshold selected by our method is 192, which is between the threshold of 205 and 157, respectively, obtained by the methods of OTSU and KSW-entropy. Comparing the Figures 6, 8, 10, and 12, we can see that when the threshold value is different, the extracted main components are not exactly the same. The larger the threshold is, the less the time-frequency components are filtered out. The threshold calculated by the method of KSW-entropy is smaller than that of OTSU, so the main components extracted by the method of KSW-entropy are less than the main components extracted by the method of OTSU. The amount of the main components extracted will affect our judgment and ability to grasp the principal information of faulty bearings.

By comparing Figures 12 and 13, it can be seen that the main time-frequency components extracted from the time-frequency images of the faulty bearing and the normal bearing are greatly different. The fault of the bearing can be judged by observing the distribution of these major components. From the main time-frequency components extracted by our method, it can be easily seen that the signal components are mainly concentrated in the frequency bands around 1300 Hz, 2800 Hz, and 3600 Hz, as shown in Figure 12. These signal components do not appear continuously, but occur at regular intervals with slight changes in energy each time. For example, there are some obviously signal components occur at 0.0085 s, 0.0272 s, 0.0455 s, and 0.0642 s, as shown in Figure 12 with red dotted lines, and the time interval between these signal components is about 0.0185 s. Between each two components with obviously high energy, there are also two signal components with slightly lower energy. That is to say, a signal component is occurred at almost every 0.0066 s or so. From this time period we can see that the frequency of the components is about 152 Hz which is close to the characteristic frequency of inner bearing ring fault. Based on the analysis of the main components of the time-frequency image, we can roughly infer that there may be a fault in the inner bearing ring.

In order to make the results more convincing, another data is analyzed to verify our method. The data is recorded on the fault size of 21 mil, and the motor load is 2 horsepower with 1750 rpm, and the sampling frequency is also 12 kHz. The waveforms of time and frequency domain are as shown in Figure 14, where 4096 samples are analyzed. The

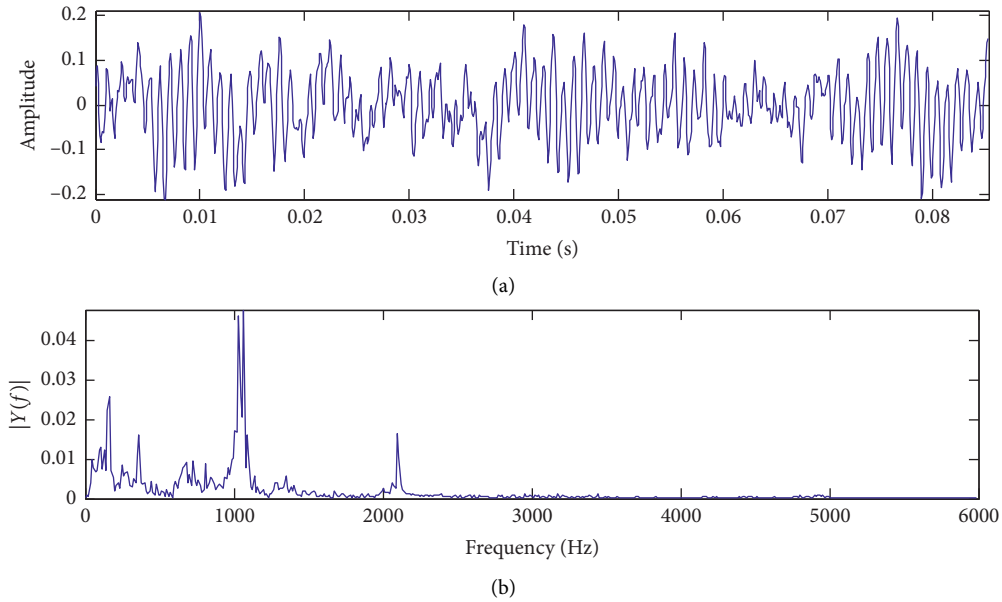


Figure 2: The waveforms of a normal bearing in time and frequency domain: (a) time domain of data Normal_0_x097_de_time_00001_01024 and (b) spectrum of data Normal_0_x097_de_time_00001_01024.

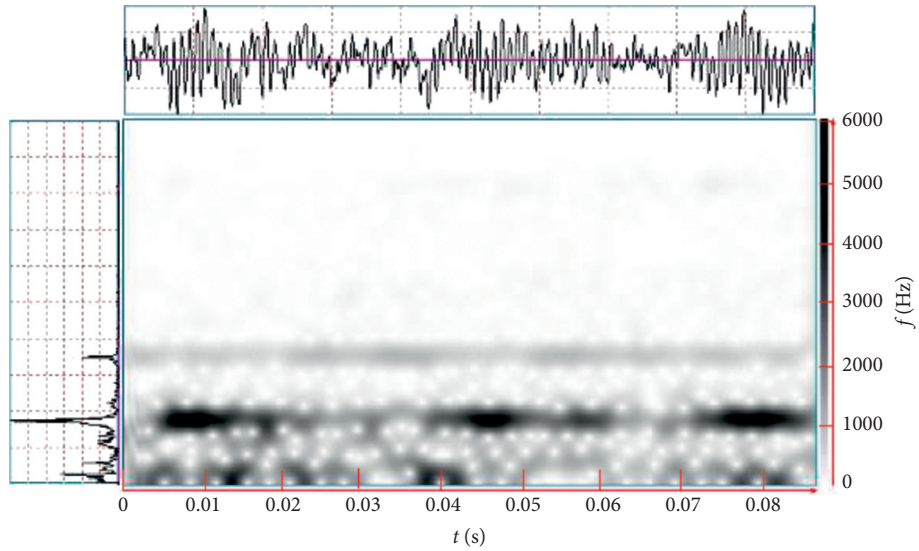


Figure 3: The joint time-frequency distribution of a normal bearing.

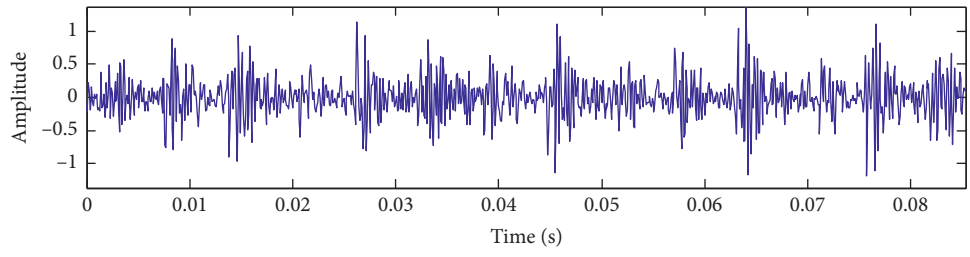
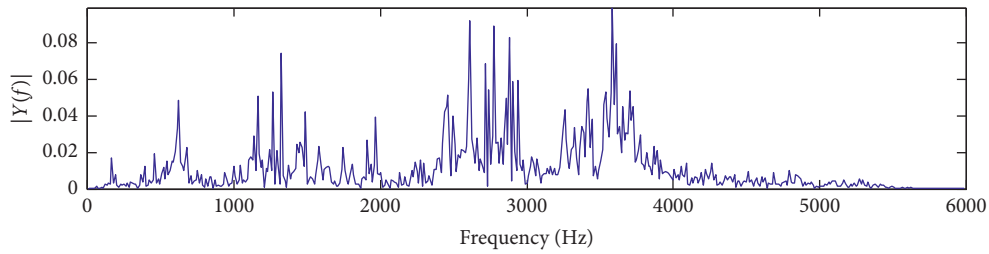


Figure 4: Continued.



(b)

Figure 4: The waveforms of a faulty bearing in time and frequency domain: (a) time domain of data IR007_0_X105_DE_time_00001_01024 and (b) spectrum of data IR007_0_X105_DE_time_00001_01024.

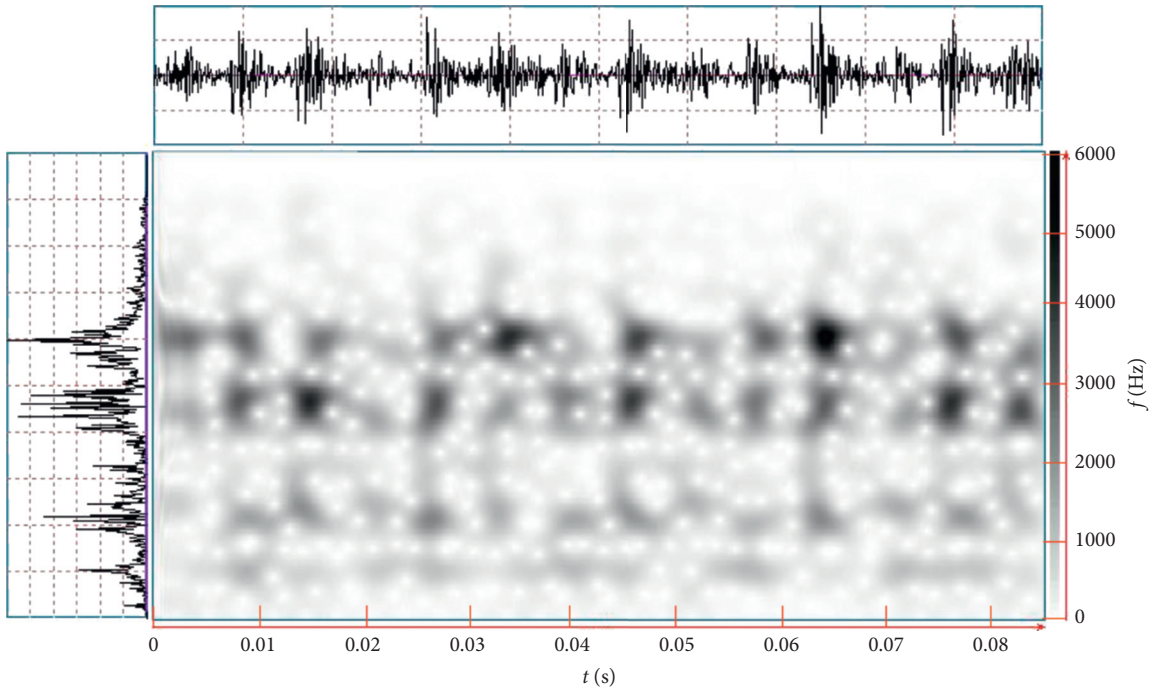


Figure 5: The joint time-frequency distribution of a faulty bearing.

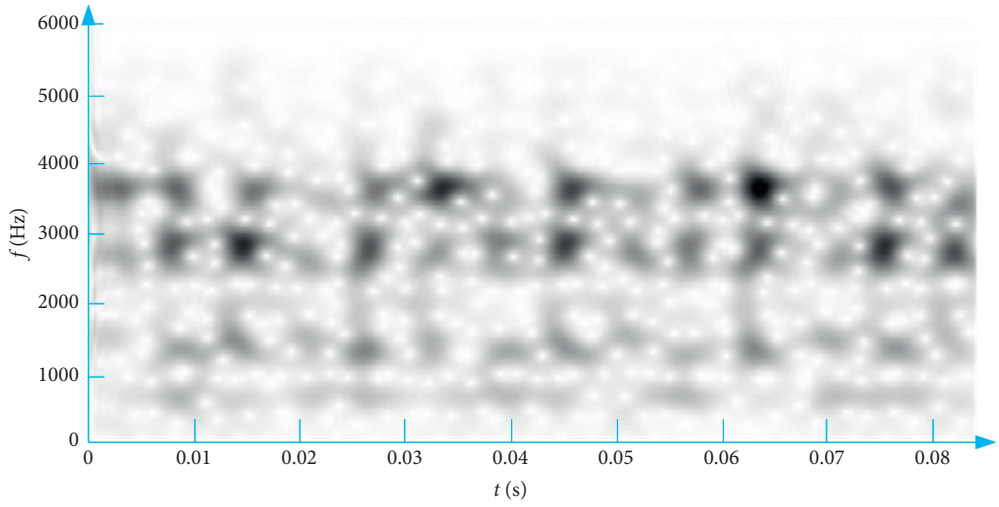


Figure 6: The original time-frequency image of a faulty bearing.

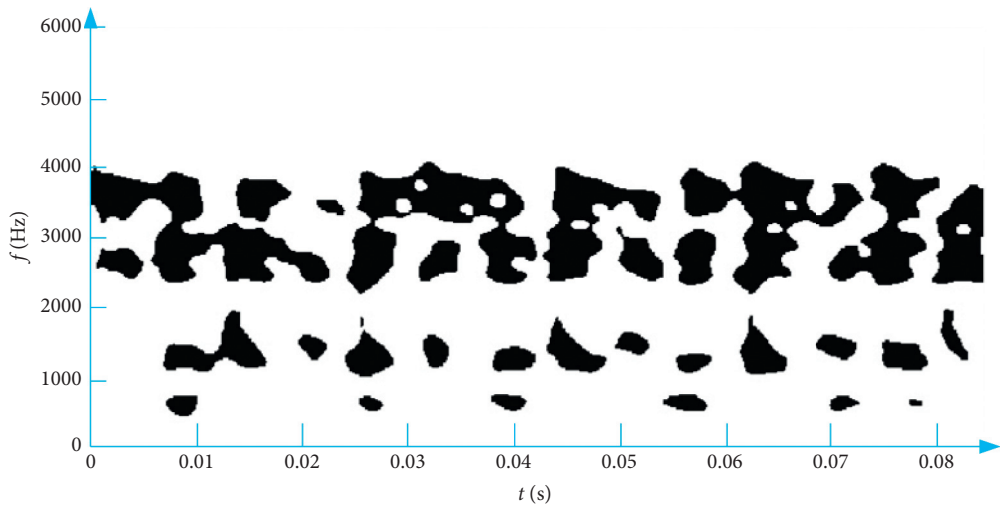


Figure 7: The mask template obtained with the method of OTSU.

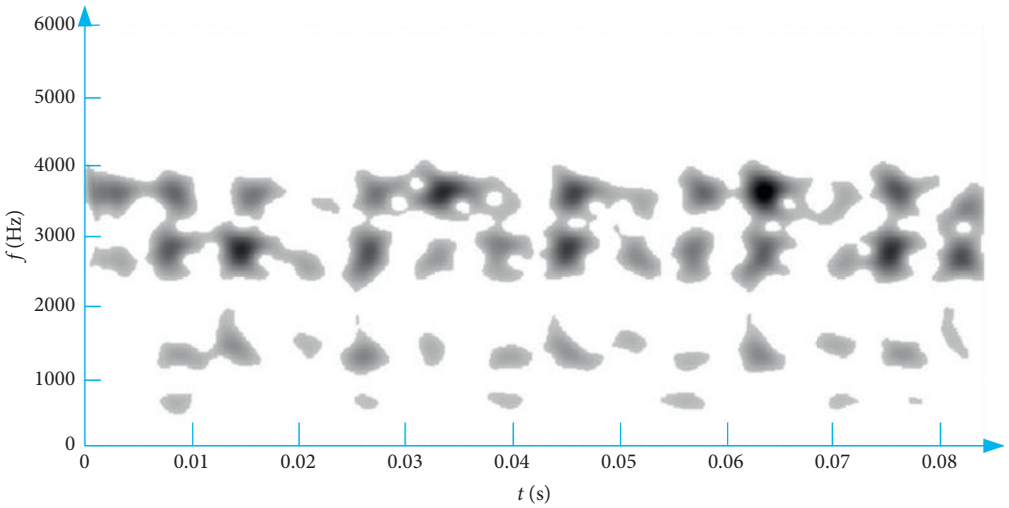


Figure 8: The main components image of the faulty bearing extracted with the mask template obtained from OTSU.

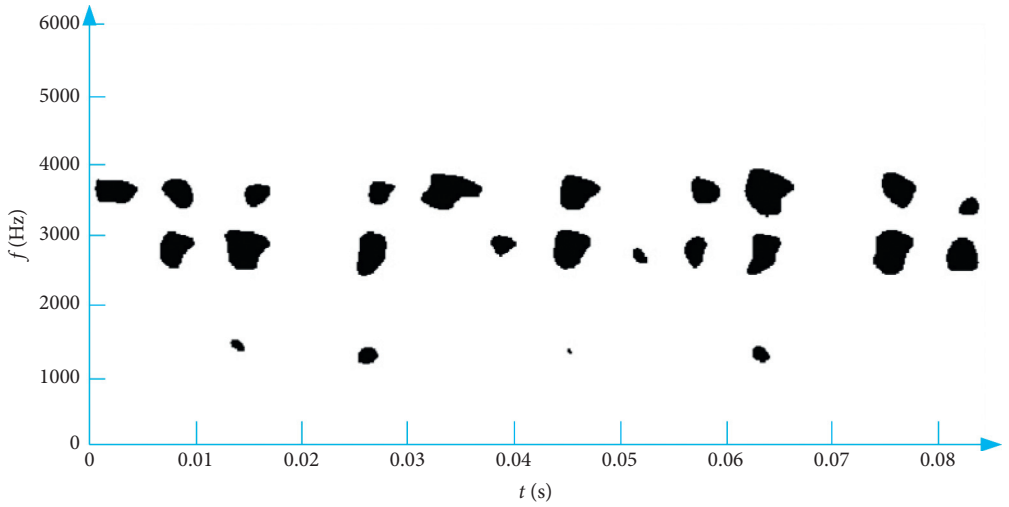


Figure 9: The mask template obtained with the method of KSW-entropy.

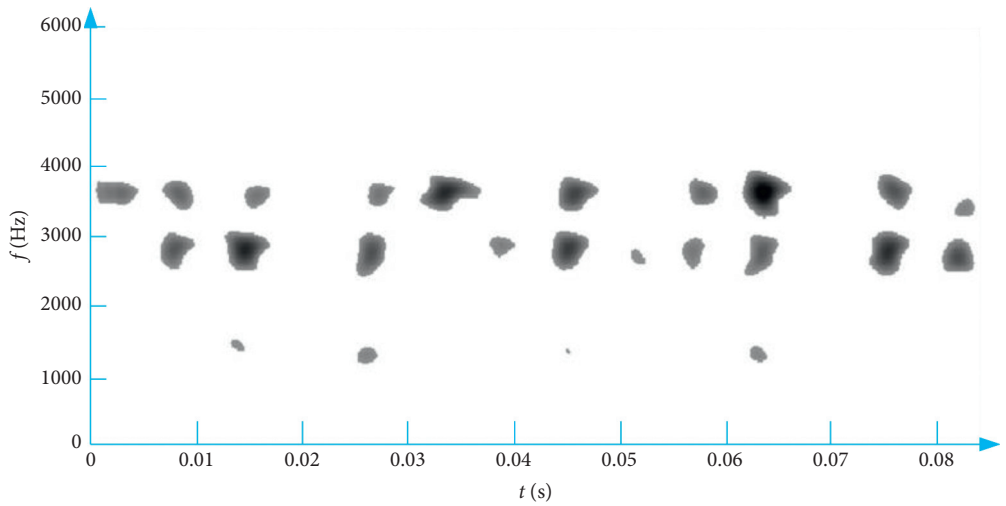


Figure 10: The main components' image of the faulty bearing extracted with the mask template obtained from KSW-entropy.

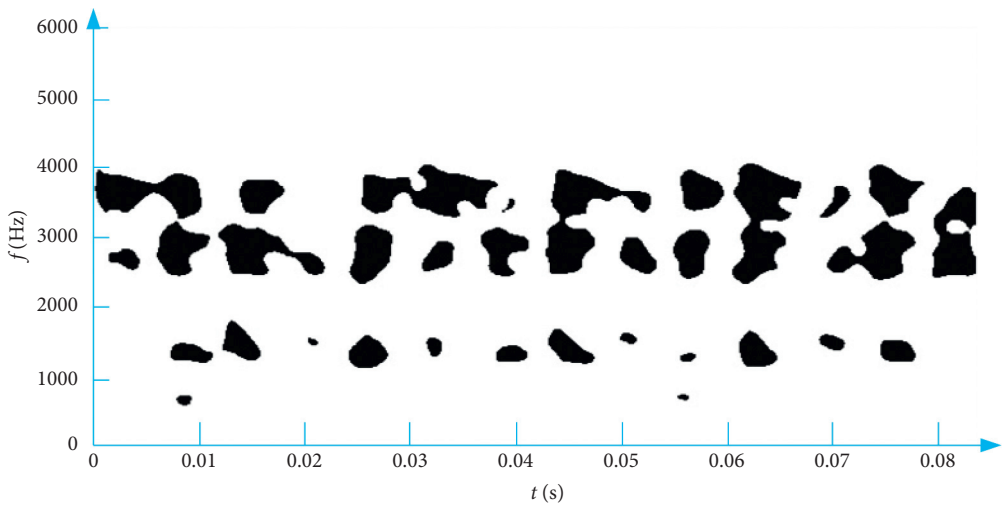


Figure 11: The mask template obtained with the method of our method.

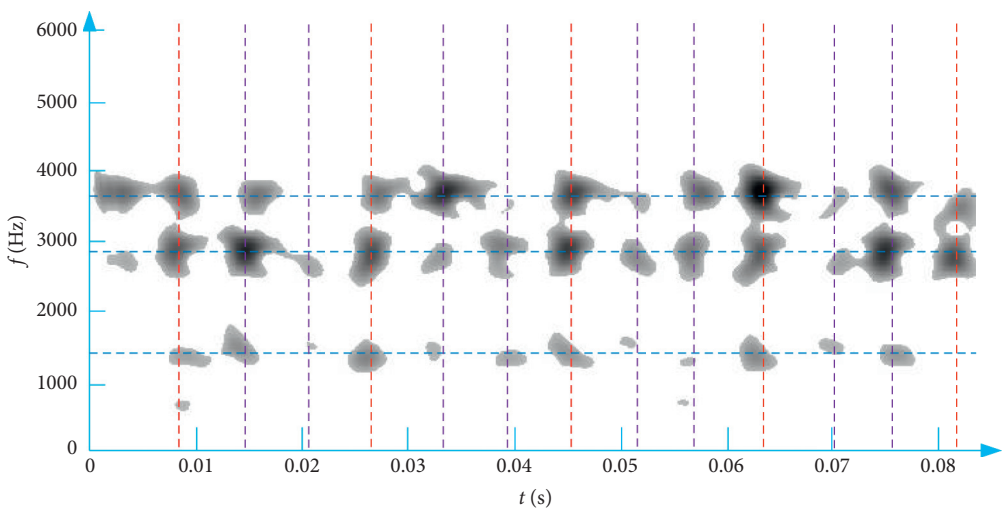


Figure 12: The main components' image of the faulty bearing extracted with the mask template obtained from our method.

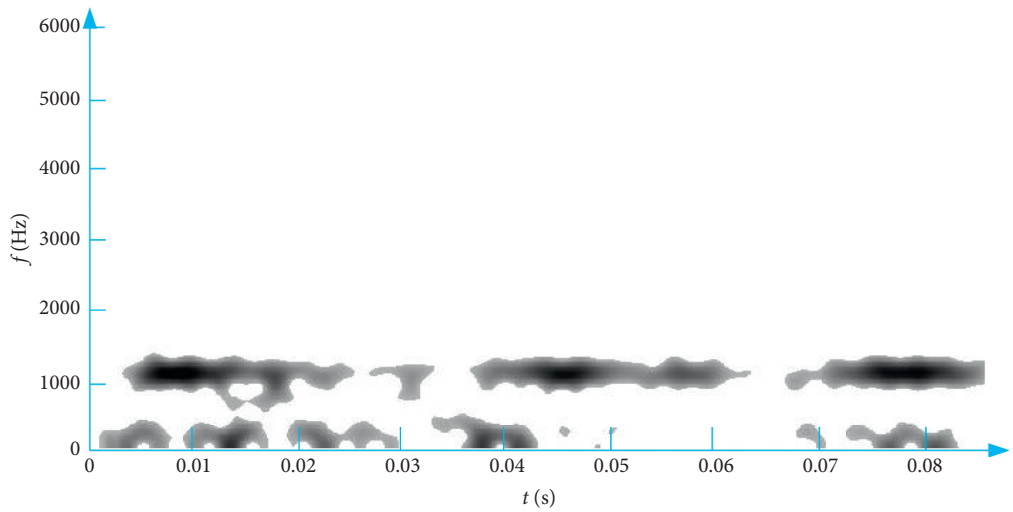


Figure 13: The main components' image of the normal bearing extracted with the mask template obtained from our method.

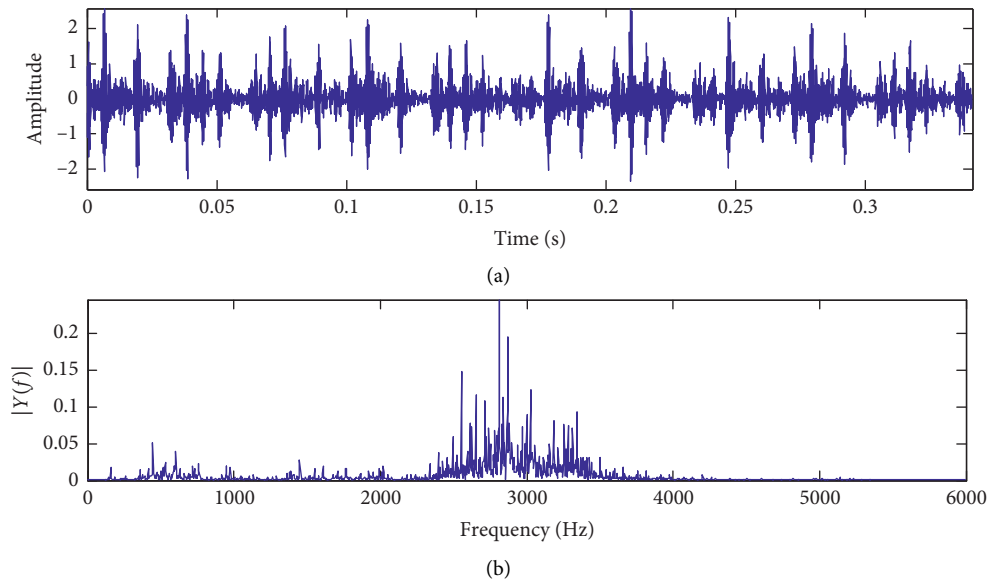


Figure 14: The waveforms of the IR021_2_X211_DE data in time and frequency domain: (a) time domain of data IR021_2_X211_DE_time_0000001_0004096 and (b) spectrum of data IR021_2_X211_DE_time_0000001_0004096.

time-frequency image of STFT and the extracted main components are shown in Figures 15 and 16.

The threshold we calculated with our method is 202. As shown in Figure 16, where the bearing failure is more serious, the signal components are still mainly concentrated in frequency band of 2400 Hz~3400 Hz. The signal components in this frequency band are very abundant and occur discontinuous. The time intervals are not constant and the intensity of the signal components are also various.

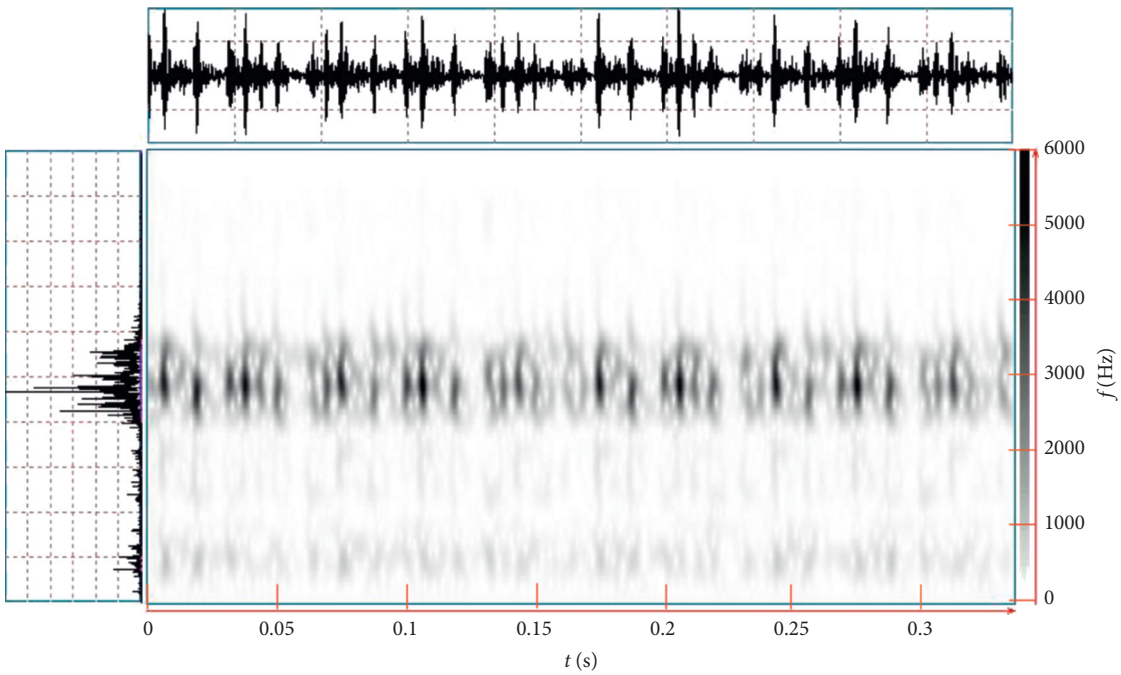


Figure 15: The joint time-frequency distribution image.

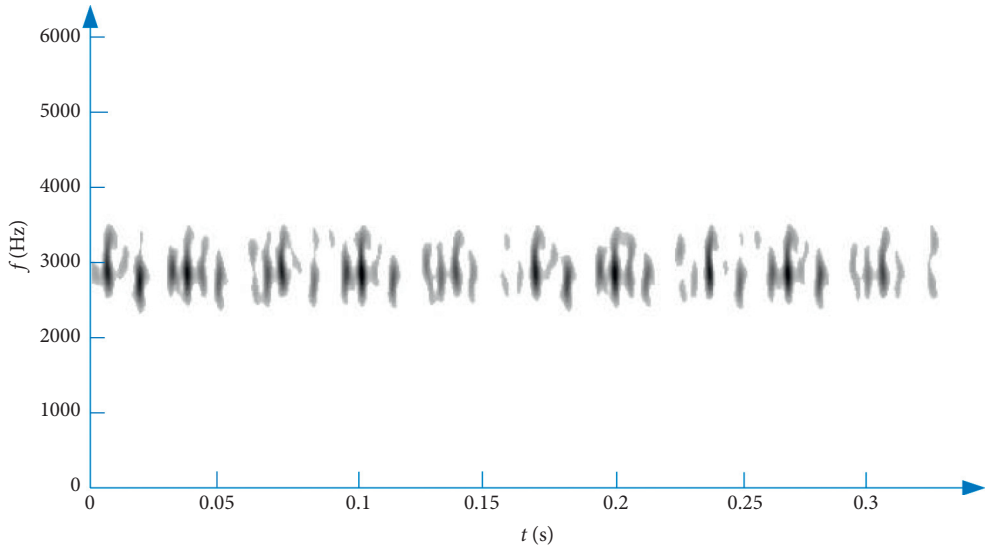


Figure 16: Main time-frequency components extracted by our method.

5. Conclusions

This paper presents an adaptive method of extracting the main components from time-frequency images, which is based on the gray histogram features of time-frequency images. In order to get a mask template, with which the main components of time-frequency images can be extracted, a threshold is firstly calculated adaptively by a method combined of OTSU and KSW-Entropy. Then, by the idea of binarization processing, the mask template and the original time-frequency image is operated with multiplication; thus, the signal components with little energy in time-frequency image can be filtered out. By this method, the main components of time-frequency images can be retained adaptively while some little details or noisy components can be filtered out, which will help us to focus on or find the characteristics of the time-frequency images obtained from the vibration signals of motor bearings. With this method, the effective pixel points of time-frequency images can be effectively reduced, and the amount of data to be processed during the later recognition processing will also be reduced, which will help us to use computers to automatically recognize or classify time-frequency images for bearing faults' diagnosis.

Data Availability

The data used to support the findings of this study can be obtained from https://csegroups.case.edu/bearingdatacenter/home.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by Sichuan Science and Technology Program (2019YFG0352). The authors would like to thank the Case Western Reserve University Bearing Data Center for providing the data for this study.

References

[1] L. Hongkun, Z. Zhixin, G. Zhenggang et al., "Rolling bearing fault diagnosis using Hough transform of time-frequency image," *Journal of Vibration, Measurement & Diagnosis*, vol. 30, no. 6, pp. 634–637, 2010.

[2] H. Isobe, S. Yamanaka, M. Okumura et al., "Fault diagnosis of reciprocating compressors using local wave time-frequency spectrum and image processing," in *Proceedings of the ISTM/2007 International Symposium on Test and Measurement*, Wuhan, China, 2007.

[3] Y. Cai, S. Cheng, Y. He, and P. Xu, "Application of image recognition technology based on fractal dimension for diesel engine fault diagnosis," in *Proceedings of the International Conference for Young Computer Scientists*, Hunan, China, 2008.

[4] D. Wei and L. Zhan-Sheng, "A fault diagnosis method based on gray level-gradient co-occurrence matrix of time-frequency image for rotating machinery," *Journal of Vibration Engineering*, vol. 20, pp. 85–91, 2009.

[5] Y. P. Cai, A. H. Li, L. S. Shi et al., "IC engine fault diagnosis method based on EMD-WVD vibration spectrum time-frequency image recognition by SVM," *Neiranji Gongcheng/Chinese Internal Combustion Engine Engineering*, vol. 33, pp. 72–78, 1979.

[6] D. Verstraete, A. Ferrada, E. L. Droguett, V. Meruane, and M. Modarres, "Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings," *Shock and Vibration*, vol. 2017, Article ID 5067651, 17 pages, 2017.

[7] J. Ville, "Theorie et application de la notion de signal analytique," *Cables et Transmission*, vol. 2A, pp. 61–74, 1948.

[8] B. Boashash and P. Black, "An efficient real-time implementation of the Wigner-Ville distribution," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 11, pp. 1611–1618, 1987.

[9] B. Boashash and P. O'Shea, "Polynomial Wigner-Ville distributions and their relationship to time-varying higher order spectra," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 216–220, 1994.

[10] R. K. Potter, G. Kopp, and H. C. Green, *Visible Speech*, Van Nostrand, New York, NY, USA, 1947.

[11] X. Hongjun, K. Jianbo, and D. Baizhen, "The STFT time frequency analysis of FH SS signal," *Journal of Guilin Institute of Electronic Technology*, 1998.

[12] J. Jing, H. Liu, and C. Lu, "Fault diagnosis of electro-mechanical actuator based on WPD-STFT time-frequency entropy and PNN," *Vibroengineering PROCEDIA*, vol. 14, pp. 130–135, 2017.

[13] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, "Wave propagation and sampling theory-part II: sampling theory and complex waves," *Geophysics*, vol. 47, no. 2, pp. 222–236, 1982.

[14] L. Debnath and J. P. Antoine, "Wavelet transforms and their applications," *Physics Today*, vol. 56, no. 4, p. 68, 2003.

[15] M. A. V. Klooster, M. Zijlmans, F. S. S. Leijten, C. H. Ferrier, M. J. A. M. Van Putten, and G. J. M. Huiskamp, "Time frequency analysis of single pulse electrical stimulation to assist delineation of epileptogenic cortex," *Brain*, vol. 134, no. Pt 10, pp. 2855–2866, 2011.

[16] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[17] Z. K. Peng, P. W. Tse, and F. L. Chu, "A comparison study of improved Hilbert–Huang transform and wavelet transform: application to fault diagnosis for rolling bearing," *Mechanical Systems & Signal Processing*, vol. 19, no. 5, pp. 974–988, 2005.

[18] C.-F. Lin and J.-D. Zhu, "Hilbert-Huang transformation-based time-frequency analysis methods in biomedical signal applications," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 226, no. 3, p. 208, 2012.

[19] H. K. Kwok and D. L. Jones, "Improved instantaneous frequency estimation using an adaptive short-time Fourier transform," *IEEE Transactions on Signal Processing*, vol. 48, no. 10, pp. 2964–2972, 2015.

[20] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[21] J. N. Kapur, P. K. Sahoo, and A. K. C Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273–285, 1985.

[22] Case Western Reserve University Bearing Data Center Website, https://csegroups.case.edu/bearingdatacenter/home.

[23] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study," *Mechanical Systems and Signal Processing*, vol. 64-65, pp. 100–131, 2015.

[24] C. Lu, Z.-Y. Wang, W.-L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Processing*, vol. 130, pp. 377–388, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5373094, 8 pages
https://doi.org/10.1155/2021/5373094

Research Article

Shield Reliability Analysis-Based Transfer Impedance Optimization Model for Double Shielded Cable of Electric Vehicle

Xiaoshan Wu,[1] **Xiaohui Shi**,[2] **Jin Jia**,[2] **Heming Zhao**,[3] **and Xu Li**[2]

[1]*School of Automotive Engineering, Chongqing University, Chongqing 401331, China*
[2]*Vehicle Engineering Institute, Chongqing University of Technology, Chongqing 400054, China*
[3]*Chongqing Qingyan Ligong Electronic Technology Co., Ltd., Chongqing 400020, China*

Correspondence should be addressed to Jin Jia; jj@cqcii.com

Received 25 August 2020; Accepted 9 October 2020; Published 8 February 2021

Academic Editor: Yong Chen

Due to the high-voltage and high-current operating characteristics of the electric drive system of electric vehicles, it forms strong electromagnetic interference during the working process. The shielding effectiveness of the high-voltage connection cable that connects the components of the electric drive system is directly related to its electromagnetic interference emissions. Therefore, the modeling and analysis of the shielding effectiveness of the connection cable is very important for the development of a connection cable with good shielding effectiveness. Firstly, the transfer impedance value representing the shielding effectiveness of the shielded cable is analyzed, and the difference between the single-layer shield and the double-layer shield cable is compared. The influence of double-layer shielded high-voltage connection cables commonly used in electric vehicles on the shielding layer DC resistance and keyhole inductance is clarified. Secondly, the transfer impedance optimization model $Z_{T\_D\text{-Desmoulins}}$ is obtained by combining with the single-layer shielded cable Desmoulins model and considering the influence of shielded layer DC resistance and keyhole inductance. Finally, three double-layer shielded cables of different types were selected for the triaxial test. The error rates of the test data and the $Z_{T\_D\text{-Desmoulin}}$ optimization model are all lower than 20% in each frequency band, which verified the correctness, universality, and great engineering application value of the optimization model.

1. Introduction

Due to the high-voltage and high-current working characteristics of the electric drive system of electric vehicles, strong electromagnetic interference is formed during the working process. The shielding effectiveness of the high-voltage connection cables that connect the components of the electric drive system is directly related to its electromagnetic interference emission level [1, 2]. The connection cables with poor shielding efficiency usually cause the electromagnetic field emission level of the entire electric vehicle to exceed the standard. Therefore, how to effectively evaluate and test the shielding effectiveness of the connector assembly is the common concern for cable and connector suppliers and vehicle manufacturers.

To solve the above problems, scholars at home and abroad have carried out extensive research on the surface transfer impedance of the connection cables. Vance [3] deeply studied the low-frequency characteristics when radiating to the cable braid and gave the most commonly used Z_d model formula. The braided inductance part was introduced by Tyni [4] to improve the accuracy when calculating the transfer impedance of high and low projection coverage cables; Demoulin and Degauque [5] proposed a new model that took into account the effects of additional fluctuations and generated additional attenuation; Marconi et al. [6] proposed a test method to measure the transfer impedance of two coaxial cables RG 213 under the same conditions and compared the test results with theoretical calculations; Xiaoling et al. [7] proposed a new model for accurately predicting the transfer impedance of braided coaxial cables by summarizing and studying the classical model; Mushtaq and Frei [8, 9] introduced the ground plate method (GPM) and capacitive voltage probe (CVP). He

compared the test results of the above methods with the triaxial injection method and line injection method. The above studies all used braided single-layer shielded cables as the research object. At present, high-voltage power cables on electric vehicles usually use double-layer shielding. The so-called double-layer shielding refers to the inner shielding layer using tinned copper wire braiding. The outer shielding layer is wrapped with aluminum-plastic composite tape (aluminum foil). However, domestic and foreign experts have not conducted in-depth research on the modeling method of double-layer shielded cables. The transfer impedance model of double-layer shielded cables is of great significance to the shielding effectiveness of vehicles. Therefore, it is necessary to carry out the research of the surface transfer impedance of the double-layer shielded cable connection cables.

In view of the above analysis, this article first introduces the difference in transfer impedance between single- and double-shielded cables, and the influence of double-layer shielded high-voltage connection cables used in electric vehicles on the shielding layer DC resistance and keyhole inductance is clarified. Then, based on the Demoulin model, the equivalent circuit diagram of the double-layer shielded cable is obtained by considering the influence of the shield layer DC resistance and keyhole inductance, and the optimization model of the double-layer shielded cable is established. Three sets of double-shielded cables are tested by the tri-coaxial method, and the correctness and generality of the optimized model are verified.

2. Surface Transfer Impedance and Its Analytical Formula

2.1. Surface Transfer Impedance. Surface transfer impedance [10, 11] is a characteristic parameter that characterizes the shielding performance of power cables. The lower the transfer impedance, the better the shielding performance of the cable and the stronger its electromagnetic immunity. It is defined [12] as the unit length of the cable. The induced voltage formed between the core wire and the shield layer when current flows through the shield layer (as shown in Figure 1), that is, the ratio of the axial voltage change rate on the braid layer to the axial current, and the calculation formula can be expressed as follows:

$$Z_T = \frac{1}{I_0}\frac{\partial V}{\partial z}, \tag{1}$$

where I_0 represents the current flowing through the outer surface of the braid; $(\partial V/\partial z)$ represents the effective value of the voltage per unit length of the uniform transmission line composed of the core wire and the shielding layer; z indicates the axial direction of the cable, as shown in Figure 1; l indicates the cable length.

2.2. Transfer Impedance Analytical Formula. Generally, the high-voltage power cables used in electric vehicles are braided shielded cables. As shown in Figure 2, the analytical model of the transfer impedance can be established by the

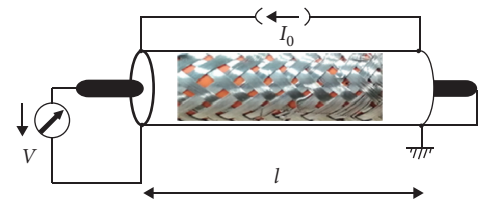


Figure 1: Schematic diagram of transfer impedance definition.

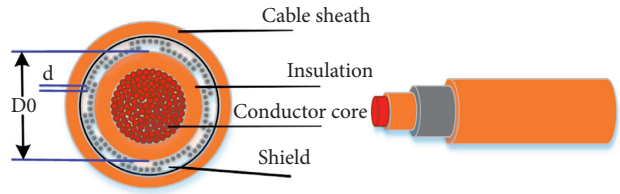


Figure 2: Schematic diagram of the shielded cable structure.

structural parameters of the cable braid and the electromagnetic field theory.

The analytical method can effectively analyze the influence of the parameterization of the shielded cable on the transfer impedance. Regarding the input parameters of the analytical model, the structural characteristics of the cable braid can be described by 7 parameters: the inner diameter of the braid D_0, the diameter of the braided wire d, the number of braid strands contained in a circle on the braided layer c, the number of wires in each braided bundle N, the braid angle α, the conductivity of the braided layer σ, and the magnetic permeability of the braided layer. After these parameters, the transfer impedance value of the shielded cable can be simulated; refer to the schematic diagram in Figure 3.

2.3. Analysis of Transfer Impedance Characteristic Curve. For power shielded cables, the equivalent circuit can be built through the RLC electrical parameters, as shown in Figure 4. The inductance parameter L is mainly composed of the inductance Lc of the core conductor, the inductance Ls of the shielding layer, and the mutual inductance M_cs between the two. In addition, the influence of small hole inductance and braided inductance Lh and Lb should be considered on the braided layer. The resistance parameter R is mainly composed of the resistance of the internal conductor Rc and the resistance of the shielding layer Rs. The resistance is affected by the skin effect and changes with frequency. The skin effect will affect the shielding effectiveness of the cable and the impedance value at the resonance frequency. The capacitance parameter C is composed of the capacitance C_cs between the core wire and the shielding layer. The transfer impedance value is mainly affected by inductive coupling, so we should pay attention to the influence of these electrical parameters on the transfer impedance of the power cable [13].

As shown in Figure 5, the composition of the Z_T curve of the shielded cable transfer impedance model is analyzed. The dotted lines in the figure are several key components that

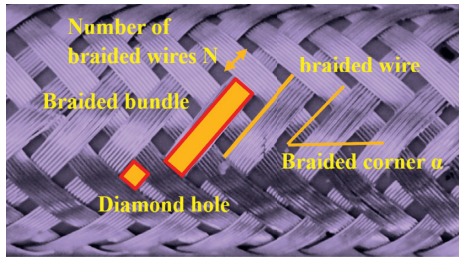


Figure 3: Schematic diagram of the structural parameters of the braid.

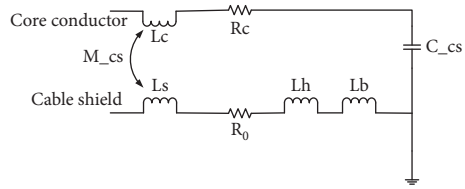


Figure 4: Shielded cable equivalent circuit diagram.

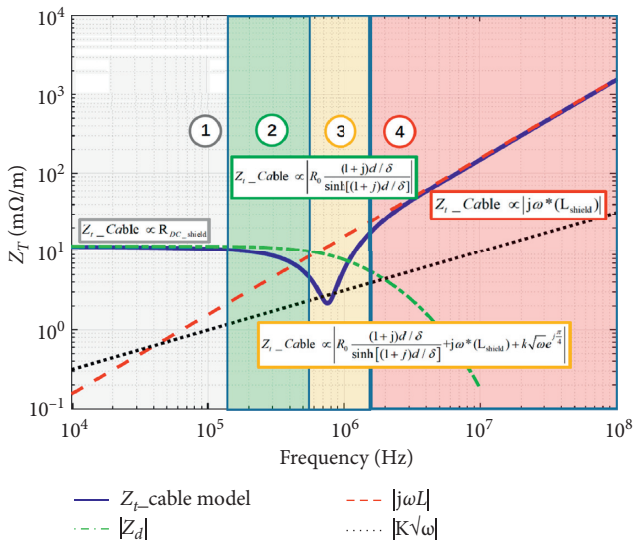


Figure 5: Z_T curve composition analysis diagram.

affect the value of transfer impedance, where Z_d is scattering impedance and $j\omega L$ is inductance and respectively, and the solid lines in the figure are the transfer impedance curves containing each component. In area 1 (gray), when the low frequency is less than 150 kHz, the current density in the braided shielding layer is evenly distributed, and the transfer impedance value is approximately the same as the DC resistance value R of the shielding layer. In zone 2 (green), the transfer impedance is mainly determined by the scattering impedance Z_d. As the frequency increases, the current density in the shielding layer becomes uneven. Due to the skin effect, the skin depth decreases according to the square root of the frequency, and the Z_T value decreases. In zone 3 (yellow) near 1 MHz, the transfer impedance value is determined by the scattering impedance Z_d, the small hole inductance $j\omega Lh$, the braided inductance $j\omega$Lb, and the additional wave attenuation, resulting in an obvious inflection point. As the frequency increases further, the magnetic field leakage caused by the diamond-shaped holes in the braided layer increases, and the inductance component of the small holes increases. The weaving of the braided bundles that intersect each other in the braid layer will cause the cutting of magnetic flux, which will also generate induced electromotive force, forming braided inductance, and increase the transfer impedance value. In the case of high frequency, the magnetic field between the inner and outer layers of the woven mesh will cause eddy current effects and additional attenuation. In area 4 (red), it is greater than 2 MHz, which is mainly determined by the small hole inductance and braided inductance. As the frequency increases, the transfer impedance value continues to increase [14, 15].

3. Double-Layer Shielding Optimization Model

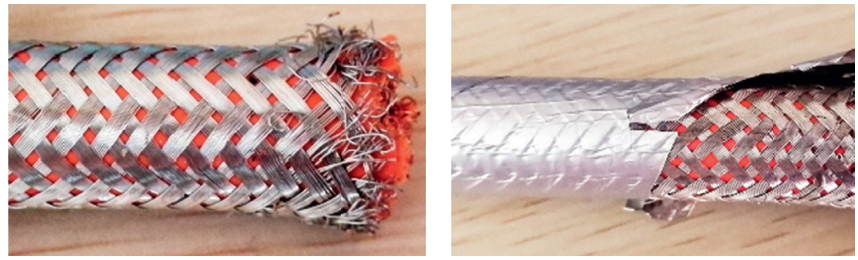
According to the above analysis of the composition of the Z_T curve, Z_T is mainly determined by the DC resistance when zone 1 (gray) is less than 150 kHz. At present, high-voltage power cables on electric vehicles usually adopt double-layer shielding. In addition to the inner tinned copper wire woven mesh, the outer layer is also wrapped with a layer of aluminum-plastic composite tape (aluminum foil), so the tested power cables also need to consider the influence of the DC resistance of the aluminum foil on the transfer impedance value at low frequencies.

In addition, it can be seen from Figure 6 that because the double-layer shielded cable [16] adds a layer of aluminum foil to the outside of the shielding layer, compared with the single-layer shielding, the diamond-shaped holes on the inside are covered by aluminum foil, which can effectively prevent the magnetic field from passing through small hole leaks, and the small hole inductance effect is greatly reduced and can be ignored. Therefore, it will affect the small hole inductance of the high-frequency part of area 4 (red) in Figure 5, and the transfer impedance value will theoretically decrease.

For the additional DC resistance value of aluminum foil, the aluminum foil layer model can be established by Q3D software for numerical analysis and calculation to extract the resistance value, as Figure 7 shows. Assuming that the aluminum foil model is ideal with a uniform thickness of 0.1 mm, any section can be selected for calculation. In order to reduce the calculation amount, the length is set to 100 mm, and the DC resistance of the aluminum foil in the cable under test is calculated. It is 0.004 Ω/m.

Demoulin proposed an analytical model of formula (12), taking into account the effects of additional volatility.

The additional wave effect is the eddy current effect caused by the magnetic field between the braided bundles of the inner and outer layers of the woven net at high frequencies, which will generate additional attenuation and lead to a decrease in the transfer impedance value in the high-frequency range. This component can be described by the eddy current caused by the tangential electric field on the shielding layer and is proportional to $\sqrt{\omega}$.



(a) (b)

Figure 6: Single-layer shielded cable (a) and double-layer shielded cable (b).

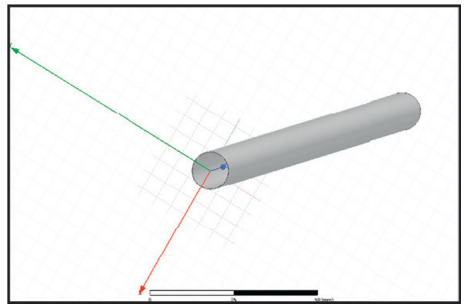


Figure 7: Aluminum foil Q3D calculation model.

$$Z_{T\_\text{Demoulin}} = Z_d + j\omega\left(L_{h2} - L_{b1}\right) + k\sqrt{\omega}e^{+j(\pi/4)}, \tag{2}$$

where

$$Z_d = R_0\frac{(1+j)d/\delta}{\sinh\left[(1+j)d/\delta\right]}, \tag{3}$$

$$\begin{aligned} R_0 &= \frac{4}{\pi d^2 NC\sigma\cos\alpha}, \\ \delta &= \frac{1}{\sqrt{\pi f\mu\sigma}}, \end{aligned} \tag{4}$$

$$L_{h2} = \frac{2\mu C}{\pi\cos\alpha}\left[\frac{b}{\pi D_M}\right]^2 \exp^{[(-\pi d/b)-2\alpha]}, \tag{5}$$

$$L_{b1} = \frac{\mu h}{4\pi D_M}\left(1 - \tan^2\alpha\right), \tag{6}$$

$$k = -\frac{1.16}{NC\,d}\cdot\arctan\frac{N}{3}\cdot\sin\left(\frac{\pi}{2} - 2\alpha\right)\cdot\sqrt{\frac{\mu}{\sigma}}. \tag{7}$$

Considering the impact of double-layer shielding, the equivalent circuit diagram of a double-layer shielded cable as shown in Figure 8 is obtained. Double-layer shielded power cables need to be considered in the analytical optimization model due to the influence of the DC resistance of the additional aluminum foil and the elimination of the small hole inductance on the transfer impedance.

In summary, the double-layer shielding optimization model is as follows:

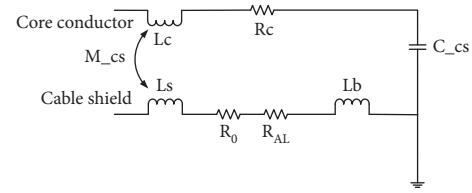


Figure 8: Equivalent circuit diagram of double-layer shielded cable.

$$Z_{T\_D\text{–Demoulin}} = Z_d' - j\omega L_{b1} + k\sqrt{\omega}e^{+j(\pi/4)}, \tag{8}$$

where

$$Z_d' = R'\frac{(1+j)d/\delta}{\sinh\left[(1+j)d/\delta\right]}, \tag{9}$$

$$R' = \frac{4}{\pi d^2 NC\sigma\cos\alpha} + R_{AL}, \tag{10}$$

$$k = -\frac{1.16}{NC\,d}\cdot\arctan\frac{N}{3}\cdot\sin\left(\frac{\pi}{2} - 2\alpha\right)\cdot\sqrt{\frac{\mu}{\sigma}}. \tag{11}$$

The single-layer shielding model and the double-layer optimization model were simulated and analyzed, as shown in Figure 9, considering increasing the DC resistance and eliminating the influence of the small hole inductance. It can be seen that the influence of the DC resistance is mainly in the low-frequency band and the influence of the inductance is in the high-frequency band. From the perspective of the overall optimization model, the transfer impedance of a double-layer shielded cable at high frequency is significantly lower than that of a single-layer shielded cable.

4. Triaxial Method Test and Model Comparison Verification

4.1. Triaxial Test. The triaxial method is a method in which the tested cable is placed in a coaxial nonferromagnetic good conductor tube for measurement, namely, the inner conductor of the cable core, the cable shielding layer, and the coaxial good conductor tube constituting a test device. The triaxial method can characterize the complex electromagnetic coupling mechanism [17] with directly measured

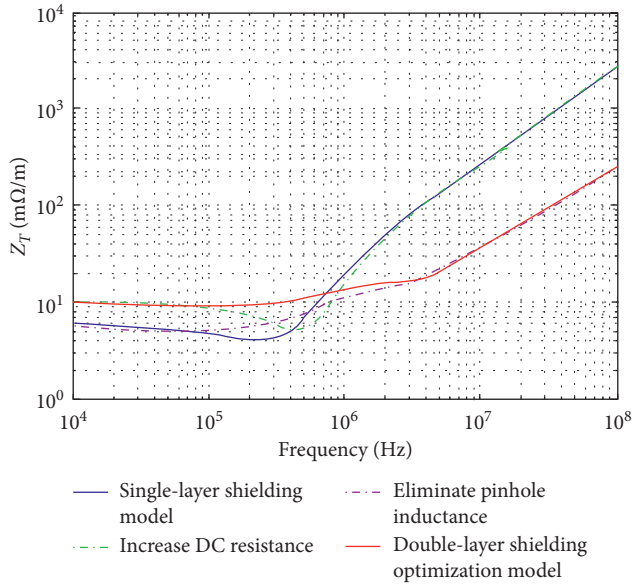


Figure 9: Comparison of single-layer shielding model and double-layer optimization model.

Table 1: Comparison of three coaxial a, b, and c methods [18].

Testing method	Test method a	Test method b	Test method c
Advantage	Test cut higher stop frequency	Measurement has more higher dynamic range	Suitable for measuring very low transfer impedance values (below 1 $\mu\Omega$/m m and lower)
Cutoff frequency	$f_{\text{cut}} * l \approx 80\,\text{MHz} * \text{m}$	$f_{\text{cut}} * l \approx 25\,\text{MHz} * \text{m}$	$f_{\text{cut}} * l \approx 30\,\text{MHz} * \text{m}$
Disadvantages	Low dynamic range of measurement	Test cut lower frequency	The effect of capacitive coupling is suppressed by the short circuit in the primary and secondary circuits, and the test is quite sensitive
Features	Matching resistance and impedance mismatch Need to connect to the matching impedance network Near-end core wire injection	Only need to connect the terminal matching resistor Near-end core injection signal	No matching resistor No need to access matched impedance network inject signal from remote test tube

circuit parameters (the electromagnetic field that affects the shielding effectiveness is replaced by surface current and surface charge equivalent), suitable for asymmetric cables and different size and structure complex connector testing, and the test results are repeatable.

Table 1 shows the comparison of triaxial methods a, b, and c. Considering the test stability, ease of operation, and commonality, this test adopts the triaxial B method; refer to the test standard: IEC62153-4-3-2013 [19]. Figure 10 shows the triaxial test layout of the power cable.

Calculation method of the triaxial b method transfer impedance value:

$$Z_t = \frac{R_1 + Z_0}{2 \cdot L_c} \cdot 10^{-\{a_{\text{means}} - a_{\text{cal}}/20\}}, \tag{12}$$

where $a_{\text{means}} = 20\log_{10}(S_{21})$ represents the measured attenuation loss, a_{cal} represents the composite loss measured during calibration, Z_0 represents the impedance of the signal generator and receiver, usually 50 Ω, and L_c represents the coupling length of the tested cable and R_1 represents terminal impedance.

4.2. Comparison and Verification of Multiple Samples. Combined with the triaxial b method test schematic diagram (as shown in Figure 11), the flowchart of the power cable test was developed (as shown in Figure 12). As shown in Table 2, three double-layer shielded cables with different parameters were selected for testing. In the test, the coupling length of the three groups of samples is 0.5 m. When the triaxial b method is used, the test cutoff frequency is 50 MHz (the maximum measurable 50 MHz). As shown in Figure 13, the test value of the sample cable in the figure produces a resonance point at 50 MHz and the trend of the transfer impedance curve changes. Comparing the simulation value of the double-layer shielding optimization model with the actual test value of the

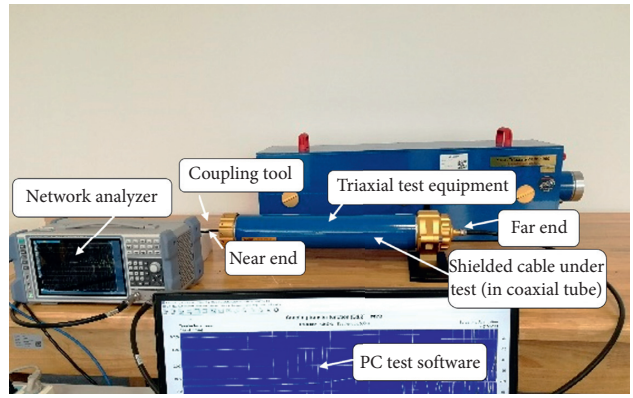


Figure 10: Physical diagram of power cable test layout.

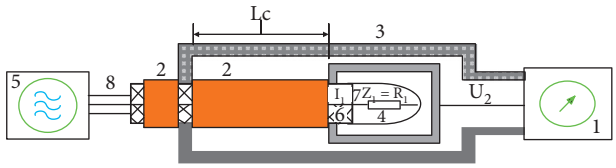


Figure 11: Schematic diagram of the test principle of the triaxial b method. (1) Network analyzer or receiver; (2) cable insulation sleeve; (3) test sleeve; (4) terminal impedance R1 length; (5) signal generator; (6) cable shield; (7) test core wire; (8) test connection line. L_c: cable coupling.

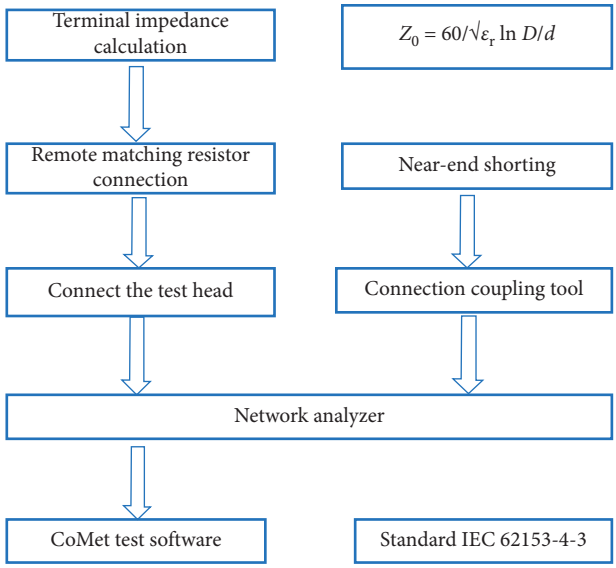


Figure 12: Power cable test flowchart.

Table 2: Sample cable parameters.

	Sample 1	Sample 2	Sample 3
Inner diameter of woven layer D_0 (mm)	10.3	11.58	10.02
Braided wire diameter d (mm)	0.15	0.15	0.15
Number of braid strands c	8	10	10
Number of wires per share n	24	24	24
Weaving angle α	35	38	38

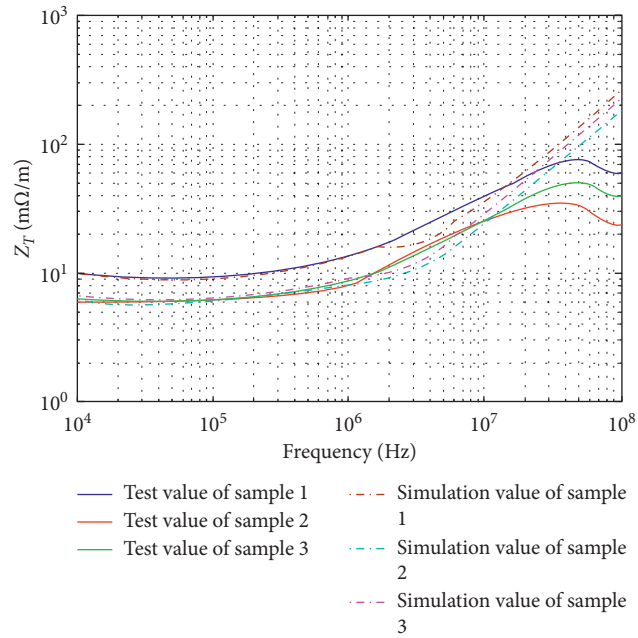


Figure 13: Comparison and analysis of simulated values and test values of samples 1, 2, and 3.

sample cable, it can be seen that the simulation models of the three samples have a good fitting effect.

Table 3 shows the numerical comparison between the optimized model and the test value of three different samples at several common frequencies. As can be seen from the data in Table 3, the simulated calculated value of the optimized model at each frequency point is very close to the actual measured value after removing the keyhole inductance and considering adding the aluminum foil DC resistance. At the frequency of 150 kHz, the error rates of the three samples were 2.16%, 0.50%, and 2.55%, which were all lower than 3%; at the frequency of 10 MHz, the error rates of sample 1 and sample 2 were 1.84% and 1.02%, respectively, both less than 2%, and the error rate of sample 3 is 9.01%. At 1 MHz and 2 MHz, the error rates of sample 1 were 7% and 8.86%, respectively, which were lower than 9%; the error rates of sample 2 and sample 3 were both lower than 12%; at 20 MHz

Table 3: Comparative analysis of optimization model and test value.

Frequency (kHz)	Sample 1 test value (mΩ/m)	Sample 2 test value (mΩ/m)	Sample 3 test value (mΩ/m)	Sample 1 optimization model (mΩ/m)	Sample 2 optimization model (mΩ/m)	Sample 3 optimization model (mΩ/m)	Sample 1 error rate (%)	Sample 2 error rate (%)	Sample 3 error rate (%)
150	9.74	6.05	6.27	9.532	6.08	6.11	2.16	0.50	2.55
1	12.72	8.06	8.27	13.61	8.72	9.13	7	8.19	10.40
2	17.04	11.15	11.33	15.53	9.88	10.61	8.86	11.39	6.35
10	37.52	24.52	25.86	36.83	24.27	28.19	1.84	1.02	9.01
20	54.94	35.33	43.09	61.09	41.4	50.28	11.19	17.18	16.69
30	68.37	48.07	57.93	78.14	57.56	69.05	14.29	19.74	19.20

and 30 MHz, the error rates of the three samples were less than 20%. In summary, the experimental data of the three samples and the calculation error rate of the optimization model in each frequency band were all less than 20%, which verified the correctness of the optimization model and the universality of the optimization model.

5. Conclusion

(1) This paper analyzed the transfer impedance value representing the shielding effectiveness of the shielded cable, compared the difference between the single-layer shielded cable and the double-layer shielded cable, and clarified the influence of the double-layer shielded high-voltage connection cable commonly used in electric vehicles on the DC resistance and keyhole inductance of the shielded cable.

(2) Considering the influence of the shielding layer's DC resistance and small hole inductance, the optimization model $Z_{T\_D\text{-Desmoulins}}$ for the transfer impedance of the double-layer shielded cable was obtained, and the single-layer shielding model and the double-layer shielding model were simulated and analyzed. The influence of DC resistance in a low-frequency band and inductance in a high-frequency band was determined.

(3) Three different types of double-shielded cables were selected for the triaxial test. The calculation error rates of the test data and the $Z_{T\_D\text{-Desmoulin}}$ optimization model in each frequency band were less than 20%. At the frequency point of 10 MHz that the enterprise focuses on, the error rates of the three double-layer shielded cables were all lower than 10%, and two of them were 1.84% and 1.02%, which almost coincided with the test data. The correctness and generality of the optimization model were verified, and it had good engineering application value.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Key R&D Program of China (no. 2018YFB0106100), Key R&D Program of Science and Technology Major Theme Project of Chongqing (no. cstc2018jszx-cyctzxX0005), and Direction Common Technology Program of Electric Vehicle Industry Technology Innovation Strategies Alliance (no. CA2019).

References

[1] G. M. Kunkel, "Transfer impedance testing of shielded cables, back shells, and connectors," *Shielding of Electromagnetic Waves*, pp. 69–71, 2020.

[2] O. Gassab, S. Bouguerra, L. Zhou, and W.-Y. Yin, "Efficient analytical model for the transfer impedance and admittance of noncoaxial/Twinax braided-shielded cables," *IEEE Transactions on Electromagnetic Compatibility*, no. 99, pp. 1–12, 2020.

[3] E. Vance, "Shielding effectiveness of braided-wire shields," *IEEE Transactions on Electromagnetic Compatibility*, vol. 17, no. 2, pp. 71–77, 1975.

[4] M. Tyni, "The transfer impedance of coaxial cables with braided conductors," in *Proceedings of the EMC Symposium*, Wroclaw, Poland, 1976.

[5] B. Demoulin and P. Degauque, "Shielding effectiveness of braids with high optical coverage," in *Proceedings of the International Symposium on EMC*, Zurich, Switzerland, 1981.

[6] K. Marconi, C. L. Andrade, V. D. Silva et al., "Evaluation of surface transfer impedance of coaxial cables," *IEEE Latin America Transactions*, vol. 18, pp. 598–603, 2020.

[7] W. Xiaoling, L. Chao, D. Hao et al., "An improved model for the transfer impedance calculations of braided coaxial cables," in *Proceedings of the 7th International Power Electronics and Motion Control Conference (IPEMC 2012)*, Harbin, China, June 2012.

[8] A. Mushtaq and S. Frei, "Alternate methods for transfer impedance measurements of shielded HV-cables and HV-cable-connector systems for EV and HEV," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 26, no. 4, pp. 359–366, 2016.

[9] K. Marconi and Andrade, "Evaluation of surface transfer impedance of coaxial cables," *IEEE Latin America Transactions*, vol. 18, no. 3, pp. 598–603, 2020.

[10] H. Schippers and J. Verpoorte, "Uncertainties in transfer impedance calculations," in *Proceedings of the 2016 ESA Workshop on Aerospace EMC (Aerospace EMC)*, Valencia, Spain, May 2016.

[11] J. L. Rotgerink, J. Verpoorte, and H. Schippers, "Uncertainties in coaxial cable transfer impedance," *IEEE Electromagnetic Compatibility Magazine*, vol. 7, no. 3, pp. 83–93, 2018.

[12] S. Bauer, C. Turk, W. Renhart et al., "Finite element analysis of cable shields to investigate the behavior of the transfer impedance with respect to fast transients," in *Proceedings of the 2019 IEEE 23rd Workshop on Signal and Power Integrity (SPI)*, Chambery, France, June 2019.

[13] J. Verpoorte, H. Schippers, and J. H. G. J. Lansink Rotgerink, "Advanced Models for the Transfer Impedance of Metal Braids in Cable harnesses," in *Proceedings of the 2018 IEEE International Symposium on Electromagnetic Compatibility and 2018 IEEE Asia-Pacific Symposium on Electromagnetic Compatibility (EMC/APEMC)*, Suntec City, Singapore, May 2018.

[14] S. Weber, S. Guttowski, E. Hoene et al., "EMI coupling from automotive traction systems," in *Proceedings of the 2003 IEEE International Symposium on Electromagnetic Compatibility*, Istanbul, Turkey, May 2003.

[15] S. Frei, A. Mushtaq, K. Hermes et al., "Current distribution in shielded cable-connector systems for power transmission in electric vehicles," in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility and 2018 IEEE Asia-Pacific Symposium on Electromagnetic Compatibility (EMC/APEMC)*, Suntec City, Singapore, 2018.

[16] N. Mora, F. Rachidi, P. Pelissou, and A. Junge, "An improved formula for the transfer impedance of two-layer braided cable shields," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 3, pp. 607–610, 2015.

[17] P. Xiao, P. A. Du, and B. Zhang, "An analytical method for radiated electromagnetic and shielding effectiveness of braided coaxial cable," *IEEE Transactions on Electromagnetic Compatibility*, vol. 61, no. 1, pp. 1–7, 2018.

[18] H. Kim and T. Jang, "Comparison of measurement results on the transfer impedance of a coaxial cable," in *Proceedings of the 2017 Asia- Pacific International Symposium on Electromagnetic Compatibility (APEMC)*, Seoul, Republic of Korea, 2017.

[19] IEC 62153-4-3, 2.0 2013-10. Metallic communication cable test methods-part 4-3: electromagnetic compatibility (EMC)-surface transfer impedance-triaxial method.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6648578, 21 pages
https://doi.org/10.1155/2021/6648578

Research Article

Adaptive Fuzzy Modified Fixed-Time Fault-Tolerant Control on SE(3) for Coupled Spacecraft

Yafei Mei,[1] **Ying Liao**,[1] **Kejie Gong**,[1] **and Da Luo**[2]

[1]*College of Aerospace Science and Engineering, National University of Defense Technology, No. 109 Deya Road, Changsha 410073, Hunan, China*
[2]*Shanghai Institute of Satellite Engineering, Shanghai 201109, China*

Correspondence should be addressed to Yafei Mei; meiyafei@nudt.edu.cn

Received 2 November 2020; Revised 18 December 2020; Accepted 28 December 2020; Published 11 January 2021

Academic Editor: Yong Chen

This paper aims to solve the control problem of coupled spacecraft tracking maneuver in the case of actuator faults, inertia parametric uncertainties, and external disturbances. Firstly, the spacecraft attitude and position coupling kinematics and dynamics model are established on the Lie group SE(3), and the coupled relative motion tracking error model is derived by exponential coordinates. Then, considering the actuator faults, an adaptive fuzzy scheme is proposed to estimate the lumped disturbances in real time, and a novel modified fixed-time terminal sliding mode fault-tolerant control law is developed to deal with the actuator faults and compensate the lumped disturbances. Next, the Lyapunov method is used to prove the stability and convergence of the system. Finally, the proposed controller can achieve fast and high-precision fault-tolerant control goals, and its effectiveness and feasibility are verified by numerical simulation.

1. Introduction

In the context of the rapid development of space technology, new and higher requirements have been put forward for the mobility and accuracy of spacecraft. The modeling and control of the attitude and trajectory of relatively moving spacecraft has always been a hot research topic in the fields of space rendezvous and docking, spacecraft formation flying (SFF) [1, 2]. Due to the strong coupling and nonlinearity of the relative motion of spacecraft's attitude and position motion, the conventional idea of dividing attitude and position motion into independent two-channel control ignores the influence of coupling between the two, although it satisfies the requirements of some space missions. However, for aerospace missions with high-precision requirements, the divide-and-conquer method will appear powerless [3]. Therefore, seeking the integrated control of spacecraft attitude and position has theoretical guidance and is of great significance to engineering practice.

Due to long-term exposure to harsh space environments such as strong radiation and ultra-low temperature, the actuator will have various types of failures. Therefore, the conventional control theory based on the normal operation of the actuator may be difficult to cope with the failure and may eventually cause the system to crash or fail. In addition, the spacecraft itself will also face the uncertainty of internal parameters and external disturbances, which brings huge challenges to the design of the control system. So it is particularly important to choose a suitable fault-tolerant control strategy for the aforementioned drawbacks, which also provides a strong guarantee for the long-term service operation of the spacecraft.

At present, many research results have been made on the problem of spacecraft attitude fault-tolerant control [4, 5]. But for the spacecraft attitude and position coupling control system, when various faults occur in the relative attitude and position actuators at the same time, the related six-degree-of-freedom (6-DOF) fault-tolerant control algorithm design is not enough [6]. Dong et al. [7, 8], studied the integrated fault-tolerant control of the spacecraft's position and attitude in the case of actuator failure based on the dual quaternion, and their numerical simulation results verified the effectiveness of the algorithm.

In recent years, the coupled modeling of rotation and translation relative motion based on different forms of rigid spacecraft has attracted widespread attention. Common spacecraft attitude and position coupling modeling and control forms mainly include dual quaternion [7, 8], Lie group SE(3) [9, 10], modified Rodrigues parameters (MRPs), and other forms [11, 12]. Although the integrated modeling method of spacecraft attitude and position based on dual quaternion is widely used, dual quaternion also has its limitations. The model based on dual quaternion uses eight parameters to describe the three-dimensional motion, so it requires unitized constraints. Sometimes improper handling of this constraint will cause problems. Moreover, since the group function corresponding to the unit quaternion is left multiplication and right multiplication, the quaternion description rotation is not unique, which will cause ambiguity, and when it is serious, it will cause unwinding problem [3]. Moreover, describing the attitude based on MRPS is nonglobal and nonunique [13]. Compared with the traditional description method in Euclidean space, the geometric framework of Lie group SE(3) is more natural and concise, the analysis results are more realistic and credible, and the designed controller is more concise, so in recent years, it received attention gradually. Lie group SE(3) can describe the three-dimensional translation and rotation of a rigid body. Lee et al. [14], Sanyal et al. [15], and Bullo and Murray [16], conducted an in-depth study on the control of the 6-DOF motion of a spacecraft on SE(3). By using the relationship between the exponential mapping function and the logarithm mapping function of Lie group and Lie algebra, the motion spinor is transformed into the corresponding spacecraft attitude and position motion equation. On this basis, various simple controllers are designed to realize the pose tracking control target [17]. In this paper, the integrated model of spacecraft attitude and position coupling is established in the framework of Lie group SE(3), which is convenient for the design of fault-tolerant controllers in the following.

Fault-tolerant control mainly includes active fault-tolerant control and passive fault-tolerant control. Among them, considering the actuator failure belongs to the integrity design category of passive fault-tolerant control, it is also a hot research direction in the field of fault-tolerant control and has obtained rich research results [18–22]. Fuzzy approximation can make full use of the information ability of fuzzy logic systems; it is easier to construct and can approximate nonlinear functions with arbitrary accuracy. When the actuator fails, the uncertainty of the system increases. For the parameter uncertain system, the adaptive law can be constructed by the Lyapunov method, and the uncertain parameters in the model can be replaced by the adaptive control based on the principle of equivalence. Finally, the adaptive law is designed for the estimated parameters to make the closed-loop system stable. This mainstream adaptive control method has been widely used in the field of spacecraft control due to its simple design and easy to understand [23]. Recently, many major achievements in the engineering application of fuzzy approximation methods have been reported, such as application of adaptive fuzzy controller in industrial process [24, 25]. At the same time, fuzzy control is also applied to robust fault-tolerant control for fault detection and actuator faults [26–29]. In addition, the fuzzy control scheme to approximate the disturbance of the spacecraft has been successfully applied, and it is effective to combine the adaptive fuzzy controller of NFTSMC in [30, 31] to reject the system uncertainty. Zhang et al. [32] applied fuzzy adaptive finite-time control to the 6-DOF SFF system and achieved success when considering the consensus control problem among the followers with signal transmission time delays.

Fixed-time control is developed on the basis of finite-time control. The difference between the two is only in the form of the sliding surface. The former can achieve fixed-time convergence without relying on the initial state, while the latter's convergence time is related to the initial state. Double-power fast terminal sliding mode control is a kind of fixed time control, which can be used to realize the fixed time stability of the system, which is more useful than the finite-time sliding mode control methods, such as terminal sliding mode (TSM) [33], fast terminal sliding mode (FTSM) [34], and nonsingular fast terminal sliding mode (NFTSM) [35]. It has faster convergence speed and better control effect. Shi et al. achieved attitude tracking control of rigid spacecraft on Lie group with fixed-time convergence [36] and global fixed time attitude tracking control for the rigid spacecraft with actuator saturation and faults [37]. Gao et al. proposed adaptive fixed time attitude tracking control for rigid spacecraft with actuator faults on MRPs [38]. Gong et al. proposed modified adaptive fixed-time terminal sliding mode control on SE(3) for coupled spacecraft tracking maneuver [3]. Mobayen et al. [39] proposed a new adaptive finite-time stabilization method based on global sliding mode to advance the steady state and transient performances of a class of chaotic flows in the presence of disturbances. Also, Mobayen and Pujol-Vázquez used robust LMI approach to deal with nonlinear feedback stabilization of continuous state-delay systems with lipschitzian nonlinearities and verified the effectiveness through experiments [40]. Jafari and Mobayen [41] combined LMI approach and second-order sliding set design for a class of uncertain nonlinear systems with disturbances.

Motivated by the facts mentioned above, this paper takes the leader-follower formation spacecraft as the research object. Firstly, a dynamic model of the relative tracking error of the spacecraft attitude and position coupling with model uncertainties, external disturbances, and actuator faults is derived on the Lie group SE(3). Then, the adaptive fuzzy method is used to design the sliding mode controller to realize the fixed-time fault-tolerant control.

The novelty of this paper is as follows: inspired by [3, 32], the proposed model in this paper takes the actuator faults into consideration. Based on the established model, a modified double-power fast terminal sliding manifold is defined by the exponential coordinates and velocity tracking errors, and then adaptive fuzzy modified fixed-time fault-tolerant control schemes is proposed, in which the adaptive fuzzy control technique is applied to reject the system lumped disturbances. Compared with finite-time stability and traditional fixed time stability, the control performance

obtained in this paper has significant advantages in convergence accuracy and effectiveness.

The structure of this paper is as follows: Section 2 introduces the mathematics preliminaries and the rigid body dynamics of the spacecraft on SE(3) with actuator faults. Section 3 adopts fuzzy adaptive method to design the sliding mode controller to realize the modified fixed-time fault-tolerant control and uses the Lyapunov method to prove the stability of the system strictly. Section 4 verifies the effectiveness of this method through numerical simulation. Section 5 draws conclusions and summarizes this paper.

2. Mathematics Preliminaries

In order to facilitate the design and stability proof and analysis of the integrated attitude and position control system, the following will give some related definitions and stability theory and lemmas.

2.1. Notations. For any column vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$, define the following vector and operation:

(1) $|\mathbf{x}|^\alpha = [|x_1|^\alpha, |x_2|^\alpha, \ldots, |x_n|^\alpha]^T$, where $|\cdot|$ is the absolute value.

(2) $\|x\|$ denotes the Euclidean norm or its induced norm.

(3) $\mathrm{sig}^\alpha(\mathbf{x}) = |\mathbf{x}|^\alpha \mathrm{sgn}(\mathbf{x}) = [|x_1|^\alpha \mathrm{sgn}(x_1), |x_2|^\alpha \mathrm{sgn}(x_2), \ldots, |x_n|^\alpha \mathrm{sgn}(x_n)]^T$, where $\mathrm{sgn}(\cdot)$ is the sign function.

(4) $(d|\mathbf{x}|^\alpha/dt) = \mathrm{diag}[\alpha \mathrm{sig}^{\alpha-1}(\mathbf{x})]\dot{x}$, $(d\mathrm{sig}^\alpha(\mathbf{x})/dt) = \mathrm{diag}[\alpha|\mathbf{x}|^{\alpha-1}]\dot{x}$.

(5) $[\cdot]^\wedge$ represents an operator that maps a vector to a Lie algebra. For $\varsigma e(3)$, it maps a vector to an skew-symmetric matrix, that is, $\mathbf{u}^\wedge = \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix}$; $[\cdot]^\vee$ represents the operator that maps the Lie algebra to a vector. For $\varsigma e(3)$, it maps the skew-symmetric matrix to a vector, which is $\begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix}^\vee = \mathbf{u}$.

2.2. Relative Coupled Dynamics of Spacecraft on Lie Group SE(3). Firstly, in order to describe the space orientation of the leader-follower spacecraft and establish the kinematics and dynamics model, we introduce three reference frames that are all orthogonal coordinate systems as shown in Figure 1. The Earth-centered inertial (ECI) reference frame with the origin at the center of the Earth is represented by $\{I\} = \{x_I, y_I, z_I\}$, which is used to describe the absolute motion of the spacecraft relative to the Earth. The body-fixed frames of the leader spacecraft and the follower spacecraft can be expressed as $\{Lb\} = \{x_{Lb}, y_{Lb}, z_{Lb}\}$ and $\{Fb\} = \{x_{Fb}, y_{Fb}, z_{Fb}\}$, respectively; their origin is at the center of mass of the spacecraft, and the axis coincides with the principal axis of inertia.

In nature, the configuration space of rigid body motion is SE(3), which can express translation and rotation of rigid body compactly. The special Euclidean group SE(3) is the semidirect product of the three-dimensional Euclidean space and the special orthogonal space, which can be expressed as $\mathrm{SE}(3) = \mathbb{R}^3 \ltimes \mathrm{SO}(3)$. $\mathbb{R}^3$ is used to describe the translation of the rigid body's center of mass; $\mathrm{SO}(3) = \{\mathbf{R} \in \mathbb{R}^{3\times3} | \mathbf{R}\mathbf{R}^T = \mathbf{I}, \det(\mathbf{R}) = 1\}$ is a rotation group composed of a three-dimensional rotation matrix, which is used to represent the rotation of the rigid body around the center of mass. Therefore, an element $\mathbf{g}$ in the Lie group SE(3) can express the configuration of the spacecraft [3]:

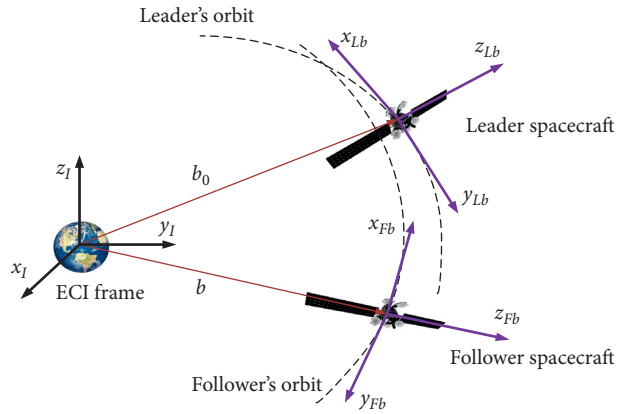


FIGURE 1: The relative motion and coordinate reference frames definition of the leader-follower spacecraft.

$$\mathbf{g} = \begin{bmatrix} \mathbf{R} & \mathbf{b} \\ 0_{1\times3} & 1 \end{bmatrix} \in \mathrm{SE}(3), \tag{1}$$

where $\mathbf{R} \in \mathrm{SO}(3)$ is the rotation matrix of the spacecraft from the body-fixed frame to the ECI reference frame and $\mathbf{b} \in \mathbb{R}^3$ is the position vector from the center of mass of the Earth to the center of mass of the spacecraft in the ECI coordinate system.

The angular velocity and translational velocity of the spacecraft are defined as

$$\boldsymbol{\xi} = \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\omega} \end{bmatrix} \in \mathbb{R}^6. \tag{2}$$

The above velocity vectors are defined in the body-fixed frame of the spacecraft. To describe the kinematics and dynamics equations below, it is necessary to introduce the Lie group SE(3) and its corresponding Lie algebra $\varsigma e(3)$ to meet the following mapping relations:

$$\boldsymbol{\xi} = \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\omega} \end{bmatrix} \in \mathbb{R}^6. \tag{3}$$

The adjoint matrix of $\mathbf{g} = \mathbf{g}(\mathbf{R}, \mathbf{b}) \in \mathrm{SE}(3)$ can be expressed as

$$\mathrm{Ad}_{\mathbf{g}} = \begin{bmatrix} \mathbf{R} & \mathbf{b}^\wedge\mathbf{R} \\ 0_{3\times3} & \mathbf{R} \end{bmatrix} \in \mathbb{R}^{6\times6}. \tag{4}$$

The Lie algebra corresponding to $\xi = [\mathbf{v}^T \ \omega^T]^T$ can be expressed as

$$\boldsymbol{\xi}^{\wedge} = \begin{bmatrix} \boldsymbol{\omega}^{\wedge} & \mathbf{v} \\ 0_{1\times 3} & 0 \end{bmatrix} \in \varsigma e(3). \tag{5}$$

The adjoint matrix of $\xi^{\wedge} = \xi^{\wedge}(\omega, \mathbf{v}) \in \varsigma e(3)$ can be expressed as

$$\mathrm{ad}_{\boldsymbol{\xi}} = \begin{bmatrix} \boldsymbol{\omega}^{\wedge} & \mathbf{v}^{\wedge} \\ 0_{3\times 3} & \boldsymbol{\omega}^{\wedge} \end{bmatrix} \in \mathbb{R}^{6\times 6}. \tag{6}$$

The co-adjoint matrix of $\xi^{\wedge} = \xi^{\wedge}(\omega, \mathbf{v}) \in \varsigma e(3)$ can be expressed as

$$\mathrm{ad}^{*}_{\boldsymbol{\xi}} = \left(\mathrm{ad}_{\boldsymbol{\xi}}\right)^{T} = \begin{bmatrix} -\boldsymbol{\omega}^{\wedge} & 0_{3\times 3} \\ -\mathbf{v}^{\wedge} & -\boldsymbol{\omega}^{\wedge} \end{bmatrix} \in \mathbb{R}^{6\times 6}. \tag{7}$$

The coupled kinematics of the spacecraft in the ECI coordinate system can be expressed as

$$\begin{cases} \dot{\mathrm{R}} = \mathbf{R}\boldsymbol{\omega}^{\wedge}, \\ \dot{\mathrm{b}} = \mathbf{R}\mathbf{v}. \end{cases} \tag{8}$$

The above kinematics equation can be simplified as follows:

$$\dot{\mathrm{g}} = \mathbf{g}\boldsymbol{\xi}^{\wedge}. \tag{9}$$

The coupled dynamics of the spacecraft in the body-fixed frame can be expressed as

$$\begin{cases} m\dot{\mathrm{v}} + m\boldsymbol{\omega}\times\mathbf{v} = \mathbf{F}_g(\mathbf{b},\mathbf{R}) + m\mathbf{R}^T\mathbf{a}_{J_2}(\mathbf{b}) + \mathbf{F}_c(\mathbf{b},\mathbf{R},\mathbf{v},\boldsymbol{\omega}) + \mathbf{F}_d, \\ \mathbf{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}\times\mathbf{J}\boldsymbol{\omega} = \mathbf{M}_g(\mathbf{b},\mathbf{R}) + \mathbf{M}_c(\mathbf{b},\mathbf{R},\mathbf{v},\boldsymbol{\omega}) + \mathbf{M}_d. \end{cases} \tag{10}$$

where m and $\mathbf{J}$ are the mass and moment of inertia of the spacecraft, respectively; $\mathbf{F}_g$ and $\mathbf{M}_g$ are the gravity gradient force and the gravity gradient moment of the spacecraft, respectively; $\mathbf{a}_{J_2}$ is the perturbation caused by the Earth's oblateness; $\mathbf{F}_c$ and $\mathbf{M}_c$ are the control force and control torque of the spacecraft, respectively; and $\mathbf{F}_d$ and $\mathbf{M}_d$ are the unknown external force and external torque caused by radiation pressure, atmosphere drag, and other bounded uncertain disturbances, respectively. The specific forms of $\mathbf{F}_g$, $\mathbf{M}_g$, and $\mathbf{a}_{J_2}$ are as follows:

$$\begin{aligned} \mathbf{F}_g &= -\frac{m\mu}{\|\mathbf{b}\|^3}\left[\mathbf{I}_3 + \frac{3}{m\|\mathbf{b}\|^2}\left(\frac{1}{2}tr(\mathbf{J})\mathbf{I}_3 + \mathbf{J} - \frac{5\mathbf{b}^T\mathbf{R}\mathbf{J}\mathbf{R}^T\mathbf{b}}{2\|\mathbf{b}\|^2}\mathbf{I}_3\right)\right]\mathbf{R}^T\mathbf{b}, \\ \mathbf{a}_{J_2} &= -\frac{3\mu J_2\mathbf{R}_e^2}{2\|\mathbf{b}\|^5}\left[\mathbf{b}_x\left(1-\frac{5\mathbf{b}_z^2}{\|\mathbf{b}\|^2}\right)\ \mathbf{b}_y\left(1-\frac{5\mathbf{b}_z^2}{\|\mathbf{b}\|^2}\right)\ \mathbf{b}_z\left(3-\frac{5\mathbf{b}_z^2}{\|\mathbf{b}\|^2}\right)\right]^T, \\ \mathbf{M}_g &= \frac{3\mu}{\|\mathbf{b}\|^5}\left(\mathbf{R}^T\mathbf{b}\right)^{\wedge}\mathbf{J}\mathbf{R}^T\mathbf{b}, \end{aligned} \tag{11}$$

where $\mu = 398,600.44\,\mathrm{km}^3\mathrm{s}^{-2}$ is the gravitational constant of the Earth, $J_2 = 0.00108263$ is the perturbation caused by the Earth's oblateness, and $\mathbf{R}_e = 6378.14\,\mathrm{km}$ is the equatorial radius of the Earth.

Then, the coupling dynamics of the spacecraft can be expressed in a compact form as follows:

$$\Xi = \mathrm{ad}^{*}_{\boldsymbol{\xi}}\,\boldsymbol{\xi} + \Gamma_g + \Gamma_c + \Gamma_d, \tag{12}$$

where $\Xi = \mathrm{diag}(m\mathbf{I}_3, \mathbf{J})$ is the unified inertia matrix of spacecraft, $\Gamma_c = \begin{bmatrix} \mathbf{F}_c^T & \mathbf{M}_c^T \end{bmatrix}^T$ is the unified control input vector, and $\Gamma_g = \begin{bmatrix} (\mathbf{F}_g + m\mathbf{R}^T\mathbf{a}_{J_2})^T & \mathbf{M}_q^T \end{bmatrix}^T$ is the unified input vector related to gravity.

Thus, combining equations (9) and (12), the coupled kinematics and dynamics of the spacecraft can be expressed compactly as

$$\begin{cases} \dot{\mathrm{g}} = \mathbf{g}\boldsymbol{\xi}^{\wedge}, \\ \Xi\dot{\boldsymbol{\xi}} = \mathrm{ad}^{*}_{\boldsymbol{\xi}}\Xi\,\boldsymbol{\xi} + \Gamma_g + \Gamma_c + \Gamma_d. \end{cases} \tag{13}$$

Next, based on the above equations, the coupled relative motion tracking error dynamics will be derived. Let $\mathbf{g}_o$ be the actual pose configuration of the leader spacecraft, which the leader spacecraft can be real or virtual; $\mathbf{g}_b$ be the actual pose configuration of the follower spacecraft. Then the actual relative pose configuration between the leader-follower spacecraft is as follows:

$$\mathbf{h} = \mathbf{g}_o^{-1}\mathbf{g}_b. \tag{14}$$

Let $\mathbf{g}_d$ be the desired pose configuration of the leader spacecraft, then the desired relative pose configuration between the leader-follower spacecraft is

$$\mathbf{h}_d = \mathbf{g}_o^{-1}\mathbf{g}_d. \tag{15}$$

Thus the pose configuration tracking error is as follows:

$$\mathbf{h}_e = \mathbf{h}_d^{-1}\mathbf{h} = \mathbf{g}_d^{-1}\mathbf{g}_o\mathbf{g}_o^{-1}\mathbf{g}_b = \mathbf{g}_d^{-1}\mathbf{g}_b. \tag{16}$$

In general, the desired relative pose configuration is a constant value, and the desired relative linear velocity and angular velocity are zero, which means that the follower spacecraft and the leader spacecraft keep relatively stationary in a fixed configuration.

The configuration tracking error of the follower spacecraft can be expressed by exponential coordinates as

$$\boldsymbol{\eta}_e = \begin{bmatrix} \boldsymbol{\rho}_e \\ \boldsymbol{\varphi}_e \end{bmatrix} \in \mathbb{R}^6. \tag{17}$$

The velocity tracking error of the follower spacecraft can be expressed by exponential coordinates as

$$\boldsymbol{\xi}_e = \begin{bmatrix} \mathbf{v}_e \\ \boldsymbol{\omega}_e \end{bmatrix} \in \mathbb{R}^6. \tag{18}$$

Then $\mathbf{h}_e$ can be expressed on SE(3) as

$$\mathbf{h}_e = \begin{bmatrix} \mathbf{R}_e & \mathbf{b}_e \\ 0_{1\times 3} & 1 \end{bmatrix}. \tag{19}$$

Through the logarithm mapping between the Lie group SE(3) and the Lie algebra $\varsigma e(3)$, we can get the following results:

$$\boldsymbol{\eta}_e = \left(\log\left(\mathbf{h}_e\right)\right)^{\vee},$$
$$\log\left(\mathbf{h}_e\right) = \begin{bmatrix} \boldsymbol{\varphi}_e^{\wedge} & \boldsymbol{\rho}_e \\ 0_{1\times 3} & 0 \end{bmatrix}, \tag{20}$$

where ρ_e is the position tracking error and φ_e is the attitude tracking error, which are expressed as follows:

$$\boldsymbol{\rho}_e = \mathbf{S}^{-1}\left(\boldsymbol{\varphi}_e\right)\mathbf{b}_e,$$
$$\boldsymbol{\varphi}_e^{\wedge} = \begin{cases} 0, & \theta = 0, \\ \dfrac{\theta}{\sin\theta}\left(\mathbf{R}_e - \mathbf{R}_e^T\right), & \theta \in (-\pi, \pi), \theta \neq 0, \end{cases} \tag{21}$$
$$\mathbf{S}^{-1}\left(\boldsymbol{\varphi}_e\right) = \mathbf{I} - \frac{1}{2}\boldsymbol{\varphi}_e^{\wedge} + \frac{1}{\theta^2}\left(1 - \frac{\theta\sin\theta}{2\left(1-\cos\theta\right)}\right)\left(\phi_e^{\wedge}\right)^2,$$

where $\theta = \arccos\left((1/2)\left[\mathrm{tr}\left(\mathbf{R}_e\right) - 1\right]\right)$, which is the norm of φ_e and corresponds to the principal rotation angle. When $\theta = 0$, it is injective; when $|\theta| < \pi$, it is bijective.

According to the relationship between Lie group and Lie algebra, it can be deduced that when the desired relative velocity is zero, the expressions of the relative velocity error and relative acceleration error of the follower spacecraft are

$$\boldsymbol{\xi}_e = \boldsymbol{\xi}_b - \mathrm{Ad}_{\mathbf{h}_e^{-1}}\boldsymbol{\xi}_d = \boldsymbol{\xi}_b - \mathrm{Ad}_{\mathbf{h}^{-1}}\boldsymbol{\xi}_o,$$
$$\dot{\xi}_e = \dot{\xi}_b + \mathrm{ad}_{\xi_r}\mathrm{Ad}_{\mathbf{h}^{-1}}\boldsymbol{\xi}_o - \mathrm{Ad}_{\mathbf{h}^{-1}}\dot{\xi}_o. \tag{22}$$

Then the coupled relative motion tracking error kinematics as given in [32] is

$$\dot{\eta}_e = \mathbf{G}\left(\boldsymbol{\eta}_e\right)\boldsymbol{\xi}_e, \tag{23}$$

where $\mathbf{G}(\eta_e)$ is expressed as a block-triangular matrix:

$$\mathbf{G}\left(\boldsymbol{\eta}_e\right) = \begin{bmatrix} \mathbf{A}\left(\boldsymbol{\varphi}_e\right) & \mathbf{T}\left(\boldsymbol{\varphi}_e, \boldsymbol{\rho}_e\right) \\ 0_{3\times 3} & \mathbf{A}\left(\boldsymbol{\varphi}_e\right) \end{bmatrix}, \tag{24}$$

where

$$\mathbf{A}\left(\boldsymbol{\varphi}_e\right) = \mathbf{I}_3 + \frac{1}{2}\boldsymbol{\varphi}_e^{\wedge} + \frac{1}{\theta^2}\left(1 - \frac{(1+\cos\theta)\theta}{2\sin\theta}\right)\left(\boldsymbol{\varphi}_e^{\wedge}\right)^2,$$
$$\mathbf{T}\left(\boldsymbol{\varphi}_e, \boldsymbol{\rho}_e\right) = \frac{1}{2}\left(\mathbf{S}\left(\boldsymbol{\varphi}_e\right)\rho_e\right)^{\wedge}\mathbf{A}\left(\boldsymbol{\varphi}_e\right) + \frac{1}{\theta^2}\left(1 - \frac{(1+\cos\theta)\theta}{2\sin\theta}\right)\left[\boldsymbol{\varphi}_e\boldsymbol{\rho}_e^T + \left(\boldsymbol{\varphi}_e^T\boldsymbol{\rho}_e\right)\mathbf{A}\left(\boldsymbol{\varphi}_e\right)\right] \tag{25}$$
$$- \frac{(1+\cos\theta)(\theta-\sin\theta)}{2\theta\sin^2\theta}\mathbf{S}\left(\boldsymbol{\varphi}_e\right)\boldsymbol{\rho}_e\boldsymbol{\varphi}_e^T + \left[\frac{(1+\cos\theta)(\theta-\sin\theta)}{2\theta^3\sin^2\theta} - \frac{2}{\theta^4}\right]\boldsymbol{\varphi}_e^T\boldsymbol{\rho}_e\boldsymbol{\varphi}_e\boldsymbol{\varphi}_e^T.$$

Taking equation (12) into equation (22) yields

$$\begin{aligned} \Xi\dot{\xi}_e &= \Xi\left(\dot{\xi}_b + \mathrm{ad}_{\xi_r}\mathrm{Ad}_{\mathbf{h}^{-1}}\boldsymbol{\xi}_o - \mathrm{Ad}_{\mathbf{h}^{-1}}\dot{\xi}_o\right) \\ &= \mathrm{ad}_{\boldsymbol{\xi}_b}^{*}\Xi\boldsymbol{\xi}_b + \Gamma_g + \Gamma_c + \Gamma_d + \Xi\left(\mathrm{ad}_{\xi_r}\mathrm{Ad}_{\mathbf{h}^{-1}}\boldsymbol{\xi}_o - \mathrm{Ad}_{\mathbf{h}^{-1}}\dot{\xi}_o\right). \end{aligned} \tag{26}$$

However, in actual space missions, the inertia matrix Ξ is uncertain due to fuel consumption and external disturbances, so the actual inertia matrix Ξ can be expressed as

$$\Xi = \Xi_0 + \Delta\Xi, \tag{27}$$

where Ξ_0 is the nominal part and $\Delta\Xi$ is the uncertainty part. Then the inverse of the inertia matrix can be expressed as

$$\Xi^{-1} = \left(\Xi_0 + \Delta\Xi\right)^{-1} = \Xi_0^{-1} + \Delta\widetilde{\Xi},$$
$$\Delta\widetilde{\Xi} = -\Xi_0^{-1}\Delta\Xi\left(\mathbf{I}_6 + \Xi_0^{-1}\Delta\Xi\right)\Xi_0^{-1}. \tag{28}$$

Therefore, (26) can be rewritten as

$$\begin{aligned} \dot{\xi}_e &= \mathbf{H} + \boldsymbol{\Gamma}_0^{-1}\Gamma_c + \Delta\mathbf{d}, \\ \mathbf{H} &= \boldsymbol{\Gamma}_0^{-1}\mathrm{ad}_{\boldsymbol{\xi}_b}^{*}\boldsymbol{\Gamma}_0\boldsymbol{\xi}_b + \mathrm{ad}_{\xi_b}\mathrm{Ad}_{\mathbf{h}^{-1}}\boldsymbol{\xi}_o - \mathrm{Ad}_{\mathbf{h}^{-1}}\dot{\xi}_o + \boldsymbol{\Gamma}_0^{-1}\Gamma_g, \\ \Delta\mathbf{d} &= \Delta\widetilde{\Xi}\left(\mathrm{ad}_{\boldsymbol{\xi}_b}^{*}\left(\boldsymbol{\Gamma}_0 + \Delta\boldsymbol{\Gamma}\right)\boldsymbol{\xi}_b + \Gamma_g + \Gamma_c\right) + \boldsymbol{\Gamma}_0^{-1}\mathrm{ad}_{\boldsymbol{\xi}_b}^{*}\Delta\boldsymbol{\Gamma}\boldsymbol{\xi}_b + \left(\boldsymbol{\Gamma}_0^{-1} + \Delta\widetilde{\Xi}\right)\Gamma_d, \end{aligned} \tag{29}$$

where $\mathbf{H}$ is a known deterministic term of the system and $\Delta\mathbf{d}$ is the lumped disturbances, including uncertainties and external disturbances. Then the coupling model of relative motion spacecraft can be expressed as follows:

$$\begin{cases} \dot{\eta}_e = \mathbf{G}\left(\boldsymbol{\eta}_e\right)\boldsymbol{\xi}_e, \\ \dot{\xi}_e = \mathbf{H} + \boldsymbol{\Gamma}_0^{-1}\Gamma_c + \Delta\mathbf{d}. \end{cases} \tag{30}$$

2.3. Actuator Configuration with Faults. In this paper, the actuator of attitude control is reaction flywheel, and the actuator of orbit control is thruster.4 reaction flywheels and 4 pairs of thrusters are used. Then the control vector can be expressed as

$$\Gamma_c = \begin{bmatrix} \mathbf{F}_c \\ \mathbf{M}_c \end{bmatrix} = \mathbf{D}\mathbf{u}, \tag{31}$$

where $\mathbf{D} \in \mathbb{R}^{6\times 12}$ is the configuration matrix, $\mathbf{u} = [u_1, u_2, u_3, \ldots, u_{12}]^T$ is the actual control vector, and u_i $(i = 1, 2, 3, \ldots, 12)$ is the torque or force that each flywheel or thruster can provide.

The four reaction flywheels adopt the traditional installation method of three orthogonal and one inclined

installation, and the configuration structure is shown in Figure 2.

The control torque distribution matrix of the reaction flywheel is

$$\mathbf{D}_1 = \begin{bmatrix} 1 & 0 & 0 & \frac{\sqrt{3}}{3} \\ 0 & 1 & 0 & \frac{\sqrt{3}}{3} \\ 0 & 0 & 1 & \frac{\sqrt{3}}{3} \end{bmatrix}. \tag{32}$$

Eight thrusters are installed symmetrically at the middle point of each edge of the cube in pairs. The installation mode of thrust passing through the center of mass is adopted. The configuration structure is shown in Figure 3.

The control force distribution matrix of the thruster is

$$\mathbf{D}_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{33}$$

Then the attitude and position coupling integrated control distribution matrix $\mathbf{D}$ is

$$\mathbf{D} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \frac{\sqrt{3}}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{\sqrt{3}}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{\sqrt{3}}{3} \end{bmatrix}. \tag{34}$$

According to the cause of the faults of the actuator, the failure of the actuator can be divided into the following categories: stuck, loose, saturated, damaged, or invalid. The abovementioned faults types can be unified as follows:

$$\mathbf{u}(t) = \mathbf{E}\mathbf{u}_c(t) + (\mathbf{I} - \mathbf{E})\overline{\mathbf{u}}, \tag{35}$$

where $\mathbf{u}_c(t)$ is the control command of the actuator; $\overline{u}$ is the stuck fault of the actuator with bounded value and satisfies the constraint $|\overline{u}_i| \le \min\{u_{i\max}, |u_{i\min}|\}$; $\overline{u}_i = \{u_{i\max}, u_{i\min}\}$ indicates that the i-th thruster is in saturated state; $\overline{u}_i = 0$ indicates that the i-th thruster is in loose position. $\mathbf{E} = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n)$ is the actuator effectiveness matrix and satisfies the constraint $0 \le \sigma_i \le 1$. $\sigma_i = 0$ denotes that the i th actuator does not supply any control output; $\sigma_i = 1$ means there is no fault for the i th actuator; and $0 < \sigma_i < 1$ implies the i th actuator has partially lost its effectiveness [38].

Taking equations (31) and (35) into equation (30), the integrated dynamic equation of relative motion spacecraft considering actuator fault can be expressed as follows:

$$\dot{\xi}_e = \mathbf{H} + \boldsymbol{\Gamma}_0^{-1}\mathbf{D}\mathbf{E}\mathbf{u}_c + \Delta\tilde{\mathrm{d}}, \tag{36}$$

where $\Delta\tilde{d} = \Xi_0^{-1}[\mathbf{D}(\mathbf{I} - \mathbf{E})\overline{u}] + \Delta\mathbf{d}$.

3. Adaptive Fuzzy Modified Fixed-Time Fault-Tolerant Controller Design and Stability Analysis

In this part, our goal is to design a fault-tolerant controller on the relative coupled dynamics so that the configuration of the spacecraft can converge to the desired state in the presence of model uncertainties and external disturbances and actuator faults in fixed time.

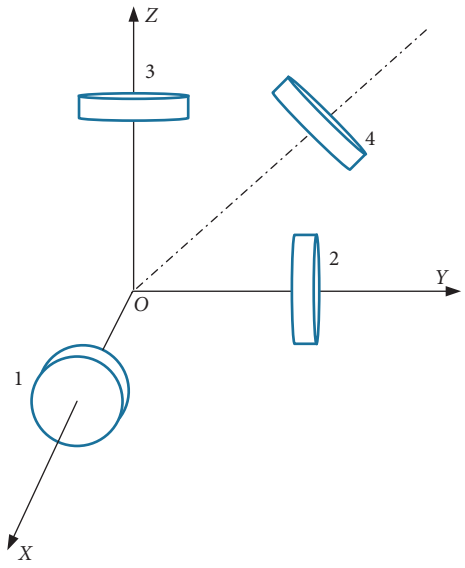


Figure 2: Configuration structure of flywheels.

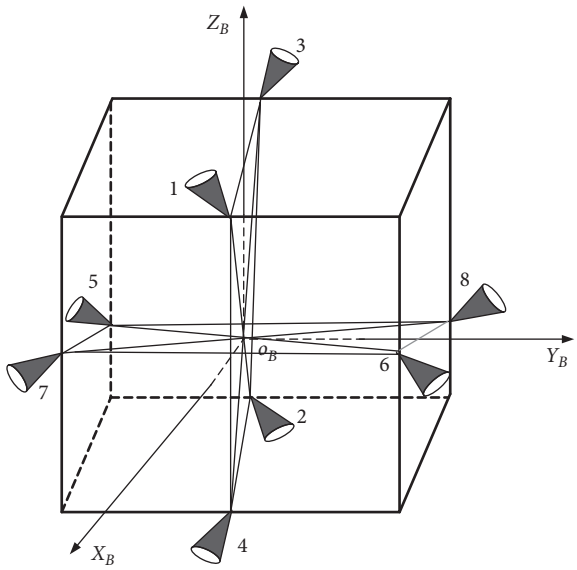


Figure 3: Configuration structure of thrusters.

3.1. Introduction of Fuzzy Approximation Technique. The spacecraft exhibits strong nonlinearity due to the influence of lumped disturbances, which will affect the performance of the controller. So it is critical to approximate the lumped disturbances with high accuracy, and fuzzy logic system (FLS) is an effective way to realize this objective. The fuzzy approximation method can make full use of fuzzy linguistic information to approximate any nonlinear continuous function. It has a good effect in fitting nonlinear function. It can approach nonlinear continuous function with arbitrary precision. The structure and basic theory of fuzzy approximation system are given below [32].

$\mathbf{X} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ is the input variable of the FLS, and M fuzzy rules are designed for each component of the input variable, then the whole system has n^M fuzzy rules, and the specific expression of each fuzzy rule is

$$\text{IF} x_1 \text{is} A_1^l \text{ and} \ldots \text{and} x_N \text{ is } A_N^l \text{ THEN } z \text{ is } B^l, \tag{37}$$

where $l_i = 1, \ldots, M$ is the number of fuzzy rules for each input variable x_k, z is the output of the fuzzy system, A_k^l is the fuzzy set of system input variables, and B^l is the fuzzy set of system output.

If the fuzzy system adopts singleton fuzzifier, center-average defuzzifier, and product inference engine, the output of fuzzy approximation system can be obtained as follows:

$$z = \frac{\overline{z}^l\left(\prod_{k=1}^N \mu_{A_k^l}(x_k)\right)}{\sum_{l=1}^M\left(\prod_{k=1}^N \mu_{A_k^l}(x_k)\right)}, \tag{38}$$

where $\mu_{A_k^l}(x_k)$ is the membership function corresponding to the input variable x_k; in this paper, Gauss membership function is used with the form

$$\mu_{A_k^l}(x_k) = a_k^l \exp\left(-\frac{1}{2}\left(\frac{x_k - \overline{x}_k^l}{b_k^l}\right)^2\right), \tag{39}$$

where $a_k^l, \overline{x}_k^l$, and b_k^l are all positive real parameters with $0 < a_k^l \leq 1$. $\overline{x}_k^l$ is the abscissa corresponding to the membership function $\mu_{A_k^l}(x_k)$ when the maximum value is 1.

Let $\mathbf{W} = (\overline{z}^1, \overline{z}^2, \ldots, \overline{z}^M)$, then equation (38) can be rewritten as follows:

$$z = \mathbf{W}\boldsymbol{\beta}, \tag{40}$$

where β is the basis function, which can be expressed as

$$\boldsymbol{\beta}(\mathbf{X}) = \frac{\prod_{k=1}^N \mu_{A_k^l}(x_k)}{\sum_{l=1}^M\left(\prod_{k=1}^N \mu_{A_k^l}(x_k)\right)}. \tag{41}$$

Based on the above introduction, the total external disturbances of the follower spacecraft can be estimated by the fuzzy approximation as

$$\Delta\widetilde{d} = \mathbf{W}^*\boldsymbol{\beta}(\mathbf{X}) + \boldsymbol{\varepsilon}, \tag{42}$$

where $\mathbf{W}^*$ is the optimal weight matrix and ε is the bounded approximation error of FLS. Let $\widehat{W}$ be the estimation of $\mathbf{W}^*$. The sliding surface $\mathbf{S}$ is the input variable of $\beta(\mathbf{X})$, and then the estimated value $\Delta\widehat{\widetilde{d}}$ of the total external disturbances $\Delta\widetilde{d}$ of the spacecraft can be expressed as

$$\Delta\widehat{\widetilde{d}} = \widehat{W}\boldsymbol{\beta}(\mathbf{X}) + \boldsymbol{\varepsilon}. \tag{43}$$

Then the estimation error of the optimal weight matrix is

$$\widetilde{W} = \widehat{W} - \mathbf{W}^*. \tag{44}$$

In order to design and analyze the controller, some assumptions are given below.

Assumption 1. The output of FLS is bounded, and the estimated value of the external total disturbances is bounded such that

$$\|\Delta\tilde{\tilde{d}}\| \le d_m, \tag{45}$$

where d_m is a positive constant.

Assumption 2. The approximation error of FLS is bounded such that

$$\|\boldsymbol{\varepsilon}\| \le \varepsilon_m, \tag{46}$$

where ε_m is a positive constant.

Assumption 3. The optimal weight matrix of FLS is bounded such that

$$\mathrm{tr}\left(\tilde{W}^T\tilde{W}\right) \le W_m, \tag{47}$$

where W_m is a positive constant.

Assumption 4. The faults of the actuators satisfy the constraint $\mathrm{rank}(\mathbf{DE}) = 6$, and this means that the redundant actuators can still combine enough control output to complete the given goal.

Remark 1. Because the mass, moment of inertia, fault amplitude, input variables of FLS, and external disturbance of the system are bounded, so Assumption 1 is reasonable; Assumptions 2 and 3 have the property that fuzzy approximation system can fit any nonlinear continuous function, and Assumption 4 does not consider underactuated system, so it is also reasonable.

3.2. Controller Design. In order to achieve the control goal of modified fixed time stability, 3 sliding surface forms are proposed as follows.

Firstly, the finite-time terminal sliding mode is denoted as

$$\mathbf{S} = \boldsymbol{\xi}_e + \mathbf{C}_1\boldsymbol{\eta}_e + \mathbf{C}_2\mathrm{sig}^{\alpha}(\boldsymbol{\eta}_e). \tag{48}$$

Then, the fixed-time terminal sliding mode is denoted as

$$\mathbf{S} = \boldsymbol{\xi}_e + \mathbf{C}_1\mathrm{sig}^{\alpha_1}(\boldsymbol{\eta}_e) + \mathbf{C}_2\mathrm{sig}^{\alpha_2}(\boldsymbol{\eta}_e). \tag{49}$$

Finally, the modified fixed-time terminal sliding mode is denoted as

$$\mathbf{S} = \boldsymbol{\xi}_e + \mathbf{C}_1\mathrm{sig}^{(1/2)+(1/2)\alpha_1+\left((1/2)\alpha_1-(1/2)\right)\mathrm{sgn}\left(|\boldsymbol{\eta}_e|-1\right)}(\boldsymbol{\eta}_e) + \mathbf{C}_2\mathrm{sig}^{\alpha_2}(\boldsymbol{\eta}_e), \tag{50}$$

where $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{6\times 6}$ are both positive definite diagonal matrices, $\alpha \in ((1/2), 1)$, and $\alpha_1 \in (1, +\infty)$, $\alpha_2 \in ((1/2), 1)$.

Remark 2. In fact, equation (50) is developed on the basis of equations (49) and (48), and it can be classified and discussed as follows:

(1) When $|\eta_e| < 1$, equation (50) can be expressed as $\mathbf{S} = \xi_e + \mathbf{C}_1|\eta_e|\mathrm{sgn}(\eta_e) + \mathbf{C}_2\mathrm{sig}^{\alpha}(\eta_e)$, and it has a similar form to equation (48).

(2) When $|\eta_e| = 1$, equation (50) can be expressed as $\mathbf{S} = \xi_e + \mathbf{C}_1\mathrm{sig}^{(1/2)+(1/2)\alpha_1}(\eta_e) + \mathbf{C}_2\mathrm{sig}^{\alpha_2}(\eta_e)$, and it has a similar form to equation (49).

(3) When $|\eta_e| > 1$, equation (50) can be expressed as $\mathbf{S} = \xi_e + \mathbf{C}_1\mathrm{sig}^{\alpha_1}(\eta_e) + \mathbf{C}_2\mathrm{sig}^{\alpha_2}(\eta_e)$, and it has the same form to equation (49).

From the above analysis and discussion, we can see that the modified fixed-time terminal sliding mode is a combination of the finite-time terminal sliding mode and the fixed-time terminal sliding mode under different conditions. To guarantee that the relative motion can spacecraft converge to the desired state in the expected time even in the case of actuator faults, the corresponding three kinds of adaptive fuzzy sliding mode controllers are designed as follows:

(1) Corresponding to (48), adaptive fuzzy finite-time fault-tolerant controller (AF-Finite) is given as

$$\mathbf{u}_c = -(\mathbf{DE})^{\dagger}\left\{\boldsymbol{\Gamma}_0\left(\mathbf{H} + \mathbf{C}_1\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e + \alpha\mathbf{C}_2\mathrm{diag}\left(|\boldsymbol{\eta}_e|^{\alpha-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e + \hat{W}\boldsymbol{\beta}(\mathbf{X})\right) + \mathbf{K}_1\mathbf{S} + \mathbf{K}_2\mathrm{sig}^{\alpha}(\mathbf{S})\right\}. \tag{51}$$

(2) Corresponding to (49), adaptive fuzzy fixed-time fault-tolerant controller (AF-Fixed) is given as

$$\mathbf{u}_c = -(\mathbf{DE})^{\dagger}\left\{\boldsymbol{\Gamma}_0\left(\begin{array}{c}\mathbf{H} + \alpha_1\mathbf{C}_1\mathrm{diag}\left(|\boldsymbol{\eta}_e|^{\alpha_1-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e + \\ \alpha_2\mathbf{C}_2\mathrm{diag}\left(|\boldsymbol{\eta}_e|^{\alpha_2-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e + \hat{W}\boldsymbol{\beta}(\mathbf{X})\end{array}\right) + \mathbf{K}_1\mathrm{sig}^{\alpha_1}(\mathbf{S}) + \mathbf{K}_2\mathrm{sig}^{\alpha_2}(\mathbf{S})\right\}. \tag{52}$$

(3) Corresponding to (50), adaptive fuzzy modified fixed-time fault-tolerant controller (AF-MFixed) is given as

$$\mathbf{u}_c = -(\mathbf{DE})^{\dagger}\left\{\boldsymbol{\Gamma}_0\left(\begin{array}{c}\left(\dfrac{1+\alpha_1}{2}+\dfrac{\alpha_1-1}{2}\operatorname{sgn}\left(|\boldsymbol{\eta}_e|-1\right)\right)\mathbf{C}_1\operatorname{diag}\left(|\boldsymbol{\eta}_e|^{\left((1+\alpha_1/2)+(\alpha_1-1/2)\operatorname{sgn}\left(|\boldsymbol{\eta}_e|-1\right)\right)-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e \\ +\alpha_2\mathbf{C}_2\operatorname{diag}\left(|\boldsymbol{\eta}_e|^{\alpha_2-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e+\widehat{W}\boldsymbol{\beta}(\mathbf{X})+\mathbf{H} \\ +\mathbf{K}_1\operatorname{sig}^{(1+\alpha_1/2)+(\alpha_1-1/2)\operatorname{sgn}(|S|-1)}(\mathbf{S})+\mathbf{K}_2\operatorname{sig}^{\alpha_2}(\mathbf{S})\end{array}\right)\right\}, \tag{53}$$

where $(\mathbf{DE})^{\dagger} = (\mathbf{DE})^T[(\mathbf{DE})(\mathbf{DE})^T]^{-1}$ is the pseudoinverse of matrix $\mathbf{DE}$; from Assumption 4, we know that $\mathbf{DE}(\mathbf{DE})^T$ is full rank, so its pseudoinverse exists; $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{6\times 6}$ are both positive definite diagonal matrices.

Then the adaptive update law of the optimal weight matrix $\widehat{W}$ is given by

$$\dot{\widehat{W}} = \dot{\widetilde{W}} = \gamma\boldsymbol{\Gamma}_0^T\mathbf{S}\boldsymbol{\beta}^T(\mathbf{X}), \tag{54}$$

where $\gamma > 0$ is an auxiliary parameter independent of control.

3.3. Stability Analysis. In this part, we will take the stability proof of AF-finite fault-tolerant controller as an example for stability analysis, and the other two (AF-Fixed and AF-MFixed) stability analysis methods are the same as is. Some lemmas are given before the stability analysis.

Lemma 1 (see [6]). *Assuming that $V(\mathbf{x})$: $\mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous positive definite function and satisfies the following differential inequality*

$$\dot{V}(\mathbf{x}) + \rho_1 V(\mathbf{x}) + \rho_2 V^{\upsilon}(\mathbf{x}) \le 0, \quad \forall t > 0, \tag{55}$$

where $\rho_1 > 0, \rho_2 > 0, \upsilon \in (0,1)$, then $V(\mathbf{x})$ can converge to the equilibrium point in finite time, and the finite time T satisfies the following constraints:

$$T \le \frac{1}{\rho_1(1-\upsilon)}\ln\frac{\rho_1 V^{1-\upsilon}(\mathbf{x}_0)+\rho_2}{\rho_2}. \tag{56}$$

Lemma 2 (see [38]). *Assuming that $V(\mathbf{x})$: $\mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous positive definite function and satisfies the following differential inequality:*

$$\dot{V}(\mathbf{x}) + \rho_1 V^{\upsilon_1}(\mathbf{x}) + \rho_2 V^{\upsilon_2}(\mathbf{x}) \le 0, \quad \forall t > 0, \tag{57}$$

where $\rho_1 > 0, \rho_2 > 0, \upsilon_1 > 1, \upsilon_2 \in (0,1)$, then $V(\mathbf{x})$ can converge to the equilibrium point in fixed time, and the fixed time T satisfies the following constraints:

$$T \le \frac{1}{\rho_1(\upsilon_1-1)} + \frac{1}{\rho_2(1-\upsilon_2)}. \tag{58}$$

Lemma 3 (see [3]). *Assuming that $V(\mathbf{x})$: $\mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous positive definite function and satisfies the following differential inequality*

$$\dot{V}(\mathbf{x}) + \rho_1 V^{(1/2)+(1/2)\upsilon_1+\left((1/2)\upsilon_1-(1/2)\right)\operatorname{sgn}(V(x)-1)}(\mathbf{x}) + \rho_2 V^{\upsilon_2}(\mathbf{x}) \le 0, \quad \forall t > 0, \tag{59}$$

where $\rho_1 > 0, \rho_2 > 0, \upsilon_1 > 1, \upsilon_2 \in (0,1)$, then $V(\mathbf{x})$ can converge to the equilibrium point in modified fixed time, and the modified fixed time T satisfies the following constraints:

$$T \le \frac{1}{\rho_1(\upsilon_1-1)} + \frac{1}{\rho_1(1-\upsilon_2)}\ln\left(1+\frac{\rho_1}{\rho_2}\right). \tag{60}$$

Because the inequality $\ln(1+(\rho_1/\rho_2)) \le (\rho_1/\rho_2)$ holds, the convergence time in Lemma 3 is shorter than that in Lemma 2.

Lemma 4 (see [32]). The eigenvalues of matrix $\mathbf{G}(\eta_e)$ are all positive.

Next, the Lyapunov method will be used to prove the reachability of sliding mode variables and the convergence of the system states.

Theorem 1. *When the nonlinear system equation (48) reaches the sliding mode surface $\mathbf{S} = 0$, the state η_e, ξ_e of the system can converge to the equilibrium point in a finite time.*

Proof. When equation (48) reaches the sliding mode surface $\mathbf{S} = 0$, such that

$$\mathbf{S} = \boldsymbol{\xi}_e + \mathbf{C}_1\boldsymbol{\eta}_e + \mathbf{C}_2\operatorname{sig}^{\alpha}(\boldsymbol{\eta}_e) = 0, \tag{61}$$

then we have

$$\boldsymbol{\xi}_e = -\mathbf{C}_1\boldsymbol{\eta}_e - \mathbf{C}_2\operatorname{sig}^{\alpha}(\boldsymbol{\eta}_e). \tag{62}$$

A candidate Lyapunov function is selected as follows:

$$V = \frac{1}{2}\boldsymbol{\eta}_e^T\boldsymbol{\eta}_e. \tag{63}$$

Taking the derivative of V with respect to time yields:

$$\begin{aligned}\dot{V} &= \boldsymbol{\eta}_e^T \dot{\eta}_e = \boldsymbol{\eta}_e^T \mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e = \boldsymbol{\eta}_e^T \mathbf{G}(\boldsymbol{\eta}_e)\left[-\mathbf{C}_1\boldsymbol{\eta}_e - \mathbf{C}_2 \mathrm{sig}^{\alpha}(\boldsymbol{\eta}_e)\right] \\ &\le -\lambda_{\min}\left(\mathbf{G}(\boldsymbol{\eta}_e)\mathbf{C}_1\right)\left\|\boldsymbol{\eta}_e\right\|_2^2 - \lambda_{\min}\left(\mathbf{G}(\boldsymbol{\eta}_e)\mathbf{C}_2\right)\left\|\boldsymbol{\eta}_e\right\|^{1+\alpha} \\ &\le -2\lambda_{\min}\left(\mathbf{G}(\boldsymbol{\eta}_e)\mathbf{C}_1\right)V - 2^{(1+\alpha/2)}\lambda_{\min}\left(\mathbf{G}(\boldsymbol{\eta}_e)\mathbf{C}_2\right)V^{(1+\alpha/2)} \\ &= -a_1 V - a_2 V^{(1+\alpha/2)},\end{aligned} \tag{64}$$

where $a_1 = 2\lambda_{\min}(\mathbf{G}(\eta_e)\mathbf{C}_1)$, $a_2 = 2^{(1+\alpha/2)}\lambda_{\min}(\mathbf{G}(\eta_e)\mathbf{C}_2)$, by using Lemma 4 we know that the eigenvalues of matrix $\mathbf{G}(\eta_e)$ are all positive; in addition, $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{6\times 6}$ are both positive definite diagonal matrices, and then we have $a_1, a_2 > 0$. By using Lemma 1, we can conclude that V will converge to equilibrium point in finite time T, such that

$$T \le \frac{2}{a_1(1-\alpha)}\ln\left(\frac{a_1 V^{(1-\alpha/2)}(0)}{a_2} + 1\right). \tag{65}$$

So, η_e can also converge to equilibrium point in finite time; according to equation (62), ξ_e will converge to equilibrium point in finite time too. □

Remark 3. Since $\alpha\mathbf{C}_2\mathrm{diag}(|\eta_e|^{\alpha-1})\mathbf{G}(\eta_e)\xi_e$ is included in equation (51), when η_e reaches the equilibrium point before ξ_e, the control output will become infinite. The singularity can be avoided by selecting $\alpha \in ((1/2), 1)$ in this paper. Because when $\eta_e = 0$, the following equation holds:

$$\begin{aligned}\alpha\mathbf{C}_2\mathrm{diag}\left(\left|\boldsymbol{\eta}_e\right|^{\alpha-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e &= \alpha\mathbf{C}_2\mathrm{diag}\left(\left|\boldsymbol{\eta}_e\right|^{\alpha-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\left(-\mathbf{C}_1\boldsymbol{\eta}_e - \mathbf{C}_2\mathrm{sig}^{\alpha}(\boldsymbol{\eta}_e)\right) \\ &= -\alpha\mathbf{C}_2\mathbf{C}_1\mathbf{G}(\boldsymbol{\eta}_e)\left|\boldsymbol{\eta}_e\right|^{\alpha} - \alpha\mathbf{C}_2^2\mathbf{G}(\boldsymbol{\eta}_e)\mathrm{sig}^{2\alpha-1}(\boldsymbol{\eta}_e) = 0.\end{aligned} \tag{66}$$

Remark 4. When we choose the other two sliding surfaces in equations (49) and (50) and use the same proof method as Theorem 1, we can also get that the system state will converge to the equilibrium point in fixed time. By using equation (49) and Lemma 2, η_e, ξ_e will converge to the equilibrium point in fixed time T_f:

$$T_f \le \frac{2}{a_{1f}(\alpha_1 - 1)} + \frac{2}{a_{2f}(1-\alpha_2)}, \tag{67}$$

where $a_{1f} = 2^{(1+\alpha_1/2)}\lambda_{\min}(\mathbf{G}(\eta_e)\mathbf{C}_1)$ and $a_{2f} = 2^{(1+\alpha_2/2)}\lambda_{\min}(\mathbf{G}(\eta_e)\mathbf{C}_2)$.

By using equation (50) and Lemma 3, η_e, ξ_e will converge to the equilibrium point in modified fixed time T_m:

$$T_m \le \frac{2}{a_{1m}(\alpha_1 - 1)} + \frac{2}{a_{1m}(1-\alpha_2)}\ln\left(1 + \frac{a_{1m}}{a_{2m}}\right), \tag{68}$$

where $a_{1m} = 2^{(1+(1/2)+(1/2)\alpha_1+((1/2)\alpha_1-(1/2))\mathrm{sgn}(|\eta_e|-1)/2)}\lambda_{\min}(\mathbf{G}(\eta_e)\mathbf{C}_1)$ and $a_{2m} = 2^{(1+\alpha_2/2)}\lambda_{\min}(\mathbf{G}(\eta_e)\mathbf{C}_2)$.

To guarantee that the sliding surface can reach $\mathbf{S} = 0$ in finite time, Theorem 2 is proposed.

Theorem 2. *For the relative motion spacecraft system with actuator faults in equation (36), the sliding surface* $\mathbf{S}$ *of the system can converge to a small region containing zero in finite time when using sliding surface in equation (48) and fuzzy adaptive control law in equations (51) and (54).*

Proof. Another candidate Lyapunov function is selected as follows:

$$V_1 = \frac{1}{2}\mathbf{S}^T\boldsymbol{\Gamma}_0\mathbf{S} + \frac{1}{2\gamma}\mathrm{tr}\left(\tilde{W}^T\tilde{W}\right). \tag{69}$$

Taking the derivative of V_1 yields

$$\begin{aligned}\dot{V}_1 &= \mathbf{S}^T\boldsymbol{\Gamma}_0\dot{S} + \frac{1}{\gamma}\mathrm{tr}\left(\tilde{W}^T\dot{\tilde{W}}\right) \\ &= \mathbf{S}^T\boldsymbol{\Gamma}_0\left(\dot{\xi}_e + \mathbf{C}_1\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e + \alpha\mathbf{C}_2\mathrm{diag}\left(\left|\boldsymbol{\eta}_e\right|^{\alpha-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e\right) + \frac{1}{\gamma}\mathrm{tr}\left(\tilde{W}^T\dot{\tilde{W}}\right) \\ &= \mathbf{S}^T\boldsymbol{\Gamma}_0\begin{pmatrix}\mathbf{H} + \boldsymbol{\Gamma}_0^{-1}\mathbf{D}\mathbf{E}\mathbf{u}_c + \Delta\tilde{d} + \mathbf{C}_1\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e \\ +\alpha\mathbf{C}_2\mathrm{diag}\left(\left|\boldsymbol{\eta}_e\right|^{\alpha-1}\right)\mathbf{G}(\boldsymbol{\eta}_e)\boldsymbol{\xi}_e\end{pmatrix} + \frac{1}{\gamma}\mathrm{tr}\left(\tilde{W}^T\dot{\tilde{W}}\right) \\ &= -\mathbf{S}^T\boldsymbol{\Gamma}_0\tilde{W}\boldsymbol{\beta}(\mathbf{X}) + \mathbf{S}^T\boldsymbol{\Gamma}_0\boldsymbol{\varepsilon} - \mathbf{S}^T\mathbf{K}_1\mathbf{S} - \mathbf{S}^T\mathbf{K}_2\mathrm{sig}^{\alpha}(\mathbf{S}) + \frac{1}{\gamma}\mathrm{tr}\left(\tilde{W}^T\dot{\tilde{W}}\right) \\ &= \frac{1}{\gamma}\mathrm{tr}\left(\tilde{W}^T\left(\dot{\tilde{W}} - \gamma\boldsymbol{\Gamma}_0^T\mathbf{S}\boldsymbol{\beta}^T(\mathbf{X})\right)\right) + \mathbf{S}^T\boldsymbol{\Gamma}_0\boldsymbol{\varepsilon} - \mathbf{S}^T\mathbf{K}_1\mathbf{S} - \mathbf{S}^T\mathbf{K}_2\mathrm{sig}^{\alpha}(\mathbf{S}).\end{aligned} \tag{70}$$

Substituting the adaptive law of equation (54) into the above equation yields

$$\begin{aligned}\dot{V}_1 &= \mathbf{S}^T\boldsymbol{\Gamma}_0\boldsymbol{\varepsilon} - \mathbf{S}^T\mathbf{K}_1\mathbf{S} - \mathbf{S}^T\mathbf{K}_2\mathrm{sig}^{\alpha}(\mathbf{S})\\ &\le -\lambda_{\min}(\mathbf{K}_1)\frac{2}{\lambda_{\max}(\boldsymbol{\Gamma}_0)}\frac{1}{2}\mathbf{S}^T\boldsymbol{\Gamma}_0\mathbf{S} - \lambda_{\min}(\mathbf{K}_2)\left[\frac{2}{\lambda_{\max}(\boldsymbol{\Gamma}_0)}\right]^{(1+\alpha/2)}\left[\frac{1}{2}\mathbf{S}^T\boldsymbol{\Gamma}_0\mathbf{S}\right]^{(1+\alpha/2)}\\ &\quad -\frac{1}{2\gamma}\mathrm{tr}\left(\widetilde{W}^T\widetilde{W}\right) - \left[\frac{1}{2\gamma}\mathrm{tr}\left(\widetilde{W}^T\widetilde{W}\right)\right]^{(1+\alpha/2)} + \Delta,\end{aligned} \tag{71}$$

where Δ is defined as

$$\Delta = \frac{1}{2\gamma}\mathrm{tr}\left(\widetilde{W}^T\widetilde{W}\right) + \left[\frac{1}{2\gamma}\mathrm{tr}\left(\widetilde{W}^T\widetilde{W}\right)\right]^{(1+\alpha/2)} + \|\mathbf{S}\|\|\boldsymbol{\Gamma}_0\|\|\boldsymbol{\varepsilon}\|. \tag{72}$$

From Assumptions 2 and 3, we know that the following inequalities are satisfied:

$$\Delta \le \frac{1}{2\gamma}W_m + \left(\frac{1}{2\gamma}W_m\right)^{(1+\alpha/2)} + \varepsilon_m\|\mathbf{S}\|\|\boldsymbol{\Gamma}_0\| = \Delta', \tag{73}$$

where χ_1 and χ_2 are defined to satisfy the following equations:

$$\begin{aligned}\chi_1 &= \min\left\{\lambda_{\min}(\mathbf{K}_1)\frac{2}{\lambda_{\max}(\boldsymbol{\Gamma}_0)}, 1\right\},\\ \chi_2 &= \min\left\{\lambda_{\min}(\mathbf{K}_2)\left(\frac{2}{\lambda_{\max}(\boldsymbol{\Gamma}_0)}\right)^{(1+\alpha/2)}, 1\right\}.\end{aligned} \tag{74}$$

Then equation (71) can be simplified as

$$\dot{V}_1 \le -\chi_1 V_1 - \chi_2 V_1^{(1+\alpha/2)} + \Delta'. \tag{75}$$

The above equation can be rewritten as

$$\begin{cases}\dot{V}_1 + \overline{\chi}_1 V_1 + \chi_2 V_1^{(1+\alpha/2)} \le 0,\\ \dot{V}_1 + \chi_1 V_1 + \overline{\chi}_2 V_1^{(1+\alpha/2)} \le 0,\end{cases} \tag{76}$$

where $\overline{\chi}_1 = \chi_1 - (\Delta'/V_1)$ and $\overline{\chi}_2 = \chi_2 - (\Delta'/V_1^{(1+\alpha/2)})$, by using Lemma 1, V_1 will converge to the equilibrium point in finite time.

When $\overline{\chi}_1 > 0$, that is, $V_1 > (\Delta'/\chi_1)$, V_1 will converge to the region Δ_1 containing zero in finite time T_1:

$$\begin{aligned}T_1 &\le \frac{2}{\overline{\chi}_1(1-\alpha)}\ln\left(\frac{\overline{\chi}_1 V^{(1-\alpha/2)}(0)}{\chi_2} + 1\right),\\ \Delta_1 &\le \frac{\Delta'}{\chi_1}.\end{aligned} \tag{77}$$

When $\overline{\chi}_2 > 0$, that is, $V_1 > (\Delta'/\chi_2)^{(2/1+\alpha)}$, V_1 will converge to the region Δ_2 containing zero in finite time T_2:

$$\begin{aligned}T_2 &\le \frac{2}{\chi_1(1-\alpha)}\ln\left(\frac{\chi_1 V^{(2/1-\alpha)}(0)}{\overline{\chi}_2} + 1\right),\\ \Delta_2 &\le \left(\frac{\Delta'}{\chi_2}\right)^{(2/1+\alpha)}.\end{aligned} \tag{78}$$

According to equations (77) and (78), we can conclude that V_1 will converge to the region $\overline{\Delta}$ containing zero in finite time T':

$$\begin{aligned}T' &= \min\{T_1, T_2\},\\ \overline{\Delta} &= \min\{\Delta_1, \Delta_2\}.\end{aligned} \tag{79}$$

Since the following inequality holds

$$\begin{aligned}\frac{1}{2}\mathbf{S}^T\boldsymbol{\Gamma}_0\mathbf{S} &\le V_1,\\ \frac{1}{2\gamma}\mathrm{tr}\left(\widetilde{W}^T\widetilde{W}\right) &\le V_1,\end{aligned} \tag{80}$$

the sliding surface $\mathbf{S}$ of the system can converge to a small region $\Delta_S = \sqrt{(2/\lambda_{\min}(\Xi_0))\overline{\Delta}}$ containing zero in finite time. The estimated value of the optimal weight matrix can also converge to the true value. □

Remark 5. When we choose the other two controllers in equations (52) and (53) and use the same adaptive update law in equation (54), we can also draw the conclusion that the sliding surface converges in fixed time. By using equation (52) and Lemma 2, V_1 will converge to the region $\overline{\Delta}_f$ containing zero in fixed time T_f'.

$$\begin{aligned}T_f' &= \min\left\{T_{1f}, T_{2f}\right\},\\ \overline{\Delta}_f &= \min\left\{\Delta_{1f}, \Delta_{2f}\right\},\end{aligned} \tag{81}$$

where $T_{1f} \le (2/\overline{\chi}_{1f}(\alpha_1 - 1)) + (2/\chi_{2f}(1-\alpha_2))$, $\Delta_{1f} \le (\Delta_f'/\chi_{1f})^{(2/1+\alpha_1)}$; $T_{2f} \le (2/\chi_{1f}(\alpha_1 - 1)) + (2/\overline{\chi}_{2f}(1-\alpha_2))$, $\Delta_{2f} \le (\Delta_f'/\chi_{2f})^{(2/1+\alpha_2)}$. The derivation process of χ_{1f}, $\overline{\chi}_{1f}$, χ_{2f}, and $\overline{\chi}_{2f}$ can be referenced with Theorem 2.

By using equation (53) and Lemma 3, V_1 will converge to the region $\overline{\Delta}_m$ containing zero in fixed time T_m':

$$\begin{aligned}T_m' &= \min\{T_{1m}, T_{2m}\},\\ \overline{\Delta}_m &= \min\{\Delta_{1m}, \Delta_{2m}\},\end{aligned} \tag{82}$$

where $T_{1m} \le (2/\overline{\chi}_{1m}(\alpha_1 - 1)) + (2/\chi_{1m}(1 - \alpha_2))\ln(\chi_{1m}/\chi_{2m})$, $\Delta_{1m} \le (\Delta_m'/\chi_{1m})^{(2/1+\alpha_1)}$; $T_{2m} \le (2/\chi_{1m}(\alpha_1 - 1)) + (2/\overline{\chi}_{1m}(1 - \alpha_2))\ln(\overline{\chi}_{1m}/\overline{\chi}_{2m})$, $\Delta_{2m} \le (\Delta_m'/\chi_{2m})^{(2/1+\alpha_2)}$. The derivation process of χ_{1m}, $\overline{\chi}_{1m}$, χ_{2m}, and $\overline{\chi}_{2m}$ can be referenced with Theorem 2.

4. Numerical Simulation Analysis

In this part, three kinds of fuzzy adaptive finite-time and fixed-time fault-tolerant control algorithms proposed in this paper are simulated to verify the effectiveness of the algorithms. Before the simulation, the input expression of the fuzzy approximation system is defined as

$$x_k = \frac{s_k}{|s_k| + 0.0001}, \quad (k = 1, 2, \ldots, 6). \tag{83}$$

Seven fuzzy membership functions are selected as follows:

$$\begin{cases} \mu_{A_k^1}(x_k) = \dfrac{1}{1 + \exp(5(x_k + \pi/4))}, \\ \mu_{A_k^2}(x_k) = \exp\left(-0.5\left(\dfrac{x_k + 1}{0.25}\right)^2\right), \\ \mu_{A_k^3}(x_k) = \exp\left(-0.5\left(\dfrac{x_k + 0.5}{0.25}\right)^2\right), \\ \mu_{A_k^4}(x_k) = \exp\left(-0.5\left(\dfrac{x_k}{0.25}\right)^2\right), \\ \mu_{A_k^5}(x_k) = \exp\left(-0.5\left(\dfrac{x_k - 0.5}{0.25}\right)^2\right), \\ \mu_{A_k^6}(x_k) = \exp\left(-0.5\left(\dfrac{x_k - 1}{0.25}\right)^2\right), \\ \mu_{A_k^7}(x_k) = \dfrac{1}{1 + \exp(5(x_k - \pi/4))}. \end{cases} \tag{84}$$

In the simulation, the mass and moment of inertia of the follower spacecraft and the leader spacecraft are chosen the same as

$$m = 110\text{kg}, \quad \mathbf{J} = \begin{bmatrix} 21.7 & -0.2 & -0.5 \\ -0.2 & 22.3 & -0.3 \\ -0.5 & -0.3 & 25.5 \end{bmatrix} \text{kg.m}^2. \tag{85}$$

The leader spacecraft moves on a Molniya orbit, and its initial orbital elements are given in Table 1 [3].

Assuming that the leader spacecraft moves along an ideal orbit, its orbit is generated by offline calculation. At the initial moment, the body-fixed frame of the leader spacecraft coincides with the orbital coordinate system, and its initial pose configuration and initial velocity are

$$\mathbf{g}_o = \begin{bmatrix} 0.8660 & -0.5 & 0 & 16490.0 \\ 0.2239 & 0.3878 & -0.8942 & 4262.8 \\ 0.4471 & 0.7744 & 0.4478 & 8512.6 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\xi}_o = \begin{bmatrix} 3.7052 & 3.6292 & 0 & 0 & 0 & 0.0011 \end{bmatrix}^T. \tag{86}$$

The position vector is expressed in the ECI system, and its unit iskm, and the velocity vector is expressed in the body-fixed coordinate system, and the units are rad/s and km/s.

The definition of the initial pose configuration and initial velocity parameters of the follower spacecraft relative to the leader spacecraft are shown in Table 2.

The desired pose configuration and desired velocity of the follower spacecraft relative to the leader spacecraft are shown in Table 3.

In other words, the control goal is to keep the attitude synchronization between the follower spacecraft and the leader spacecraft, hover under it, and the relative velocity of the two is zero.

The uncertain part of the mass and inertia matrix and the external disturbances are selected as follows:

$$\Delta\boldsymbol{\Gamma} = \text{diag}(\sin(0.5t)\mathbf{I}_3, 0.1\sin(0.5t)\mathbf{I}_3), \quad \Gamma_d = \begin{bmatrix} 0.05\sin(0.5t)\text{N} \\ 0.05\sin(0.5t)\text{N} \\ -0.05\sin(0.5t)\text{N} \\ 0.005\sin(0.15t)\text{N} \cdot \text{m} \\ 0.005\sin(0.25t)\text{N} \cdot \text{m} \\ -0.005\sin(0.2t)\text{N} \cdot \text{m} \end{bmatrix}. \tag{87}$$

During the simulation, the maximum output of the reaction flywheel and thruster are 1 N.m and 10 N, respectively. That is, the boundary of control force is $[-10, 10]$N, and the control torque is limited to $[-1, 1]\text{N} \cdot \text{m}$. The parameters of the controllers are chosen as in Table 4.

The specific fault types of each flywheel and each thruster are shown in Table 5.

Figures 4–7 show the output of AF-MFixed and its comparison with AF-Finite and AF-Fixed under the normal condition of the actuator; Figures 8–11 show the output with the actuator fault. The above output results verify the stability analysis of the proposed control scheme.

Figure 4 illustrates the pose configuration and the velocity tracking error of AF-MFixed without actuator fault. It can be seen that the attitude and angular velocity tracking errors quickly converge to the equilibrium state within 18 s, and the convergence accuracy is finally maintained within 1×10^{-4}deg and 2×10^{-5}deg/s, respectively; the position and translational velocity tracking errors quickly converge to the equilibrium state within 60 s, and the convergence

Table 1: Initial orbital elements of the leader.

Orbital element	Value
Semimajor axis (km) a (km)	26628
Eccentricity e	0.7417
Inclination i (deg)	63.4
RAAN Ω (deg)	0
Argument of perigee ω (deg)	270
True anomaly f (deg)	120

Table 2: Initial state of the follower spacecraft relative to the leader spacecraft.

Initial relative parameters	Values
Initial relative position (m)	$[15 \ \ 15 \ \ 15]^T$
Initial relative linear velocity (m/s)	$[-0.051 \ \ -0.247 \ \ -0.075]^T$
Initial relative attitude (rad)	$2\pi/3$
Initial relative principal rotation axis	$[-2 \ \ -2 \ \ -3]^T$
Initial relative angular velocity (rad/s)	$[0.009 \ \ 5.98 \ \ -9.31]^T \times 10^{-4}$

accuracy is finally maintained within 8×10^{-5}m and 1×10^{-6}m/s, respectively.

Figure 5(a) shows the output torque of each flywheel and Figure 5(b) shows the output force of each thruster under normal conditions. It is evident that all the actuator outputs are bounded.

Figure 6 shows the comparison of the pose configuration and the velocity tracking error norms among AF-Finite, AF-Fixed, and AF-MFixed without actuator fault. It can be seen from Figures 6(a) and 6(b) that the convergence rates and convergence accuracy of the three methods are almost the same for attitude and angular velocity tracking errors. From Figures 6(c) and 6(d), as for position and translational velocity tracking errors, AF-MFixed and AF-Fixed are superior to AF-Finite in convergence rates, AF-Finite have the highest convergence accuracy of position tracking error, and the minimum overshoot of translational velocity tracking error, but it has the lowest convergence accuracy of translational velocity tracking error, and the control performance of AF-Fixed is between the other two. In summary, AF-MFixed has more obvious advantages in terms of rapidity than AF-Finite and accuracy of control performance than AF-Fixed, which also confirms the analysis and discussion of AF-MFixed in Remark 2.

Figure 7 shows the comparison of the integration of control force and torque of the three methods without actuator fault. The integral of the control output often represents the control energy. From Figure 7(a), we can know that AF-Finite has the largest energy consumption of control force (integration of control force) and AF-Fixed and AF-MFixed have the similar control force energy consumption. However, it can be seen from Figure 7(b) that there is no significant difference in the energy consumption of control torque (integration of control torque) among the three. Therefore, AF-MFixed also has great advantages in reducing control energy consumption.

Figure 8 illustrates the pose configuration and the velocity tracking error of AF-MFixed with actuator fault. It can

Table 3: Desired state of the follower spacecraft relative to the leader spacecraft.

Desired relative parameters	Values
Desired relative position (m)	$[5 \ \ 0 \ \ 0]^T$
Desired relative linear velocity (m/s)	$[0 \ \ 0 \ \ 0]^T$
Desired relative attitude (rad)	0
Desired relative angular velocity (rad/s)	$[0 \ \ 0 \ \ 0]^T$

Table 4: Control parameters for simulation.

Parameter name	Values
Sliding surface	$\alpha = 0.6, \alpha_1 = 1.2, \alpha_2 = 0.6$ $\mathbf{C}_1 = \mathrm{diag}(0.054, 0.054, 0.054, 0.18, 0.18, 0.18)$ $\mathbf{C}_2 = \mathrm{diag}(0.054, 0.054, 0.054, 0.18, 0.18, 0.18)$
Controller parameters	$\mathbf{K}_1 = \mathrm{diag}(1200, 800, 800, 10, 10, 10)$ $\mathbf{K}_2 = \mathrm{diag}(1200, 800, 800, 10, 10, 10)$ $\gamma = 0.001$

Table 5: Fault conditions of actuators.

Actuator	Fault expression
Flywheel 1	$u_1 = \begin{cases} u_{1c} & t < 25\,\mathrm{s} \\ 0.65u_{1c} & t \geq 25\,\mathrm{s} \end{cases}$
Flywheel 2	$u_2 = \begin{cases} u_{2c} & t < 25\,\mathrm{s} \\ 0.5u_{2c} & t \geq 25\,\mathrm{s} \end{cases}$
Flywheel 3	$u_3 = 0.8u_{3c}, t > 0\,\mathrm{s}$
Flywheel 4	$u_4 = 0.6u_{4c}, t > 0\,\mathrm{s}$
Thruster 1	$u_5 = \begin{cases} 0\,\mathrm{N} & t < 15\,\mathrm{s} \\ 0.4u_{5c} & t \geq 15\,\mathrm{s} \end{cases}$
Thruster 2	$u_6 = 0.9u_{6c}, t > 0\,\mathrm{s}$
Thruster 3	$u_7 = \begin{cases} 0\mathrm{N} & t < 15\,\mathrm{s} \\ 0.6u_{7c} & t \geq 15\,\mathrm{s} \end{cases}$
Thruster 4	$u_8 = 0.75u_{8c}, t > 0\,\mathrm{s}$
Thruster 5	$u_9 = \begin{cases} u_{9c} & t < 1\,\mathrm{s} \\ 0.8u_{9c} & t \geq 15\,\mathrm{s} \end{cases}$
Thruster 6	$u_{10} = 0.3u_{10c}, t > 0\,\mathrm{s}$
Thruster 7	$u_{11} = \begin{cases} u_{11c} & t < 15\,\mathrm{s} \\ 0.6u_{11c} & t \geq 15\,\mathrm{s} \end{cases}$
Thruster 8	$u_{12} = 0.45u_{12c}, t > 0\,\mathrm{s}$

be seen that the attitude and angular velocity tracking errors quickly converge to the equilibrium state within 20 s, and the convergence accuracy is finally maintained within 3×10^{-4}deg and 2×10^{-5}deg/s, respectively. The position tracking errors quickly converge to the equilibrium state within 60 s, and the convergence accuracy is finally maintained within 4×10^{-3}m and tends to decrease; the translational velocity tracking errors quickly converge to the equilibrium state within 75 s, and the convergence accuracy is finally maintained within 2×10^{-5}m/s.

Figure 9(a) shows the output torque of each flywheel, and Figure 9(b) shows the output force of each thruster under fault conditions. It is evident that all the actuator outputs are bounded, and the output curves well reflect the types of the fault. It is worth noting that the control torques and forces do not vanish completely when the control goal is

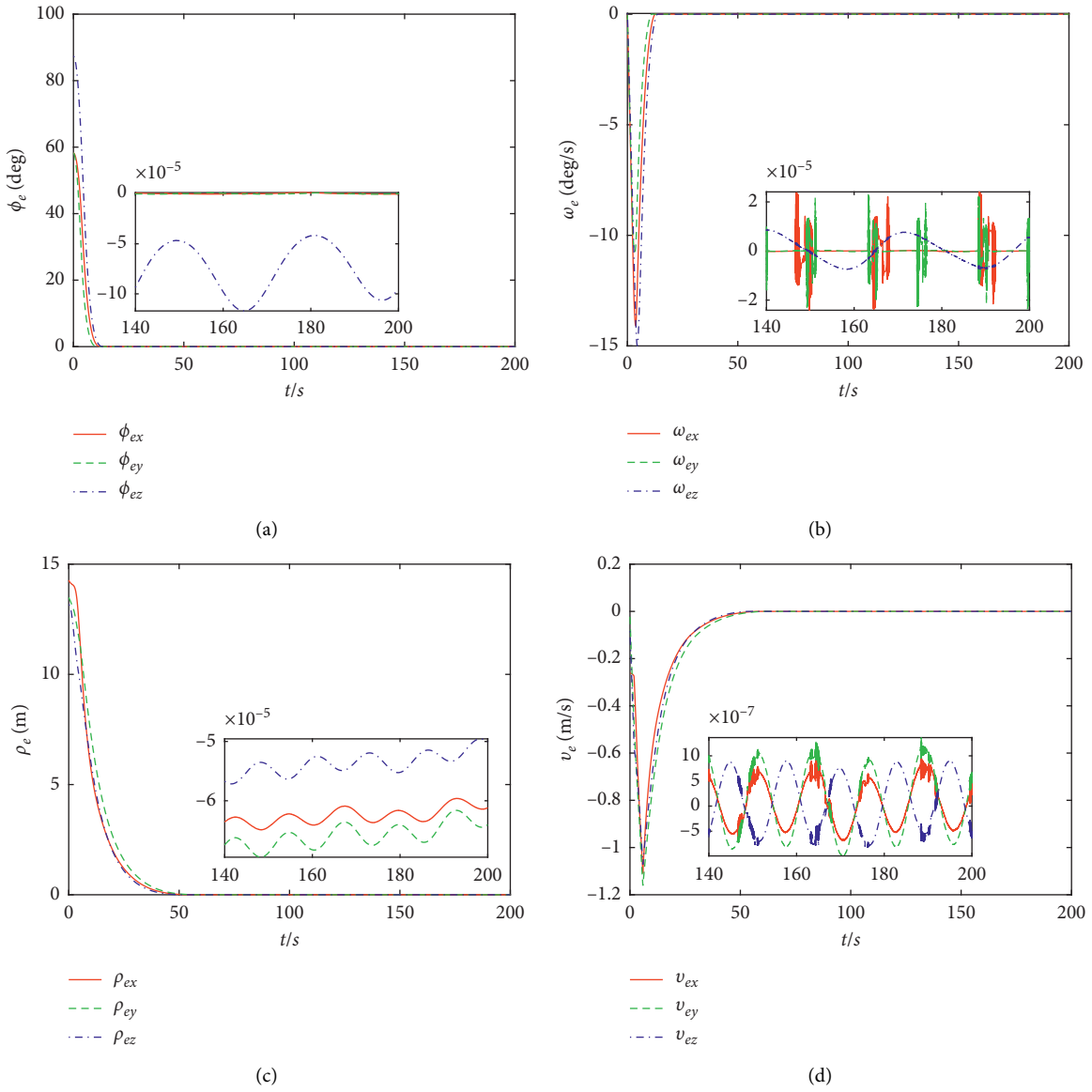
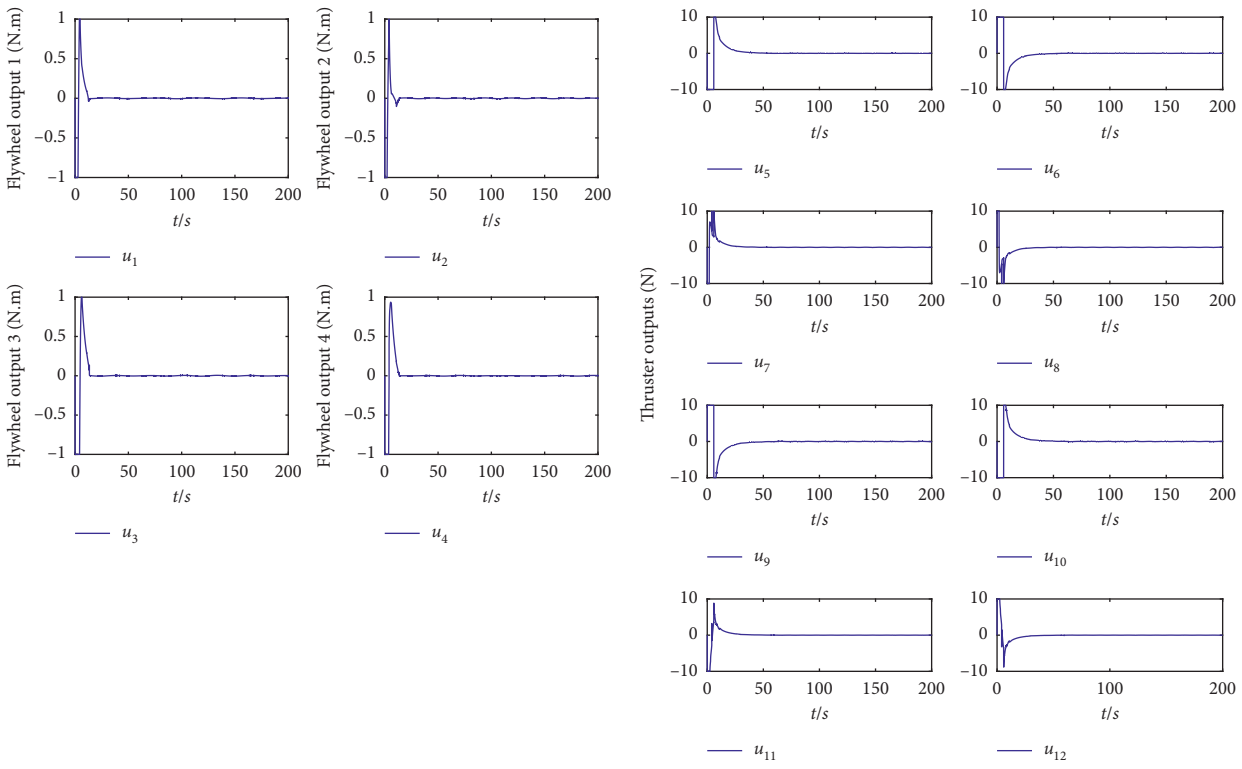


Figure 4: Tracking errors of AF-MFixed under normal condition.

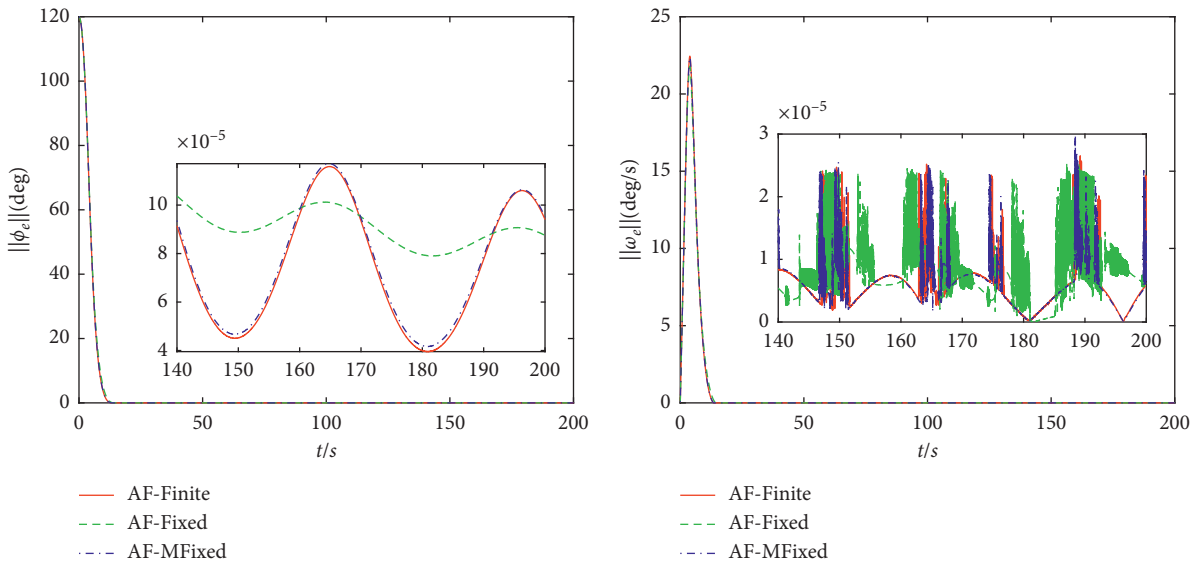
achieved because they also need to compensate for the total disturbances and actuator faults to keep the relative pose configuration between the follower spacecraft and the leader spacecraft.

Figure 10 shows the comparison of the pose configuration and the velocity tracking errors norms among AF-Finite, AF-Fixed, and AF-MFixed with actuator fault. It can be seen from Figures 10(a) and 10(b) that the convergence rates of the three methods are almost the same for attitude and angular velocity tracking errors, but AF-MFixed has a significant advantage over AF-Fixed in terms of convergence accuracy for attitude tracking error. We can know from Figures 10(a) and 10(c) that AF-Finite has the highest convergence accuracy for relative pose configuration tracking error. From Figures 10(c) and 10(d), as for position and translational velocity tracking errors, AF-MFixed and AF-Fixed are superior to AF-Finite in convergence rates, AF-Finite have the highest convergence accuracy of position and translational velocity tracking error and the minimum overshoot of translational velocity tracking error, the control performance of AF-Fixed is between the other two, and it can realize the translational velocity tracking error control with high accuracy and convergence rates, but AF-MFixed can converge more



(a) (b)

Figure 5: Tracking errors of AF-MFixed under normal condition.



(a) (b)

Figure 6: Continued.

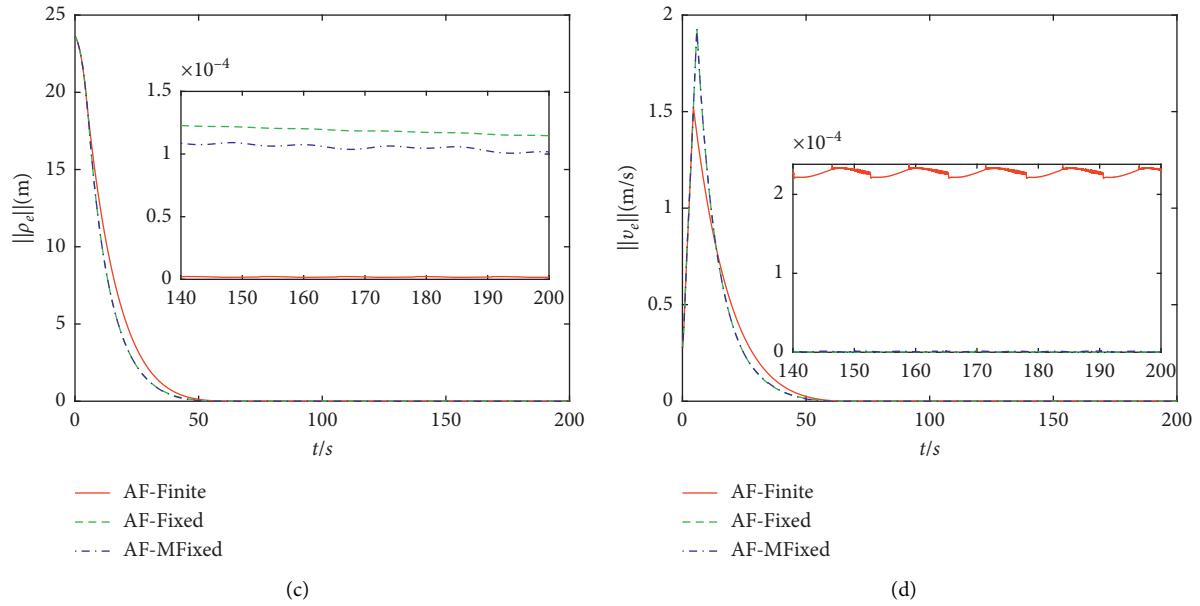


Figure 6: Comparison of the tracking error output norms of the three methods under normal condition.

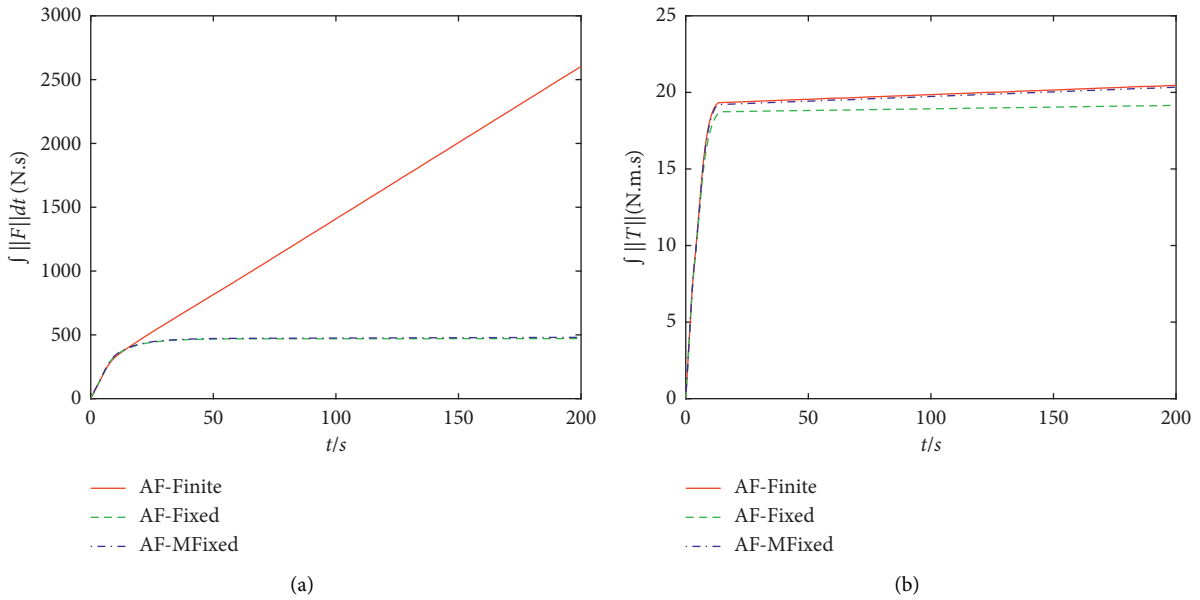


Figure 7: Comparison of the integration of control output of the three methods under normal condition.

accurately than AF-Fixed at the same convergence speed. In summary, AF-MFixed has more obvious advantages in terms of rapidity and accuracy of control performance, which also confirms the analysis and discussion of AF-MFixed in Remark 2.

Figure 11 shows the comparison of the integration of control force and torque of the three methods with actuator fault. From Figure 11(a), we can know that AF-Finite has the lowest energy consumption of control force, and AF-MFixed is slightly higher than that of AF-Fixed in order to achieve fast convergence performance. However, it can be seen from Figure 11(b) that AF-Fixed has the lowest energy consumption of control torque, and AF-Finite is slightly higher than that of AF-Fixed, but the difference in the

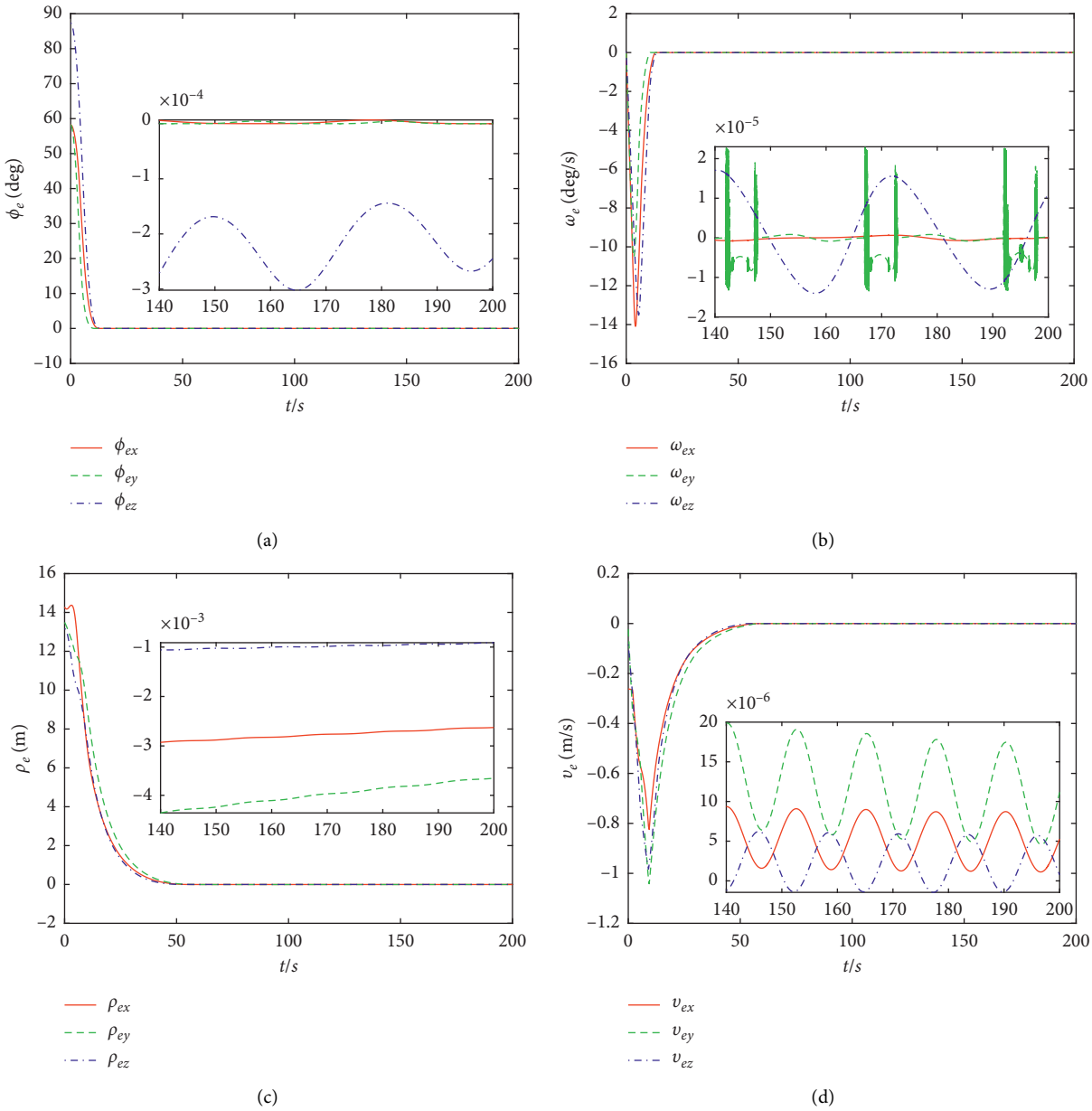
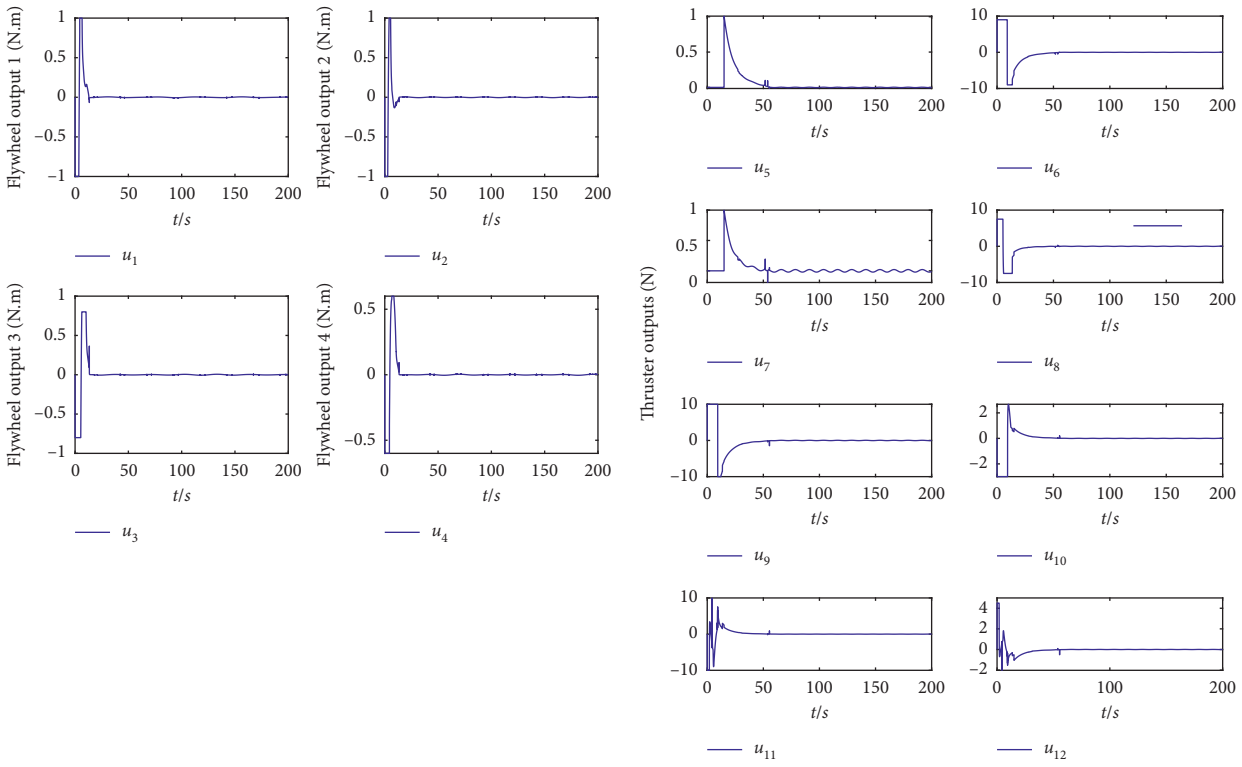
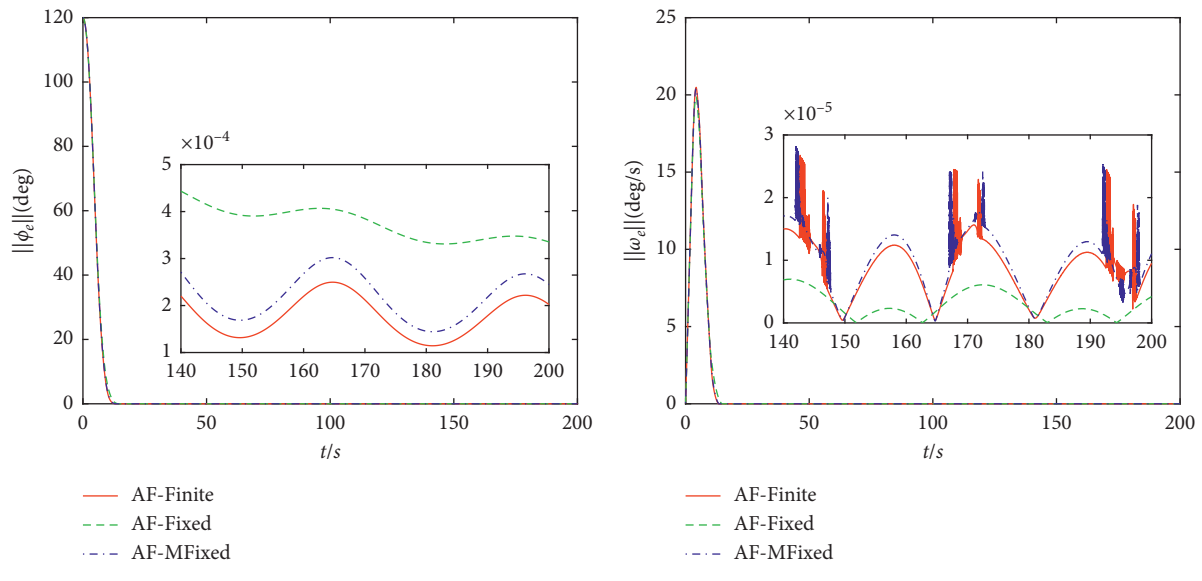


Figure 8: Tracking errors of AF-MFixed with actuator fault.



(a) (b)

Figure 9: Output of actuators under fault condition.



(a) (b)

Figure 10: Continued.

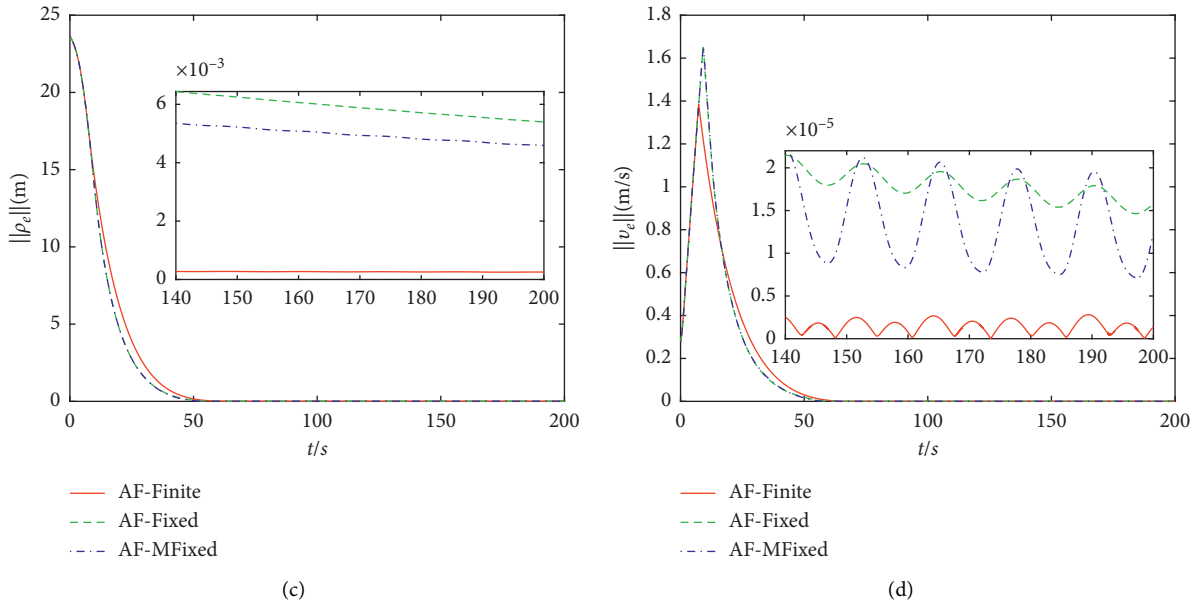


Figure 10: Comparison of the tracking error output norms of the three methods with actuator fault.

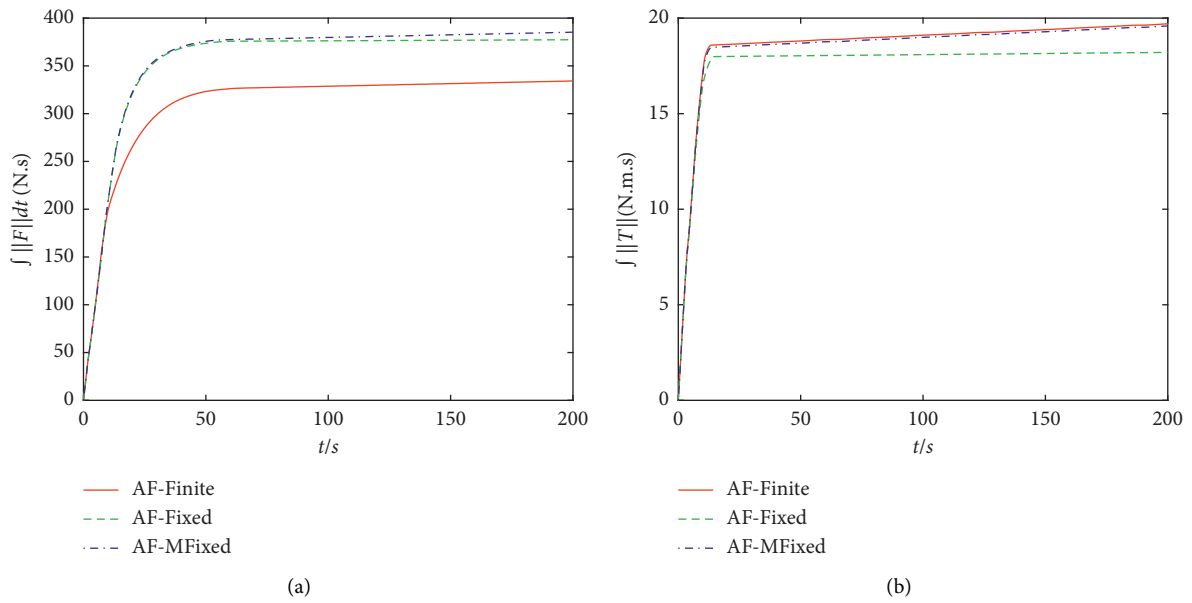


Figure 11: Comparison of the integration of control output of the three methods with actuator fault.

energy consumption of control torque is not obvious among the three compared with energy consumption of control force.

5. Conclusions

In this paper, adaptive fuzzy modified Fixed-time fault-tolerant control schemes on SE(3) for coupled spacecraft were proposed to solve the attitude and position tracking problem with external disturbances, model uncertainties, and actuator faults simultaneously. From the comparative analysis of the three control strategies, we can see that AF-MFixed can achieve the control goals of fast convergence and higher tracking accuracy; the settling time of the closed-loop tacking system can be independent of the initial states. The integrated attitude and position modeling method based on Lie group SE(3) is simple and can be applied to solve the problem of 6-DOF in practical aerospace engineering. The

fuzzy adaptive control scheme can estimate the total disturbances and fault information with high accuracy. The parameter tuning of the proposed algorithm is simple, avoiding the tedious adjustment of too many parameters. Moreover, the algorithm is suitable for both actuator failure and normal condition, and the control performance under fault condition will be slightly lower than that under normal condition, which also shows that the robustness proposed in this paper has a strong advantage, and it has potential engineering application value. However, the practical problem is that the actuator fault information is difficult to obtain in real time. The establishment of a fault observer for fault diagnosis will be the work of the author in the future, and the estimated fault information will be applied to the design of fault-tolerant controllers. In addition, the next research content of this paper will consider the influence of sensor measurement noise and fault, and design a more robust fault-tolerant controller.

Data Availability

The data used to support the findings of this study are included in this paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Shanghai Aerospace Science and Technology Innovation Foundation, Grant No. SAST2020-023.

References

[1] M. S. De Queiroz, V. Kapila, and Q. Yan, "Adaptive nonlinear control of multiple spacecraft formation flying," *Journal of Guidance, Control, and Dynamics*, vol. 23, no. 3, pp. 385–390, 2000.

[2] D. Lee, "Spacecraft coupled tracking maneuver using sliding mode control with input saturation," *Journal of Aerospace Engineering*, vol. 28, no. 5, Article ID 04014136, 2014.

[3] K. Gong, Y. Y. Liao, and Y. Wang, "Adaptive fixed-time terminal sliding mode control on SE(3) for coupled spacecraft tracking maneuver," *International Journal of Aerospace Engineering*, vol. 2020, pp. 1–15, 2020.

[4] Z. Chen, Q. Chen, X. He et al., "Adaptive finite-time command filtered fault-tolerant control for uncertain spacecraft with prescribed performance," *Complexity*, vol. 2018, 2018.

[5] F. Song and S. Qin, "robust fault-tolerant control for satellite attitude stabilization based on active disturbance rejection approach with artificial bee colony algorithm," *Mathematical Problems in Engineering*, vol. 2014, no. 4, 17 pages, Article ID 512707, 2014.

[6] J. Zhang, D. Ye, Z. Sun et al., "Extended state observer based robust adaptive control on SE(3) for coupled space-craft tracking maneuver with actuator saturation and misa-lignment," *Acta Astronautica*, vol. 143, 2018.

[7] H. Dong, Q. Hu, M. I. Friswell, and G. Ma, "Dual-quaternion-based fault-tolerant control for spacecraft tracking with finite-time convergence," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 4, pp. 1231–1242, 2017.

[8] H. Dong, Q. Hu, and G. Ma, "Dual-quaternion based fault-tolerant control for spacecraft formation flying with finite-time convergence," *ISA Transactions*, vol. 61, no. Mar., pp. 87–94, 2016.

[9] D. Lee, A. K. Sanyal, and E. A. Butcher, "Asymptotic tracking control for spacecraft formation flying with decentralized collision avoidance," *Journal of Guidance, Control, and Dynamics*, vol. 38, no. 4, pp. 587–600, 2015.

[10] D. Lee and G. Vukovich, "Almost global finite-time stabilization of spacecraft formation flying with decentralized collision avoidance," *International Journal of Control, Automation and Systems*, vol. 15, no. 3, pp. 1167–1180, 2017.

[11] M. R. Binette, C. J. Damaren, and L. Pavel, "Nonlinear H∞ attitude control using modified Rodrigues parameters," *Journal of Guidance, Control, and Dynamics*, vol. 37, no. 6, pp. 2017–2021, 2014.

[12] Y. Huang and Y. Jia, "Robust adaptive fixed-time tracking control of 6-DOF spacecraft fly-around mission for noncooperative target," *International Journal of Robust and Nonlinear Control*, vol. 28, no. 6, pp. 2598–2618, 2018.

[13] J. Zhang, D. Ye, J. D. Biggs, and Z. Sun, "Finite-time relative orbit-attitude tracking control for multi-spacecraft with collision avoidance and changing network topologies," *Advances in Space Research*, vol. 63, no. 3, pp. 1161–1175, 2019.

[14] D. Lee, A. Sanyal, E. Butcher, and D. Scheeres, "Finite-time control for spacecraft body-fixed hovering over an asteroid," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 1, pp. 506–520, 2015.

[15] A. Sanyal, L. Holguin, and S. P. Viswanathan, "Guidance and control for spacecraft autonomous chasing and close proximity maneuvers," *IFAC Proceedings Volumes*, vol. 45, no. 13, pp. 753–758, 2012.

[16] F. Bullo and R. M. Murray, *Proportional Derivative (Pd) Control on the Euclidean Group*, vol. 2, pp. 1091–1097, European Control Conference Citeseer, Zurich, Switzerland, 1995.

[17] L. Jiang, Y. Wang, and S. Xu, "Integrated 6-DOF orbit-attitude dynamical modeling and control using geometric mechanics," *International Journal of Aero-Space Engineering*, vol. 2017, no. 1, 13 pages, Article ID 4034328, 2017.

[18] J. Ackermann, "Parameter space design of robust control systems," *IEEE Transactions on Automatic Control*, vol. 25, no. 6, pp. 1058–1072, 1980.

[19] D. Zhao, H. Yang, B. Jiang, and L. Wen, "Attitude stabilization of a flexible spacecraft under actuator complete failure," *Acta Astronautica*, vol. 123, pp. 129–136, 2016.

[20] H. Qian, M. Y. Peng, and M. Cui, "Adaptive observer-based fault-tolerant control design for uncertain systems," *Mathematical Problems in Engineering*, vol. 2015, no. 3, 16 pages, Article ID 429361, 2015.

[21] G. Q. Wu, S. N. Wu, and Z. G. Wu, "Robust finite-time control for spacecraft with coupled translation and attitude dynamics," *Mathematical Problems in Engineering*, vol. 2013, no. 4, 7 pages, Article ID 707485, 2013.

[22] L. Zheng, Q. X. Dong, and X. M. YangZeng, "Robust adaptive sliding mode fault tolerant control for nonlinear system with actuator fault and external disturbance," *Mathematical Problems in Engineering*, vol. 2019, Article ID 6349510, 13 pages, 2019.

[23] K. Zhou and Y. Xia, "Adaptive attitude tracking control for rigid spacecraft with finite-time convergence," *Automatica*, vol. 49, no. 12, pp. 3591–3599, 2013.

[24] T. Wang, J. Qiu, S. Yin, H. Gao, J. Fan, and T. Chai, "Performance-based adaptive fuzzy tracking control for networked industrial processes," *IEEE Transactions on Cybernetics*, vol. 46, no. 8, pp. 1760–1770, 2016.

[25] T. Wang, H. Gao, and J. Qiu, "A combined adaptive neural network and nonlinear model predictive control for multirate networked industrial process control," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 2, pp. 416–425, 2015.

[26] L. Li, S. X. Ding, J. Qiu, Y. Yang, and Y. Zhang, "Weighted fuzzy observer-based fault detection approach for discrete-time nonlinear systems via piecewise-fuzzy Lyapunov functions," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 6, pp. 1320–1333, 2016.

[27] L. Li, S. X. Ding, J. Qiu, and Y. Yang, "Real-time fault detection approach for nonlinear systems and its asynchronous T–S fuzzy observer-based implementation," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 283–294, 2017.

[28] Y. Wei, J. Qiu, and H.-K. Lam, "A novel approach to reliable output feedback control of fuzzy-affine systems with time delays and sensor faults," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, p. 1808, 2017.

[29] Y. Wei, J. Qiu, H. K. Lam, and L. Wu, "Approaches to T–S fuzzy-affine-model-based reliable output feedback control for nonlinear ITO stochastic systems," *IEEE Transactions on Fuzzy Systems*, vol. 25, 2017.

[30] A.-M. Zou and K. D. Kumar, "Adaptive fuzzy fault-tolerant attitude control of spacecraft," *Control Engineering Practice*, vol. 19, no. 1, pp. 10–21, 2011.

[31] B. Huo, Y. Xia, K. Lu, and M. Fu, "Adaptive fuzzy finite-time fault-tolerant attitude control of rigid spacecraft," *Journal of the Franklin Institute*, vol. 352, no. 10, pp. 4225–4246, 2015.

[32] J. Zhang, M. D. Ye, and Z. Sun, "Adaptive fuzzy finite-time control for spacecraft formation with communication delays and changing topologies," *Journal of the Franklin Institute*, vol. 354, no. 11, pp. 4377–4403, 2017.

[33] S. T. Venkataraman and S. Gulati, "Control of nonlinear systems using terminal sliding modes," *Journal of Dynamic Systems, Measurement, and Control*, vol. 115, no. 3, pp. 554–560, 1993.

[34] X. Yu and M. Zhihong, "Fast terminal sliding-mode control design for nonlinear dynamical systems," *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, vol. 49, no. 2, pp. 261–264, 2002.

[35] L. Yang and J. Yang, "Nonsingular fast terminal sliding-mode control for nonlinear dynamical systems," *International Journal of Robust and Nonlinear Control*, vol. 21, no. 16, pp. 1865–1879, 2011.

[36] X. N. Shi, Z. G. Zhou, and D. Zhou, "Adaptive fault-tolerant attitude tracking control of rigid spacecraft on Lie group with fixed time convergence," *Asian Journal of Control*, vol. 22, no. 1, pp. 423–435, 2020.

[37] X.-N. Shi, Y.-A. Zhang, D. Zhou, and Z.-G. Zhou, "Global fixed-time attitude tracking control for the rigid spacecraft with actuator saturation and faults," *Acta Astronautica*, vol. 155, pp. 325–333, 2019.

[38] J. W. Gao, Z. Fu, and S. Zhang, "Adaptive fixed-time attitude tracking control for rigid spacecraft with actuator faults," *IEEE Transactions on Industrial Electronics*, vol. 66, 2019.

[39] S. Mobayen, J. Ma, G. Pujol-Vazquez, L. Acho, and Q. Zhu, "Adaptive finite-time stabilization of chaotic flow with a single unstable node using a nonlinear function-based global sliding mode," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 43, no. S1, pp. 339–347, 2019.

[40] S. Mobayen and G. Pujol-Vázquez, "A robust LMI approach on nonlinear feedback stabilization of continuous state-delay systems with lipschitzian nonlinearities: experimental validation," *Iranian Journal of Science and Technology, Transactions of Mechanical Engineering*, vol. 43, no. 3, pp. 549–558, 2019.

[41] M. Jafari and S. Mobayen, "Second-order sliding set design for a class of uncertain nonlinear systems with disturbances: an LMI approach," *Mathematics and Computers in Simulation*, vol. 156, pp. 110–125, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6615839, 11 pages
https://doi.org/10.1155/2021/6615839

Research Article

Multiparty Homomorphic Machine Learning with Data Security and Model Preservation

Fengtian Kuang,[1] **Bo Mi**,[2] **Yang Li**,[2] **Yuan Weng**,[2] **and Shijie Wu**[3]

[1]*Chongqing Jiaotong University, Mathematics and Statistics, Chongqing 400074, China*
[2]*Chongqing Jiaotong University, Information Science and Engineering, Chongqing 400074, China*
[3]*Unit 78156 of the Chinese People's Liberation Army, Chengdu 610000, China*

Correspondence should be addressed to Bo Mi; mi_bo@163.com

Received 1 November 2020; Revised 29 November 2020; Accepted 22 December 2020; Published 11 January 2021

Academic Editor: Yong Chen

With the widespread application of machine learning (ML), data security has been a serious issue. To eliminate the conflict between data privacy and computability, homomorphism is extensively researched due to its capacity of performing operations over ciphertexts. Considering that the data provided by a single party are not always adequate to derive a competent model via machine learning, we proposed a privacy-preserving training method for the neural network over multiple data providers. Moreover, taking the trainer's intellectual property into account, our scheme also achieved the goal of model parameter protection. Thanks to the hardness of the conjugate search problem (CSP) and discrete logarithm problem (DLP), the confidentiality of training data and system model can be reduced to well-studied security assumptions. In terms of efficiency, since all messages are coded as low-dimensional matrices, the expansion rates with regard to storage and computation overheads are linear compared to plaintext implementation without accuracy loss. In reality, our method can be transplanted to any machine learning system involving multiple parties due to its capacity of fully homomorphic computation.

1. Introduction

With the continuous development of artificial intelligence, data have become precious resources due to their value for mining. Nevertheless, numerous private information is embodied as data, which may be abused to violate personal privacy, business secrets, or even state secrets. For example, once a patient's medical record is exposed to insurance companies, they may never sell him some kind of medical insurance [1]. Similarly, many other machine learning applications have also caught sight of privacy infringements, such as financial analysis, product customization, and public opinion surveillance [2–4]. On the other hand, any data-driving mechanism heavily relies on the quantity and quality of information, which brings about the conflict between data usability and data confidentiality. Fortunately, secure multiparty computation (SMC) [5–7] and homomorphic encryption (HE) [8, 9] provide us powerful tools to process data in a concealed manner. Therefore, the remaining problem to address is how to devise a cryptosystem that is applicable for machine learning in consideration of storage and computation overheads.

As a cryptographic technology orienting decentralized systems, secure multiparty computation aims at data confidentiality for distributed participants. Despite the privacy concern, the involved parties can still figure out a public output as they wish. Based on such cryptosystem, F. Ö. Çatak et al. [10] proposed a privacy-preserving learning protocol for classification in virtue of vertically segmented data from multiple parties. Since the data is just partially shared without concealing, semantic security is unachievable as plain data dose. The first provable secure ML protocol of this kind is presented by R. Devin et al. [11] for text classification. Howbeit, their research only focused on the privacy of data classification and left the learning process unaddressed.

Oriented at centralized systems, homomorphic encryption is another way towards secure machine learning, which is capable of performing specific operations over ciphertexts. Researches of applying FE for data privacy during machine learning have developed rapidly since the significant innovation [12] appeared in 2016. Y. Aono et al. [13] combined the additive homomorphism with deep learning to narrow the gap between system functionality and data security, by applying FH technology to asynchronous stochastic gradient descent algorithm. F. Bourse et al. [14] improved the FHE structure of Chillotti et al. [15] and proposed a homomorphic neural network evaluation framework, namely, FHE-DiNN. Its complexity is strictly linear in network depth, but the model parameters must be proactively predefined. Based on a multikey variant of two HE schemes [16, 17] with ciphertexts packed, H. Chen et al. [18] provided a suite of interfaces for secure machine learning which also exploited bootstrapping for arbitrary circuit evaluation. As matter of fact, almost all existing FHE-based machine learning algorithms are based on the algebraic structure of lattice, such as BGV [19–21], CKKS [22–24], and NTRU [25–27]. These methods suffer from a common defect that decryption may fail due to noise growth. Though bootstrapping can be deemed as an effective tool for noise control, its extra computational burden is hardly acceptable. Surprisingly, J. Li et al. [28] discovered an alternative tool, saying Conjugate Search Problem, to actualize full homomorphism without noise interference. They also applied such cryptosystem for privacy-preserving data training, which achieved the same accuracy as the plaintexts used for learning.

Though more comprehensible and effective than lattice-based secure machine learning, Li's scheme can only be applied to the scenario of a signal data provider. Ordinarily, one party can always provide a small quantity of data which may incur an overfitted model. To ensure the generalization of machine learning, data from diverse sources should be gathered for a specific learning task. In the circumstances of multiparty secure machine learning, each data provider may conceal their information by a dependent key. Therefore, a training framework that operates over heterogeneous (i.e., encrypted by different keys) ciphertext is desiderated. Conversely, the parameters of the system model should be taken as assets held by the trainer as in general business operation. Thus, we should also make sure that the machine is concealed, even when not thoroughly trained.

To preserve the privacy of all participants, this paper presents a complete machine learning mechanism in virtue of CSP and DLP hardness. Our contributions are summarized as follows.

1.1. Contributions

(1) We coded float-type data as low-dimensional upper triangular matrices that are homomorphic under the operations of addition, subtraction, multiplication, division, and comparison. With the help of CSP, the plain matrices can also be projected to semantically secure ciphertexts homomorphically under the same kind of operations. That is to say, our basic cryptosystem is fully homomorphic, since addition and multiplication are simultaneously implemented. Therefore, we can realize secure training and classification/regression once private data are provided under the same key.

(2) We constructed a cyclic group by lifting the plain matrices to a Galois domain. Thereafter, key switching (switch a ciphertext encrypted by one key to another) is made possible via DLP for the purpose of cooperative training.

(3) We combined the two aforementioned technologies and devised a secure machine learning protocol under semihonest model, which preserves the privacy of multiple data providers as well as that of the trainer.

2. System Model

Neural network (NN) is employed as the engineering background and verification model in this paper due to its extensive application. Nevertheless, it is worth mentioning that our scheme can be applied to most machine learning algorithms if privacy is significant to multiple participants.

2.1. Neural Network Model. A typical neural network contains three or more layers, which turns into a deep learning model if hidden layers are multiple [29]. The certain principle of NN lies in the fact that numerous neurons can automatically extract features of the inputs layer by layer. Besides the topology of NN, the most important factors that defined it are the weights and bias designated to each link and neuron. As for learning, the essence is how to adjust these parameters in virtue of training data via iterative forward-/back-propagation. Thereafter, to securely implement a neural network model, we should homomorphically evaluate the following functions.

Forward calculation (e.g., sigmoid):

$$f_{fw} = \text{sigmoid}\left(\mathbf{a}_i, \mathbf{w}_i + b_i\right), \tag{1}$$

where $\mathbf{a}_i$ and $\mathbf{w}_i$ are the input and weight vectors corresponding to the proactive links of neuron i, while b_i represents its bias.

Backward calculation:

Loss function (e.g., quadratic loss function):

$$f_{\text{bw\_loss}} = L(Y|ft(X)) = \sum_n \left(t_n - o_n\right)^2, \tag{2}$$

where t_n is the target value and o_n is the actual value.

Parameter adjusting (e.g., gradient descent):

$$f_{\text{bw\_adj}} = \text{old}\, w_{j,k} - \Delta w_{j,k}, \tag{3}$$

$$\Delta w_{j,k} = \alpha \cdot E_k \cdot O_k\left(1 - O_k\right) \cdot O_j^T, \tag{4}$$

where E_k is the error vector between the target value and the actual value, O_j^T is the transpose of the output of the current layer node, and O_k is the output of the node of the next layer.

2.2. System Model and Security Goal. In our system, a powerful trainer expects to acquire a neural network whose topology is predefined. To ensure the completeness of the resultant model, they may request multiple parties for training data. However, the data providers concern about privacy leakage though they have strong wills to cooperate. Meanwhile, the trainer also worries that the system parameters may expose and infringe their intellectual property. Therefore, we should preserve the privacy of all participants and guarantee the functionality of machine learning at the same time. Moreover, taking the trained neural network as a service, a user may not only desire to designate a classification/regression task to the server but also be anxious about data abuse.

2.3. Adversary Model. Suppose that the trainer and all data providers are honest but curious during the whole process. That is to say, they will completely follow the protocol to avoid unnecessary disputes but may be interested in the privacy contained within the data. Furthermore, it is reasonable to assume that both the trainer and data owner are provided with PPT (probabilistic polynomial time) computational power. However, since the trainer is always better equipped than data providers, the hypothesis that they have the accessibility to a quantum machine may also be valid. To define the success of privacy violation, we exploit the concept of symmetric IND-CPA (indistinguishability under chosen-plaintext attack) as below.

2.4. Symmetric IND-CPA [30]. Define an experiment under symmetric cryptosystem SE = (SE.KeyGen, SE.Enc, SE.Dec) as

$$\begin{aligned}
&\mathrm{Exp}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\kappa):\\
&k\leftarrow \mathrm{HE.KeyGen}(\kappa),\\
&(M_0, M_1)\leftarrow \mathscr{A}^{\mathrm{HE.Enc}_k(\cdot)}(\cdot),\ \text{for } |M_0| = |M_1|,\\
&\gamma\leftarrow_R\{0, 1\},\ C^* = \mathrm{HE.Dec}(M_\gamma),\\
&\gamma\prime = \mathscr{A}(C^*),\\
&\text{Output 1, if } \gamma = \gamma\prime, \text{and 0 otherwise,}
\end{aligned} \tag{5}$$

for any PPT adversary $\mathscr{A}$ that queries the oracle $\mathrm{HE.Enc}_k(\cdot)$ polynomial times. Thus, the adversary's advantage can be expressed by

$$\mathrm{Adv}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\kappa) = \left|\Pr\left[\mathrm{Exp}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\lambda) = 1\right] - \frac{1}{2}\right|. \tag{6}$$

Then the cryptosystem SE is IND-CPA-secure if $\mathrm{Adv}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\kappa) < \epsilon(\kappa)$, where $\epsilon(\kappa)$ stands for a negligible function in the security parameter κ.

3. Cryptographic Construction

Focusing on the security goals presented in the system model, we are now ready to construct our cryptographic building blocks. In this part, we first explore the homomorphism of conjugate search problem to underpin the functionality of training over homogeneous (i.e., encrypted by the same key) ciphertexts. Then, we present a key switching technology that can convert a ciphertext encrypted by one key to be decryptable by another.

Conjugate search problem is a special form of group factorization problem (GFP) [31], defined as follows.

3.1. Conjugate Search Problem (CSP) [31]. Given $(C, M) \in \Psi \times \Psi$ over a nonabelian algebraic structure Ψ, it is intractable to solve $H \in \Psi$ such that $C = \mathrm{HMH}^{-1}$.

B. Evgeni [32] proved that the CSP is postquantum secure over the general linear group $\mathrm{GL}_d(R)$ (R means real number field) if $d \geq 4$. Hence, to assure system security, we should code the message as a matrix with degree larger than 4.

To protect the privacy of data providers without affecting the accuracy of training, we resort to homomorphic encryption that is capable of actualizing the forward-/backward-propagation processes covertly. Thereafter, we devised a way that makes CSP semantically secure and homomorphic. It is worth noting that the conjugate search problem is resistant to quantum attacks, which dispels the privacy concern for data providers even if the trainer is extremely equipped.

A typical homomorphic encryption algorithm can be noted as a tetrad HE = (HE.KeyGen, HE.Enc, HE.Dec, HE.Eval), standing for the functions of key generation, encryption, decryption, and evaluation, respectively.

For any data m over the message space R, we first code it as an upper triangular matrix $M \in R^{6\times 6}$ as follows.

3.2. Encoding. Convert the message m into three pairs of random numbers (a_1, a_2), (a_3, a_4), and (a_5, a_6), satisfying $a_1 + a_2 = m$, $a_3 + a_4 = a_5 + a_6 = r$, and $(a_3^2 - a_4^2)(a_5^2 - a_6^2) = 1$, where r is a constant random number of the system. Thus, we can construct the following matrices:

$$\begin{aligned}
M_1 &= \begin{pmatrix} a_1 & a_2 \\ a_2 & a_1 \end{pmatrix},\\
M_2 &= \begin{pmatrix} a_3 & a_4 \\ a_4 & a_3 \end{pmatrix},\\
M_3 &= \begin{pmatrix} a_5 & a_6 \\ a_6 & a_5 \end{pmatrix}.
\end{aligned} \tag{7}$$

Combining the above matrices, the message m is finally coded as

$$M = \begin{pmatrix} M_1 & R_1 & R_2 \\ 0 & M_2 & R_3 \\ 0 & 0 & M_3 \end{pmatrix}, \tag{8}$$

where 0 represents the 2×2 all-zero matrix and R_i $(i = 1, 2, 3)$ stands for random matrices uniformly sampled from $R^{2\times2}$.

For clarity, we denote the space of coded messages as Γ. It is interesting that Γ naturally constitutes a multiplicative cyclic group (excluding the elements whose determinants are zero) and $R \sim \Gamma$ (homomorphic). Furthermore, it is well known that all square matrices with the same dimension compose a ring. Though $\Gamma \subseteq R^{6\times6}$ and its elements are commutative for multiplication, there is an overwhelming probability that a matrix P uniformly sampled from $R^{6\times6}$ is noncommutative with the coded message M. Thereupon, a CSP-based fully homomorphic encryption algorithm can be actualized as below.

3.3. Key Generation. HE.KeyGen(1^κ): uniformly sample a matrix from $R^{9\times9}$, which can also be represented as a combination of nine 6×6 random matrices, namely,

$$P = \begin{pmatrix} P_1 & P_2 & P_3 \\ P_4 & P_5 & P_6 \\ P_7 & P_8 & P_9 \end{pmatrix}, \quad \text{for } P_i = \begin{pmatrix} p_{i1} & p_{i2} \\ p_{i3} & p_{i4} \end{pmatrix}, (i = 1, 2, \ldots, 9). \tag{9}$$

The probability that P is communitive with elements in Γ should be negligible. Then, the algorithm takes $k = P$ as the symmetric key.

3.4. Encryption. HE.Enc(P, M): output

$$C = \mathrm{PMP}^{-1} = P\begin{pmatrix} M_1 & R_1 & R_2 \\ 0 & M_2 & R_3 \\ 0 & 0 & M_3 \end{pmatrix}P^{-1} \tag{10}$$

as the ciphertext of message m (coded as a matrix M).

3.5. Decryption. HE.Dec(P, C): compute $M = P^{-1}CP$ to obtain

$$M_1 = \begin{pmatrix} a_1 & a_2 \\ a_2 & a_1 \end{pmatrix}. \tag{11}$$

Then, figure out $m = a_1 + a_2$ to recover the plaintext.

3.6. Evaluation. HE.Eval$(f, C_1, \ldots, C_l)$: We describe the very basic operations underpinning formulae (2)-(4) in advance. Suppose that C_1 and C_2 are ciphertexts corresponding to m_1 and m_2 under the same key; the additive and multiplicative arithmetic can be simply carried out by $C_{\text{add}} = C_1 + C_2$ and $C_{\text{mul}} = C_1C_2$. These two operations can be trivially assembled to realize the functions for backward propagation. However, since the exponential operation cannot be implemented directly via homomorphic addition and multiplication, some activation functions of forward propagation such as *sigmoid* should be approximated as the form of polynomials. Thereby, we resort to a specific conversion [32–34],

$$\mathrm{sigmod}(x) = \begin{cases} 0.000734x^4 + 0.014222x^3 + 0.108706x^2 + & \\ 0.392773x + 0.571859, & -\infty < x \le -1.5, \\ 0.002083x^5 + 0.020833x^3 + 0.25x + 0.5, & -1.5 < x \le 1.5, \\ -0.000734x^4 + 0.014222x^3 - 0.108706x^2 + & \\ 0.3922773x + 0.428141, & 1.5 < x < \infty, \end{cases} \tag{12}$$

to replace

$$\mathrm{sigmod}(x) = \frac{1}{1 + e^{-x}}. \tag{13}$$

Noting that the aforementioned formula is expressed as a piecewise function, to homomorphically decide which subfunction should be carried out, we can encrypt the numbers of -1.5 and 1.5 and compare them with x for branching.

To program a piecewise function, J. Li et al. [28] presented a homomorphic algorithm that covertly compares the size between two ciphertexts. Though our scheme is similar to that of [28], we argue that their cryptosystem is not semantically secure because $a_{2i-1} + a_{2i} = m$ and a_{2i-1} is always bigger than a_{2i} for $i = 1, 2, 3$.

3.7. Security Analysis of [28]. By computing

$$\det\left(C^* - TC\prime\right) = \det(P)\det\left(\begin{pmatrix} M_1^* - tM_1' & R_1^* & R_2^* \\ 0 & M_2^* - tM_2' & R_3^* \\ 0 & 0 & M_2^* + tM_2' \end{pmatrix}\right)\det\left(P^{-1}\right), \tag{14}$$

where R_i^* is also a random matrix, for

$$T = \begin{pmatrix} t & R_1 & R_2 \\ 0 & t & R_3 \\ 0 & 0 & t \end{pmatrix}, \tag{15}$$

where

$$t = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{16}$$

and $R_i\,(i = 1, 2, 3)$ is uniformly sampled from $R^{2\times 2}$, the adversary carries out a chosen-plaintext attack such as the following.

$\mathrm{Exp}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\kappa)$: ,
$k \leftarrow \mathrm{HE.KeyGen}(\kappa)$,
$(M_0, M_1, C') \leftarrow \mathscr{A}^{\mathrm{HE.Enc}_k(\cdot)}(\cdot)$,,
for $|M_0| = |M_1|, C' = \mathrm{HE.Enc}_k(M')$, and $M_0 < M' < M_1$,
$\gamma \leftarrow_R \{0, 1\}$, $C^* = \mathrm{HE.Dec}(M_\gamma)$,
$\gamma' = 1$ if $\det(C^* - \mathrm{TC}')$, and $\gamma' = 0$ otherwise,
Output 1, if $\gamma = \gamma'$, and 0 otherwise.
Considering that

$$\begin{aligned} \det(C^* - \mathrm{TC}') = \det(P)\det(M_1^* - tM_1')\det(M_2^* - tM_2') \\ \det(M_3^* - tM_3')\det(P^{-1}), \quad (i = 1, 2, 3), \end{aligned} \tag{17}$$

where

$$\begin{aligned} \det(M_i^* - tM_i') &= (a_{2i-1}^* - a_{2i}')^2 - (a_{2i}^* - a_{2i-1}')^2, \\ &= (m^* - m\prime)((a_{2i-1}^* - a_{2i}^*)t + n(a_{2i-1}' - a_{2i}')), \end{aligned} \tag{18}$$

since $a_{2i-1} > a_{2i}$ is guaranteed throughout Li's scheme [28], $((a_{2i-1}^* - a_{2i}^*)t + n(qa_{2i-1}'h - a\,2i)$ must be positive. It is obvious that $\det(P)\det(P^{-1}) = 1$; hence, the adversary can easily determine whether $m^* = m_0$ or $m^* = m_1$ by checking the sign of $\det(C^* - \mathrm{TC}\prime)$. That is to say,

$$\mathrm{Adv}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\kappa) = \left|\Pr\left[\mathrm{Exp}_{\mathrm{HE},\mathscr{A}}^{\mathrm{CPA}}(\kappa) = 1\right] - \frac{1}{2}\right| = 1. \tag{19}$$

It seems that the conflict between piecewise function evaluation and IND-CPA security is infeasible to address. However, we can introduce a specific form of ciphertext which can be used to encrypt a designated number and compare it with any other normal ciphertext. Our construction is given below.

The data provider randomly chooses a nonzero number $k \in R - \{0\}$ and encrypts $m\prime$ as

$$C\prime = \mathrm{PM}\prime P^{-1} = P\begin{pmatrix} kM_1' & R_1 & R_2 \\ 0 & kM_2' & R_3 \\ 0 & 0 & kM_3' \end{pmatrix}P^{-1}, \tag{20}$$

for

$$M_i' = \begin{pmatrix} a_{i1}' & a_{i2}' \\ a_{i3}' & a_{i4}' \end{pmatrix}, \quad (i = 1, 2, 3), \tag{21}$$

which satisfies

$$\begin{cases} a_{i1}' + a_{i4}' = k \neq 0, & i = 1, 2, 3, \\ a_{i2}' + a_{i3}' = -k \neq 0, & i = 1, 2, 3, \\ a_{i1}'a_{i4}' - a_{i2}'a_{i3}' = m\prime, & i = 1, \\ a_{i1}'a_{i4}' - a_{i2}'a_{i3}' = r, & i = 2, 3. \end{cases} \tag{22}$$

To compare C' with general cipher C^* without decryption, the evaluator computes

$$\Delta = \frac{\det(C^*C')}{\det(C')} - \det(C^* - C') = k^2(m^* - m') \tag{23}$$

and thus achieves

$$C_{\mathrm{comp}} = \begin{cases} m^* > m' & \text{if } \Delta > 0, \\ m^* > m' & \text{if } \Delta = 0, \\ m^* > m' & \text{if } \Delta < 0. \end{cases} \tag{24}$$

3.8. Correctness. The correctness of encryption and decryption algorithms is straightforward, so we only focus on the homomorphism of evaluation.

Homomorphic addition: since

$$\begin{aligned} C_{\mathrm{add}} &= C_1 + C_2, \\ &= P^{-1}\begin{pmatrix} M_{11} & R_{11} & R_{12} \\ 0 & M_{12} & R_{13} \\ 0 & 0 & M_{13} \end{pmatrix}P + P^{-1}\begin{pmatrix} M_{21} & R_{21} & R_{22} \\ 0 & M_{22} & R_{23} \\ 0 & 0 & M_{23} \end{pmatrix}P, \\ &= P^{-1}\begin{pmatrix} M_{11} + M_{21} & R_{11} + R_{21} & R_{12} + R_{22} \\ 0 & M_{12} + M_{22} & R_{13} + R_{23} \\ 0 & 0 & M_{13} + M_{23} \end{pmatrix}P, \end{aligned} \tag{25}$$

we can decrypt it as

$$\begin{aligned} M_{\mathrm{add}} &= P(C_1 + C_2)P^{-1}, \\ &= \begin{pmatrix} M_{11} + M_{21} & R_{11} + R_{21} & R_{12} + R_{22} \\ 0 & M_{12} + M_{22} & R_{13} + R_{23} \\ 0 & 0 & M_{13} + M_{23} \end{pmatrix} \end{aligned} \tag{26}$$

because

$$M_{11} + M_{21} = \begin{pmatrix} a_{11} + a_{21} & a_{12} + a_{22} \\ a_{12} + a_{22} & a_{11} + a_{21} \end{pmatrix}. \tag{27}$$

The addition of m_1 and m_2 can be decoded as

$$m_{\mathrm{add}} = a_{11} + a_{21} + a_{12} + a_{22} = m_1 + m_2. \tag{28}$$

Homomorphic multiplication: because

$$
\begin{aligned}
C_{\text{mul}} &= C_1 C_2, \\
&= P^{-1}\begin{pmatrix} M_{11} & R_{11} & R_{12} \\ 0 & M_{12} & R_{13} \\ 0 & 0 & M_{13} \end{pmatrix}\begin{pmatrix} M_{21} & R_{21} & R_{22} \\ 0 & M_{22} & R_{23} \\ 0 & 0 & M_{23} \end{pmatrix} P, \\
&= P^{-1}\begin{pmatrix} M_{11}M_{21} & R_1^* & R_2^* \\ 0 & M_{12}M_{22} & R_3^* \\ 0 & 0 & M_{13}M_{23} \end{pmatrix} P, \\
M_{11}M_{21} &= \begin{pmatrix} a_{11}a_{21} + a_{12}a_{22} & a_{11}a_{22} + a_{12}a_{21} \\ a_{12}a_{21} + a_{11}a_{22} & a_{12}a_{22} + a_{11}a_{21} \end{pmatrix},
\end{aligned} \tag{29}
$$

we can deduce that

$$
(a_{11}a_{21} + a_{12}a_{22}) + (a_{11}a_{22} + a_{12}a_{21}) = (a_{11} + a_{12})(a_{21} + a_{22}) = m_1 m_2. \tag{30}
$$

Homomorphic comparison: on the premise of $\det(P)\det(P^{-1}) = 1$, it can be seen that

$$
\begin{aligned}
\Delta &= \frac{\det\left(C^* C\prime\right)}{\det(C\prime)} - \det\left(C^* - C\prime\right), \\
&= \frac{\det\left(M^* M\prime\right)}{\det(M\prime)} - \det\left(M^* - M\prime\right)a, \\
&= \prod_{i=1}^{3} \frac{\det\left(M_i^* k M_i'\right)}{\det\left(k M_i'\right)} - \prod_{i=1}^{3} \det\left(M_i^* - k M_i'\right).
\end{aligned} \tag{31}
$$

According to formula (21), we have

$$
\begin{aligned}
\frac{\det\left(M_i^* k M_i'\right)}{\det\left(k M_i'\right)} &= a_{2i-1}^* 2 - a_{2i}^* 2, \\
\det\left(M_i^* - M_i'\right) &= \left(a_{2i-1}^* 2 - a_{2i}^* 2\right) - k^2 2i \\
&\quad \left(a_{2i-1}^* + a_{2i}^* - \det\left(M_i'\right)\right), \quad i = 1, 2, 3.
\end{aligned} \tag{32}
$$

Recall that $a_{2i-1}^* + a_{2i}^* = m^*$ and $a_{i1}' a_{i4}' - a_{i2}' a_{i3}' = m\prime$ when $i = 1$, while $a_{2i-1}^* + a_{2i}^* = a_{i1}' a_{i4}' - a_{i2}' a_{i3}' = r$ when $i = 2, 3$. In terms of the condition that $(a_3^2 - a_4^2)(a_5^2 - a_6^2) = 1$, we can reduce formula (30) to

$$
\Delta = k^2\left(m^* - m\prime\right). \tag{33}
$$

It is obvious that the signs of Δ and $m^* - m\prime$ are exactly the same, since $k^2 > 0$, which determines the relationship between m^* and m' without decryption.

3.9. Security. Thanks to the hardness of Conjugate Search Problem, an adversary must find P such that $P^{-1}\text{CP} = M$ to recover the plaintext. As for the semantic security of our scheme, it can be seen that $((a_{2i-1}^* - a_{2i}^*) + (a_{2i-1}' - a_{2i}'))$ in formula (17) is not always positive due to arbitrary relationship between a_{2i-1} and a_{2i}. Therefore, when an adversary executes a chosen-plaintext attack as mentioned before, their advantage is negligible. Noting that any normal ciphertext can just be compared with specifically encrypted messages without decryption, the data provider has full control over their privacy and permits exact comparisons only if necessary.

After each training, the neural network coefficients are concealed by the key of the data provider. When multiple data providers take part in the training process, those semimanufactured parameters should also be re-encrypted under the key of subsequent data holder for homomorphic computation. Therefore, we devised a way to decrypt and re-encrypt the machine coefficients without exposing them to data providers, in consideration of the trainer's property right. Our key switching scheme is based on the hardness of Discrete Logarithm Problem (DLP).

3.10. Discrete Logarithm Problem. Given a cyclic group G, a generator $g \in G$, and a random element $h \in G$, it is difficult to find the discrete logarithm a such that $g^a = h$.

Accordingly, if an adversary has obtained a ciphertext $y = h^b = g^{ab} \in G$, it is hard for them to recover h because of the confidentiality on ab [35]. However, in light of the Lagrange theorem [36], we can exploit a trapdoor to reverse y back to h.

3.11. Lagrange Theorem. Denote H as a subgroup of finite G; then, $|H|||G|$, for $|H|$ and $|G|$ are the orders of groups H and G.

Since any $h \in G$ generates a subgroup $H \subseteq G$ via $H = \{h^a | a \in Z\}$, we can conclude that $h^{|G|} = e$ in terms of the Lagrange theorem, where e is the identity of group G.

Based on the aforementioned mathematical tools, we are now ready to construct our key switching scheme as a triad KS = (KS.KeyGen, KS.CSPtoDLP, KS.DLPtoCSP). Without loss of generality, we denote $k_t = (b, s)$, $k_A = P_A$, and $k_B = P_B$ as secret keys belonging to the trainer T and two data providers A and B, respectively. Then KS.KeyGen can be used to generate the encryption/decryption key pair for the trainer, while KS.CSPtoDLP is used to convert a ciphertext C_A encrypted by k_A to be decryptable by k_t and KS.CSPtoDLP is utilized to modify C_t (encrypted under k_t) as C_B whose corresponding key is k_B.

3.12. Key Generation. KS.KeyGen(1^κ): as mentioned before, we denote the space of coded messages as Γ. Suppose that the precision of matrix elements in HE is l-bits whose integer part is m-bits and the decimal part is n-bits. We can multiply any coded plaintext M by 2^n to lift it over $Z_{2^l}^{6\times 6}$. Accordingly, the message space is changed to a cyclic group $\Gamma\prime$ for $|\Gamma\prime| = 2^{12l}(2^l - 1)^3$. Moreover, for each $2^n M_i$, it composes a group Γ_i' satisfying $|\Gamma_i'| = 2^l - 1$. Thereby, we uniformly sample an odd number $b \in Z_{2^l-1}$ and compute $s \in Z_{2^l-1}$ such that $s \cdot b = 1 \bmod (2^l - 1)$. Output $k_t = (b, s)$ as the key to the trainer.

3.13. Switching $\mathbf{C_A}$ *to* $\mathbf{C_t}$ *KS.CSPtoDLP*(C_A). The trainer T changes the encrypted model parameters C_A as $C_A' = 2^n C_A \bmod 2^l \in Z_{2^l}^{6\times 6}$ and sends $C_{At} = (C_A')^b \bmod 2^l$ to data provider A. On receiving C_{At}, A computes $C_t = P_A^{-1} C_{At} P_A \bmod 2^l$ as their response.

3.14. Switching $\mathbf{C_t}$ *to* $\mathbf{C_B}$ *KS.DLPtoCSP*(C_t). On receiving C_t from the trainer T, the data provider B computes their response as $C_{tB} = P_B C_t P_B^{-1} \bmod 2^l$. Therefore, the trainer T can reverse C_{tB} back to a ciphertext $C_B = P_B M P_B^{-1}$ purely encrypted under k_B via $C_B' = (C_{tB})^s \bmod 2^l$ and then right-shift its elements by n-bits.

3.15. Correctness. Since $C_A' = P_A(2^n M)P_A^{-1} \bmod 2^l$, we have

$$C_{At} = \underbrace{P_A(2^n M)P_A^{-1} P_A(2^n M)P_A^{-1}, \ldots, P_A(2^n M)P_A^{-1}}_{b \text{ times}} = P_A(2^n M)^b P_A^{-1} \bmod 2^l. \tag{34}$$

Thus, $C_t = P_A^{-1} P_A (2^n M)^b P_A^{-1} P_A = (2^n M)^b \bmod 2^l$.
Similarly, because $C_{tB} = P_B (2^n M)^b P_B^{-1} \bmod 2^l$,

$$C_B' = P_B (2^n M)^{b\cdot s} P_B^{-1} \bmod 2^l, = P_B \begin{pmatrix} (2^n M_1)^{1+k_1(2^l-1)} & R_1^* & R_2^* \\ 0 & (2^n M_2)^{1+k_3(2^l-1)} & R_3^* \\ 0 & 0 & (2^n M_3)^{1+k_3(2^l-1)} \end{pmatrix} P_B^{-1}, \tag{35}$$

where k_i are integers for $i = 1, 2, 3$.

During the encoding process in HE, it is easy to choose a_{2i-1} such that $2^n a_{2i-1} \neq 0 \bmod 2^l$. Considering that $a_1 + a_2 = m$ and $a_{2i-1} + a_{2i} = r$ for $i = 1, 2$, the space of $2^n M_i$ must be a cyclic group Γ_i' for $|\Gamma_i'| = 2^l - 1$. According to the Lagrange Theorem, it can be seen that $(2^n M_i)^{1+k_i(2^l-1)} = 2^n M_i \bmod 2^l$; thus $C_B' = P_B(2^n M)P_B^{-1} \bmod 2^l$. By right-shifting n-bits on C_B', we obtain $C_B = P_B M P_B^{-1}$.

3.16. Security. Note that, after receiving C_{At}, the trainer can trivially compute $M = (C_t^s \bmod 2^l)/2^n$ to recover the message. Nevertheless, since the model parameters are of their intellectual property, such operation does not conflict with our security goal.

As for data providers, they can just witness an exponential form of the plaintext (i.e., $C_t = (2^n M)^b \bmod 2^l$). According to the hardness of DLP, the information about message M will not be exposed.

4. Privacy-Preserving Machine Learning with Multiple Data Providers

To preserve privacy for machine learning, many cryptographic training and classification/regression methods have been proposed in the scene of a single data provider. In most cases, data should be sourced from multiple providers to guarantee the generality of training. Therefore, we present a secure machine learning mechanism with the capacity of training as well as classification/regression in consideration of data and parameter privacy.

As for training, the cloud is supposed to obtain model parameters with the help of labeled data. During the initialization phase, the trainer T computes $k_t \leftarrow \text{KS.KeyGen}(1^\kappa)$ for key switching and each data provider i generates $k_i \leftarrow \text{HE.KeyGen}(1^\kappa)$ for homomorphic training.

Denote the encoded training data owned by provider i as M_i and the system parameters as M_t. The server primarily encrypts the initialized system coefficients (may contain some private intellectual property information) as $C_t = (2^n M_t)^b \bmod 2^l$ to the first data provider who executes $C_1 \leftarrow \text{HE.Enc}(k_1, M_1)$ and $\widetilde{C}_1 \leftarrow \text{KS.DLPtoCSP}(C_t)$ as their response. On encrypted data C_1 and $\widetilde{C}_1$ corresponding to the same key k_1, the cloud can thus achieve $\overline{C}_1 \leftarrow \text{HE.Eval}(f_{\text{training}}, \widetilde{C}_1)$ which are updated system parameters decryptable by k_1.

For clarity, we describe the above processes as shown in Table 1.

Note that KS.DLPtoCSP$(\cdot)$ is a protocol that should be carried out by both the data provider and the cloud.

To make the updated coefficients homomorphically computable with data encrypted by the following providers, we can exploit the key switching scheme to re-encrypt it. Without loss of generality, the updated parameters under key k_i will be represented as $\overline{C}_i$. By means of $C_t \leftarrow \text{KS.CSPtoDLP}(\overline{C}_i)$ and $\widetilde{C}_{i+1} \leftarrow \text{KS.DLPtoCSP}(C_t)$, the cloud can obtain the re-encrypted coefficients $\widetilde{C}_{i+1}$ with the help of successive providers. After receiving C_{i+1} from the next provider, they can compute $\overline{C}_{i+1} \leftarrow \text{HE.Eval}(f_{\text{training}}, C_{i+1}, \widetilde{C}_{i+1})$ since both ciphertexts are encrypted by k_{i+1}. In consideration of the final parameters $\overline{C}_N$, the cloud needs to execute $C_t \leftarrow \text{KS.CSPtoDLP}(\overline{C}_N)$ with the last provider and then computes $M = (C_t^s \bmod 2^l)/2^n$ to restore the plain parameters.

The subsequent training and recovering processes are presented in Table 2.

The classification/regression process is straightforward that, on encrypted data $C_u \leftarrow \text{HE.Enc}(k_u, M_u)$ and system parameters $\widetilde{C}_u \leftarrow \text{KS.DLPtoCSP}(C_t)$ for $C_t = (2^n M_t)^b \bmod 2^l$, the cloud can homomorphically compute $\overline{C}_u \leftarrow \text{HE.Eval}(f_{\text{cla/reg}}, C_u, \widetilde{C}_u)$. By decrypting the received $\overline{C}_u$, the user obtains the classification/regression result such that $M_{\text{cla/reg}} \leftarrow \text{HE.Dec}(k_u, \overline{C}_u)$. This process can be found in Table 3.

5. Experiment Analysis

We drew support from the power load data of Chongqing Tongnan Electric Power Co., Ltd., dating from May 4 to May 10 in 2015, to verify the effectiveness of our training method. A short-term electrical load prediction model is also testified in virtue of 96 historical data pieces sampled during 4 consecutive days. The original machine learning model is exactly the same as that of [29], which has considered nothing about privacy. Our experiment environment is shown in Table 4.

Table 1: Initialization and first training.

Key generation		
Data providers		Cloud
$k_i \leftarrow \text{HE.KeyGen}(1^\kappa)$		$k_t \leftarrow \text{KS.KeyGen}(1^\kappa)$
First training		
Data provider 1		Cloud
Receives C_t	$\Leftarrow$	$C_t = (2^n M_t)^b \bmod 2^l$
$\widetilde{C}_1 \leftarrow \text{KS.DLPtoCSP}(C_t)$	$\Rightarrow$	Receives $\widetilde{C}_1$
$C_1 \leftarrow \text{HE.Enc}(k_1, M_1)$	$\Rightarrow$	Receives C_1
		$\overline{C}_1 \leftarrow \text{HE.Eval}(f_{training}, C_1, \widetilde{C}_1)$

Table 2: Subsequent training and recovering.

Subsequent training		
Data providers i		Cloud
Receives $\overline{C}_i$	$\Leftarrow$	$\overline{C}_i$
$C_t \leftarrow \text{KS.CSPtoDLP}(\overline{C}_i)$	$\Rightarrow$	Receives C_t
Data providers $i+1$		Cloud
Receives C_t	$\Leftarrow$	C_t
$\widetilde{C}_{i+1} \leftarrow \text{KS.DLPtoCSP}(C_t)$,	$\Rightarrow$	Receives $\widetilde{C}_{i+1}$
$C_{i+1} \leftarrow \text{HE.Enc}(k_{i+1}, M_{i+1})$	$\Rightarrow$	Receives C_{i+1}
		$\overline{C}_{i+1} \leftarrow \text{HE.Eval}(f_{\text{training}}, C_{i+1}, \widetilde{C}_{i+1})$
Recovering		
Data provider N		Cloud
Receives $\overline{C}_N$	$\Leftarrow$	$\overline{C}_N$
$C_t \leftarrow \text{KS.CSPtoDLP}(\overline{C}_N)$	$\Rightarrow$	Receives C_t
		$M_t \leftarrow (C_t^s \bmod 2^l)/2^n$

Table 3: Classification and regression.

Classification/regression		
User u	Cloud	
Receives C_t	$\Leftarrow$	$C_t = (2^n M_t)^b \bmod 2^l$
$\widetilde{C}_u \leftarrow \text{KS.DLPtoCSP}(C_t)$	$\Rightarrow$	Receives $\widetilde{C}_u$
$C_u \leftarrow \text{HE.Enc}(k_u, M_u)$	$\Rightarrow$	Receives C_u
Receives $\overline{C}_u$	$\Leftarrow$	$\overline{C}_u \leftarrow \text{HE.Eval}(f_{\text{cla/reg}}, C_u, \widetilde{C}_u)$
$M_{\text{cla/reg}} \leftarrow \text{HE.Dec}(k_u, \overline{C}_u)$		

Table 4: Experiment environment.

CPU	OS	RAM (GB)	Programming language
i5-10210 U 1.60 GHz	Win10 64-bit	16	Python

To simulate the scenario of multiparty machine learning, we divide the data into three parts and realize the training process corresponding to 3 different keys in HE. To prove that our method is not harmful to the accuracy of the trained network, as is shown in Figure 1, we compared the prediction result directly achieved via original model (without privacy-preserving) with that of ours (privacy-preserving scheme). Figure 1 illustrates that the two results are completely consistent.

The experimental results are shown in Table 5; our scheme can perform encryption training and prediction for multiple data providers in general machine learning. As for the efficiency of training and prediction, our scheme is 73578 and 12000 times slower than its plain version. Nevertheless, since the server is always powerful on computational capacity and the data providers only have to carry out trivial multiplications over $R^{6\times6}$, our scheme is practical in cloud environments. Moreover, if the accuracy is tolerable, we can shorten the ciphertext to make it more efficient.

In terms of communication overheads, encrypted data for training or prediction are 18 times larger than plain messages. In each iteration, the cloud should also exchange the ciphertexts of system parameters with two successive data providers, which are also 18 times of original

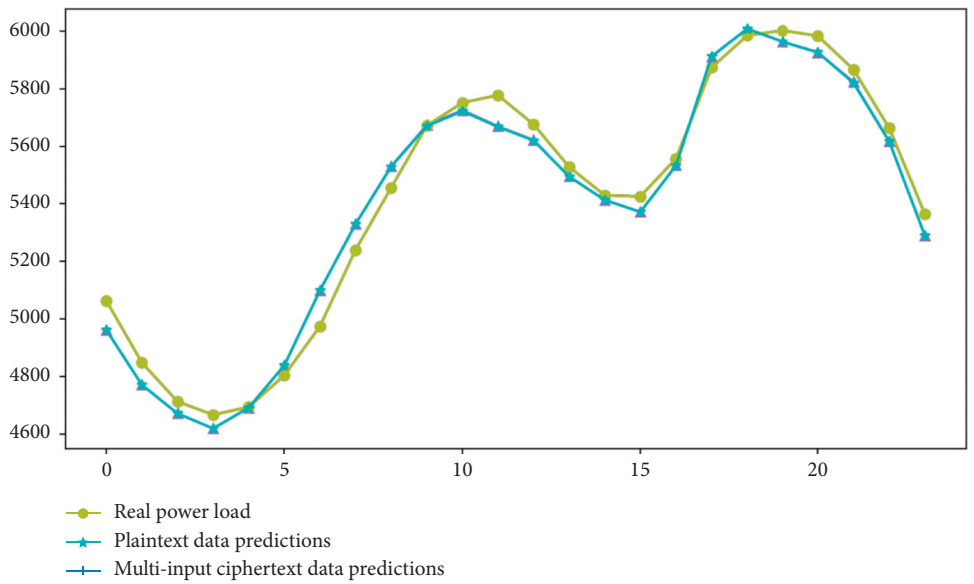


Figure 1: Comparison of prediction results.

Table 5: Computational efficiency.

Model	Training time	Predicted time	Epochs
Original model	0.19 s	4 ms	400
Our model	233.33 min	48 s	400

coefficients. Considering that the expansion rate is not big and system parameters are quite limited, the communication burden causes just little performance degradation.

6. Conclusions

We presented a privacy-preserving machine learning method that works over multiple data providers in this paper. Thanks to the hardness of the conjugate search problem, data can be homomorphically processed for training or classification/regression under the same key. It is worth mentioning that we solved the intrinsic conflict between IND-CPA security and homomorphic comparision (without decryption), by specifically encoding the data which is allowed to be compared. To support training among multiple data providers, a key switching technology is also proposed based on the difficulty of the discrete logarithm problem and Lagrange theorem, which evaded the necessity of multikey homomorphic computation. Experiment illustrated that the accuracy of machine learning cannot be affected by the privacy capability of our scheme. The expansion rate of computation/communication complexity is small enough, which makes the scheme practical in cloud environments.

Data Availability

Our dataset comes from Chongqing Tongnan Electric Power Co., Ltd. (telephone: 023-44559308; official website: http://www.12398.gov.cn/html/information/753078881/753078881201200006.shtml).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Darong Huang and Yang Liu for their comments and suggestions. This work was supported in part by the National Natural Science Foundation of P.R. China under Grants 61573076, 61703063, and 61903053, the Science and Technology Research Project of the Chongqing Municipal Education Commission of P.R. China under Grants KJZD-K201800701, KJ1705121, and KJ1705139, and the Program of Chongqing Innovation and Entrepreneurship for Returned Overseas Scholars of P.R. China under Grant cx2018110.

Supplementary Materials

data.txt： contains the data used for training and prediction in this research; it is from the power load data of Chongqing Tongnan Electric Power Co., Ltd., dating from May 4 to May 10 in 2015. (*Supplementary Materials*)

References

[1] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.

[2] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," *Expert Systems with Applications*, vol. 120, pp. 103–115, 2019.

[3] B. Lutnick, B. Ginley, D. Govind et al., "An integrated iterative annotation technique for easing neural network training in medical image analysis," *Nature Machine Intelligence*, vol. 1, no. 2, pp. 112–119, 2019.

[4] L. Zhao, Q. Wang, Q. Zou et al., "Privacy-preserving collaborative deep learning with unreliable participants," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1486–1500, 2019.

[5] Q. Feng, D. He, Z. Liu, H. Wang, and K.-K. R. Choo, "SecureNLP: a system for multi-party privacy-preserving natural language processing," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3709–3721, 2020.

[6] N. Agrawal, S. A. Shahin, M. J. Kusner et al., "QUOTIENT: two-party secure neural network training and prediction," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1231–1247, London, UK, November 2019.

[7] S. Sayyad, "Privacy Preserving Deep Learning Using Secure Multiparty Computation," in *Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 139–142, Coimbatore, September 2020.

[8] M. Izabachène, R. Sirdey, and M. Zuber, "Practical Fully Homomorphic Encryption for Fully Masked Neural Networks," in *Proceedings of the International Conference on Cryptology and Network Security*, pp. 24–36, Fuzhou, China, October 2019.

[9] T. N. Yelina, S. V. Bezzateev, and V. A. Mylnikov, "The Homomorphic Encryption in Pipelines Accident Prediction by Using Cloud-Based Neural Network," in *Proceedings of the 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*, pp. 1–4, IEEE, Saint Petersburg, Russia, June 2019.

[10] F. Ö. Çatak, "Secure multi-party computation-based privacy preserving extreme learning machine algorithm over vertically distributed data," in *Proceedings of the International Conference on Neural Information Processing*, pp. 337–345, Springer, Istanbul, Turkey, November 2015.

[11] D. Reich, A. Todoki, R. Dowsley et al., "Privacy-preserving classification of personal text messages with secure multiparty computation," *Advances in Neural Information Processing Systems*, pp. 3757–3769, 2019.

[12] R. Gilad-Bachrach, N. Dowlin, K. Lain et al., "Cryptonets: applying neural networks to encrypted data with high throughput and accuracy," in *Proceedings of the International Conference on Machine Learning*, pp. 201–210, New York, NY, USA, June 2016.

[13] Y. Aono, T. Hayashi, L. Wang et al., "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2017.

[14] F. Bourse, M. Minelli, M. Minihold et al., "Fast Homomorphic Evaluation of Deep Discretized Neural Networks," in *Proceedings of the Annual International Cryptology Conference*, pp. 483–512, Springer, Santa Barbara, CA, USA, August 2018.

[15] I. Chillotti, N. Gama, M. Georgieva et al., "Faster Fully Homomorphic Encryption: Bootstrapping in Less than 0.1 Seconds," in *Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security*, pp. 3–33, Springer, Hanoi, Vietnam, December 2016.

[16] Z. Brakerski, "Fully Homomorphic Encryption without Modulus Switching from Classical GapSVP," in *Proceedings of the Annual Cryptology Conference*, pp. 868–886, Springer, Santa Barbara, CA, USA, August 2012.

[17] J. H. Cheon, A. Kim, M. Kim et al., "Homomorphic Encryption for Arithmetic of Approximate Numbers," in *Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security*, pp. 409–437, Springer, Hong Kong, China, December 2017.

[18] H. Chen, W. Dai, M. Kim et al., "Efficient multi-key homomorphic encryption with packed ciphertexts with application to oblivious neural network inference," in *Proceedings of the Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 395–412, London, UK, November 2019.

[19] F. Bu, Y. Ma, Z. Chen et al., "Privacy Preserving Back-Propagation Based on BGV on Cloud," in *Proceedings of the 2015 IEEE 17th International Conference on High Performance Computing and Communications*, pp. 1791–1795, IEEE, Munich, Germany, September 2015.

[20] K. Sarpatwar, N. K. Ratha, K. Nandakumar et al., "Privacy Enhanced Decision Tree Inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34-35, 2020.

[21] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(Leveled) fully homomorphic encryption without bootstrapping," *ACM Transactions on Computation Theory*, vol. 6, no. 3, pp. 1–36, 2014.

[22] F. Boemer, R. Cammarota, D. Demmler et al., "MP2ML: a mixed-protocol machine learning framework for private inference," in *Proceedings of the 15th International Conference on Availability*, pp. 1–10, Boras, Sweden, March 2020.

[23] F. Boemer, A. Costache, R. Cammarota et al., "nGraph-HE2: a high-throughput framework for neural network inference on encrypted data," in *Proceedings of the 7th ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, pp. 45–56, London, UK, November 2019.

[24] M. Azraoui, M. Bahram, B. Bozdemir et al., "SoK: Cryptography for Neural Networks," in *Proceedings of the IFIP International Summer School on Privacy and Identity Management*, pp. 63–81, Springer, 2019.

[25] E. Shishniashvili, L. Mamisashvili, and L. Mirtskhulava, "Enhancing IoT security using multi-layer feedforward neural network with tree parity machine elements," *International Journal of Simulation--Systems, Science & Technology*, vol. 21, no. 2, pp. 371–375, 2020.

[26] M. S. Sruthi and A. A. TV, "Protected entry design for data encryption and decryption using big data in cloud," *International Journal of Emerging Technology and Innovative Engineering*, vol. 5, no. 3, pp. 1–7, 2019.

[27] A. López-Alt, E. Tromer, and V. Vaikuntanathan, "On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption," in *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, pp. 1219–1234, NY, New York, May 2012.

[28] J. Li, X. Kuang, S. Lin, X. Ma, and Y. Tang, "Privacy preservation for machine learning training and classification based on homomorphic encryption schemes," *Information Sciences*, vol. 526, pp. 166–179, 2020.

[29] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.

[30] S. F. Sun, X. Yuan, J. K. Liu et al., "Practical backward-secure searchable encryption from symmetric puncturable encryption," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 763–780, Toronto, Canada, October 2018.

[31] L. Gu and S. Zheng, "Conjugacy systems based on nonabelian factorization problems and their applications in cryptography," *Journal of Applied Mathematics*, vol. 2014, pp. 1–10, Article ID 630607, 2014.

[32] O. Çetin, F. Temurtaş, and Ş. Gülgönül, "An application of multilayer neural network on hepatitis disease diagnosis using approximations of sigmoid activation function," *Dicle Medical Journal/Dicle Tip Dergisi*, vol. 42, no. 2, pp. 150–157, 2015.

[33] K. Zhang, K. Peng, S. X. Ding, Z. Chen, and X. Yang, "A correlation-based distributed fault detection method and its application to a hot tandem rolling mill process," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 3, pp. 2380–2390, 2020.

[34] K. Zhang, H. Hao, Z. Chen, S. X. Ding, and K. Peng, "A comparison and evaluation of key performance indicator-based multivariate statistics process monitoring approaches," *Journal of Process Control*, vol. 33, pp. 112–126, 2015.

[35] N. P. Smart, "The discrete logarithm problem on elliptic curves of trace one," *Journal of Cryptology*, vol. 12, no. 3, pp. 193–196, 1999.

[36] G. Panti, "A general Lagrange theorem," *American Mathematical Monthly*, vol. 116, no. 1, pp. 70–74, 2009.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6687248, 9 pages
https://doi.org/10.1155/2021/6687248

Research Article

Equipment Operational Reliability Evaluation Method Based on RVM and PCA-Fused Features

Linbo Zhu,[1] **Dong Chen**,[2] **and Pengfei Feng**[2]

[1]*School of Chemical Engineering and Technology, Xi'an Jiaotong University, Xi'an, China*
[2]*Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an Jiaotong University, Xi'an, China*

Correspondence should be addressed to Linbo Zhu; linbozhu@mail.xjtu.edu.cn

Received 1 December 2020; Revised 23 December 2020; Accepted 30 December 2020; Published 11 January 2021

Academic Editor: Yong Chen

Reliability assessment is of great significance in ensuring the safety and reducing maintenance cost of equipment. The traditional statistical method is widely used to estimate the reliability of mass equipment; however, it cannot efficiently predict the overall reliability of single or small batch equipment due to lack of failure data. This paper introduced the operational reliability concept to describe the running condition of single or small batch equipment and proposed a method based on the combination of Relevance Vector Machines (RVMs) and Principal Component Analysis (PCA) to evaluate the operational reliability. Some representative characteristic indexes of operating equipment were firstly selected, and PCA was applied to obtain a hybrid index of the equipment's running condition. Then, a RVM prediction model was trained to predict the development of the hybrid index and corresponding probability density function (PDF). Based on this, the operational reliability of the equipment was calculated by the interval integral defined by the failure threshold and the predicted value of the hybrid index. The approach was validated using the experimental test conducted on the aero-engine rotor bearings. The results show a good agreement in the evaluations of the failure time between the proposed method and the experimental test.

1. Introduction

Reliability assessment is of great significance in ensuring the safety and reducing maintenance cost of equipment. In general, the reliability of equipment includes overall reliability and operational reliability. The former is obtained by statistical analysis of a large number of equipment failure data [1], and it reflects the overall reliability of the equipment throughout the lifetime. The latter is obtained from the performance degradation information of the equipment, and it reflects the real time reliability of the equipment [2]. In engineering field, single or small batch equipment is widely used, and failure data are very scarce in this case, so the traditional statistical method is not suitable for the reliability evaluation of this type of mechanical equipment, such as high precision NC machine tools, nuclear power facilities, aircraft [3–5], and so on. In contrast, the operational reliability is more of practical significance to ensure the safety of this type of equipment because it can be obtained in real time when the condition monitoring data are sampled continuously.

The operational reliability of the equipment is evaluated based on the performance degradation information, which is normally extracted from the monitoring data of the equipment [6]. The operating condition information obtained from the monitoring data is supposed to effectively reflect the evolution process of the dynamic operating characteristics of the equipment. In general, the equipment is considered to be failed or unreliable when some of its important performance parameters (such as vibration, noise, and so on) gradually reduced to the critical threshold [7, 8]; therefore, the relationship between performance parameters and operational reliability can be established.

There are three basic steps in evaluating the operational reliability of the equipment based on the performance degradation information. First, the characteristic index which reflects the operating condition is selected, and then the operational reliability calculation model based on the

characteristic index is established. Third, the reliability of the equipment is predicted based on various prediction algorithms, such as artificial neural network [9], Support Vector Machines (SVMs) [10], Relevance Vector Machines (RVMs) [11], and so on. For the problem of characteristic index selection, the most related characteristic index to the health of the equipment is usually selected. For example, Casandra et al. [12] chosen the kurtosis as the bearing status indicators to evaluate the reliability of aero-engine rotor bearings. Shaban et al. [13] selected wear amount of tool as the performance degradation characteristic of cutting tools. Zi et al. [14] used the radial runout of the spindle end as the characteristic index of the electric spindle performance degradation. Li et al. [15] selected the waveform index of bearing vibration signal and used the energy obtained from wavelet packet decomposition as the degradation index of bearings to establish the reliability model. In the above references, only single characteristic index is used in operational reliability evaluation; this may result in the lack of robustness of the assessment method. Several references show that the reliability estimation accuracy can be improved by using multiple characteristic parameters. Widodo and Yang [11] introduced multiple parameters in reliability evaluation of aero-engine rotor bearings by using the Principal Component Analysis (PCA) method to fuse peak, kurtosis, and entropy to a hybrid index. Zheng et al. [16] also employed PCA to combine 10 variables, such as power system equipment availability coefficient, power supply reliability rate, capacity-to-load ratio, and so on, to estimate the reliability of the power supply system, and the reliability prediction accuracy was proved to be improved by adopting the hybrid index.

After selection of the characteristic index, the operational reliability can be obtained by several methods. One widely used method is to calculate the interval integral of the probability density function (PDF) of the selected index between the failure threshold and the observed value of the index [9]. In this method, the model of PDF of the selected index should be assumed and estimated firstly. Wang and Dragomir-Daescu [17] assumed that the PDF of wear data of the bearing in induction generators is a two-parameter Weibull distribution. Zi et al. [14] compared normal distribution, the logarithmic distribution, and the Weibull distribution in estimation the operational reliability of the spindle. Schömig and Infineon [18] obtained the fault data through simulation experiments and compared the accuracy of the reliability evaluation when the fault data obey the gamma distribution, exponential distribution, and Weibull distribution. The results show that the Weibull distribution is the best distribution in evaluation the operational reliability of semiconductor manufacturing equipment. It is easy to know that the accurate evaluation of PDF of the selected index is a critical issue in operational reliability estimation; however, the estimation accuracy of PDF is highly dependent on the number of samples.

Another commonly used method of operational reliability assessment is the K-M evaluator, in which the continued product of the ratio between the number of normal samples and the total number of samples is taken as the operational reliability of the equipment [19]. The main advantage of the K-M evaluator is that its computation process does not involve the estimation of PDF of samples, and the evaluation of operational reliability of equipment is simplified. Heng et al. [9] took the difference between the suspended data and CM (condition monitoring) data into consideration and combined the K-M and PDF methods to calculate the reliability of centrifugal pumps. He et al. [20] used the K-M estimator and the proportional hazards model to estimate the reliability of the engine exhaust valve based on mean air pressure, maximum coolant temperature, maximum fuel temperature, and other indexes. Fang et al. [21] used the K-M estimator to evaluate the reliability of the CNC honing hydraulic system based on pump output flow value. Based on tools wear amount, Chen et al. [22] estimated the reliability of CNC machine tools by combining Bayes and K-M estimators. However, the accuracy of the K-M evaluator still depends on the number of observed failure samples, so it is also limited in applying in the single or small batch equipment.

In order to overcome the shortcomings of the traditional reliability evaluation methods, the RVM and PCA methods are introduced in this paper to develop a new evaluation method of the operational reliability for single or small patch equipment. RVM is a Bayesian-based machine learning method proposed by Tipping [23]; it is often adopted as a predictor in reliability prediction. For example, it is used in bearing reliability prediction [11], battery reliability prediction [24], and software reliability prediction [25]. In fact, RVM can also estimate the posterior probability of the predicted object at each prediction step. If it is applied to predict the characteristic index of equipment, the value of the characteristic index as well as its PDF can be obtained simultaneously. Based on the predicted PDF and the preset failure threshold, the operational reliability of the equipment can be obtained. At the same time, the PCA method is used to combine several characteristic indexes into one hybrid index to increase the robustness and the accuracy of the operational reliability prediction. The steps of method include the following. First, some representative characteristic indexes of running equipment are selected, and PCA is applied to these indexes to get the hybrid index. Then, the series of the hybrid index of long-term monitoring is used to train a single-step prediction RVM model to predict the value and probability density function (PDF) of the next step. Third, the operational reliability of the equipment is calculated by the interval integral defined by the failure threshold and the predicted value of the hybrid index.

In this method, only the performance degradation information of the object equipment is required, and there is no any other information from the same type equipment used, so the method is suitable for the reliability estimation problem of single equipment. The rest of the paper is organized as follows. Section 2 discusses RVM, PCA, and operational reliability calculation model and presents the framework of the methodology. Section 3 describes validation experiments, and Section 4 shows the results and discussion. Section 5 gives the conclusions.

2. Theoretical Method

2.1. RVM Regression Model. Regression problem is defined as follows. Given an input x_i ($i = 1,2,\ldots, N$), by using the regression model, we can get its output t_i($i = 1,2,\ldots, N$), that is, a set of sample set $\{x_i, t_i\}_{i=1}^{N}$ satisfies the following relationship:

$$t_i = y(x_i, \boldsymbol{\omega}) + \varepsilon_i, \tag{1}$$

where ε_i is the prediction noise of x_i and is assumed to be zero-mean Gaussian distribution with variance σ^2. Moreover, $y(x_i, \boldsymbol{\omega})$ can be expressed [26] as follows:

$$y(x_i, \boldsymbol{\omega}) = \sum_{i=1}^{N} \omega_i K(x, x_i) + \omega_0, \tag{2}$$

where ω_i is an adjustable weight, ω_0 is bias, and $K(x, x_i)$ is the corresponding kernel function; it is used to map the inputs x_i from nonlinear space to high dimensional space and perform the linear regression in this space. The likelihood function of the training sample set is as follows:

$$p\left(\mathbf{t}|\boldsymbol{\omega}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-(N/2)} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\omega}\|^2\right\}, \tag{3}$$

where $\Phi = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]^T$ is the $N \times (N+1)$ design matrix with $\phi(x_i) = [1, K(x_i, x_1), K(x_i, x_2), \ldots, K(x_i, x_N)]^T$, x_i, $i = 1, 2, \ldots, N$. In order to avoid the problem of over-learning, hyperparameter α_i is introduced to each weight coefficient ω_i.

Supposing ω_i obeys a zero-mean Gaussian with variance α^{-1}, then we can get the following:

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{i=0}^{N} N\left(\omega_i|0, \alpha_i^{-1}\right). \tag{4}$$

According to the Bayesian formula, we can get the posterior distribution function of the weight ω with the likelihood function equation (3) and a priori distribution function equation (4). That is,

$$p\left(\boldsymbol{\omega}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2\right) = \frac{p\left(\mathbf{t}|\boldsymbol{\omega}, \sigma^2\right)p(\boldsymbol{\omega}|\boldsymbol{\alpha})}{p\left(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2\right)} = N(\boldsymbol{\omega}|\mu, \textstyle\sum), \tag{5}$$

where the posterior covariance and mean are, respectively, as follows:

$$\begin{aligned} \Sigma &= \left(\sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathbf{A}\right)^{-1}, \\ \mu &= \sigma^{-2}\textstyle\sum \boldsymbol{\Phi}^T\mathbf{t}. \end{aligned} \tag{6}$$

In equation (6), $\mathbf{A} = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_N)$. The optimized hyperparameters α_{MP} and σ_{MP} can be obtained by maximizing the marginal likelihood function $p(\mathbf{t}|\alpha, \sigma^2)$ with respect to α and σ [11]:

$$\max_{\alpha_{MP}, \sigma_{MP}} p\left(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2\right) = \int p\left(\mathbf{t}|\boldsymbol{\omega}, \sigma^2\right)p(\boldsymbol{\omega}|\boldsymbol{\alpha})d\boldsymbol{\omega} = (2\pi)^{-(N/2)}\left|\sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\Phi^T\right|^{-(1/2)} \exp\left\{-\frac{1}{2}t^T\left(\sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\Phi^T\right)^{-1}\mathbf{t}\right\}. \tag{7}$$

Given a new input value x_*, the target output is as follows:

$$p\left(t_*|x_*, \alpha_{MP}, \sigma_{MP}^2\right) = \int p\left(t_*|\boldsymbol{\omega}, \sigma_{MP}^2\right)p\left(\boldsymbol{\omega}|\mathbf{t}, \alpha_{MP}, \sigma_{MP}^2\right)d\boldsymbol{\omega}. \tag{8}$$

which can be easily computed since both integrated terms obeys Gaussian distribution, that is,

$$p\left(t_*|x_*, \alpha_{MP}, \sigma_{MP}^2\right) = N\left(t_*|y_*, \sigma_*^2\right), \tag{9}$$

where the mean and variance are, respectively, as follows:

$$\begin{aligned} y_* &= \mu^T\phi(x_*), \\ \sigma_*^2 &= \sigma_{MP}^2 + \phi(x_*)^T\textstyle\sum \boldsymbol{\Phi}(x_*). \end{aligned} \tag{10}$$

Therefore, the RVM learning method based on the Bayesian framework can be used to predict the probability and obtain the forecast value and its probability distribution. If the input and output data are from a time series (for example, precipitation data), RVM can be used to predict the future value of the time series.

2.2. Operational Reliability Calculation Model. The traditional reliability theory can only provide the overall reliability assessment for mass equipment, but in engineering field, single or small batch equipment is widely used and failure data are very scarce in this case; people are more concerned about the reliability of the particular equipment in operation. To solve the reliability evaluation problem of single or small batch equipment, Heng et al. [9] proposed an operational reliability model based on the equipment running characteristic index, and the model is described as follows.

Let $Y_i(t)$ be the value of the condition characteristic index for equipment i at operating age t, and $\mathbf{Y}(t) = [Y_1(t), Y_2(t), \ldots, Y_m(t)]^T$ containing the condition values from all of the m historical equipment in interval t, and Y_{thresh} be the threshold of the characteristic index; if $Y_i(t) > Y_{\text{thresh}}$, the equipment is considered to be failed. $f(Y|t)$ is the corresponding PDF of $\mathbf{Y}(t)$; the overall reliability at time t is defined as follows:

$$R(t) = P\left(Y(t) < Y_{\text{thresh}}\right) = \int_0^{Y_{thresh}} f(Y|t)dY. \tag{11}$$

The operational reliability of equipment i at time $t + k\Delta t$ is defined as follows:

$$R_i(t + k\Delta t) = \prod_{j=1}^{k} \frac{P\left[y_{\text{thresh}} > Y_i(t + j\Delta) \geq y_{i,t+j\Delta} | y_{\text{thresh}} > Y_i(t + (j-1)\Delta) \geq y_{i,t+(j-1)\Delta}, \cdots\right]}{P\left[Y_i(t + j\Delta) \geq y_{i,t+j\Delta} | y_{\text{thresh}} > Y_i(t + (j-1)\Delta) \geq y_{i,t+(j-1)\Delta}, \cdots\right]} = \prod_{j=1}^{k} \frac{\int_{y_{i,t+j\Delta}}^{y_{\text{thresh}}} f(y|t + j\Delta)dy}{\int_{y_{i,t+j\Delta}}^{\infty} f(y|t + j\Delta)dy}, \quad (12)$$

where $\int_{y_{i,t+j\Delta}}^{y_{\text{thresh}}} f(y|t + j\Delta)dy$ is the integral of the PDF between the observed degradation index of device i and the threshold and $\int_{y_{i,t+j\Delta}}^{\infty} f(y|t + j\Delta)dy$ is the integral of the PDF over all possible values equal to or higher than the observed degradation index of device i.

In summary, if the value of the characteristic index and the corresponding PDF can be obtained, the operational reliability of equipment can be calculated by equation (12).

2.3. PCA. In the process of reliability analysis, several vibration characteristics of the equipment can be obtained simultaneously. It will be too complicated to evaluate the reliability if we choose all of them. To improve the stability of reliability estimation, an intuitive approach is to fuse all characteristics to conduct the reliability analysis [27]. In this paper, the PCA method is adopted to fulfill this requirement. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. If the vibration characteristics have similar trends with respect to time, the first principal component will contain the most information of all characteristics and can be used as the reliability index.

The principal component analysis method is obtained by projecting the original vector as a new coordinate space consisting of the eigenvectors of the covariance matrix of the original variables [28, 29]. For a given vibration characteristic vector set of m features $\mathbf{X} = \{x_1, \cdots, x_m\}$, $x_i \in R^n$ which consists of feature vectors, the covariance matrix C is given as equation (13), in which μ is mean value of x_i:

$$C = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T. \quad (13)$$

Then, the eigenvalues $\lambda_i (i = 1, 2, \ldots, n)$ and corresponding eigenvectors $v_i (i = 1, 2, \ldots, n)$ of C are given as follows:

$$Cv_i = \lambda_i v_i. \quad (14)$$

Sort the eigenvalues in decreasing order $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_n$, and composite the first k eigenvalues $\Delta = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ and the corresponding eigenvectors $\mathbf{V} = (v_1, v_2, \ldots, v_k)$ and then transform the original data $\mathbf{X}$ onto the new subspace $\mathbf{V}$ and get the transformed data $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{V}^T \mathbf{X}. \quad (15)$$

The number of principle component is selected depending on the cumulative contribution R_k, which is usually set more than 85%–90%:

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}. \quad (16)$$

2.4. Operational Reliability Prediction Methodology. As described above, the method employs performance degradation data of equipment which are obtained from the continuously monitoring of equipment's running condition. Some representative characteristic indexes of equipment, which can describe the degradation process of equipment from normal to failure, are selected. PCA is applied to deal with these characteristic indexes to obtain the hybrid index. The RVM regression model is then trained to obtain the hybrid index. Based on the RVM regression method, the posterior distribution function and the predicted value of the characteristic index in different time can be obtained. Finally, the operational reliability can be calculated by the mentioned operational reliability calculation model (Section 2.2). The total process is illustrated in Figure 1.

3. Experimental Investigation

The data of fatigue life experiments of aero-engine rotor bearings [30] are adopted to validate the proposed method. In experiments, four bearings were installed on a shaft. The rotation speed was kept constant at 2000 rpm, radial load of 6000 lbs is applied onto the shaft, and all bearings were force lubricated. Three different types failure are tested. The first failure type is inner ring failure; it comes from the 3# bearing (it is named A bearing in the paper). The second failure type is composite failure of outer ring and rolling element; it comes from the of 4# bearing (named B bearing), and the third failure is outer ring spalling of 1# bearing (named C bearing). The data were collected every 20 minutes, and the sampling frequency is 12 kHz in each measurement cycle. The measurement time is 43000 minutes, and the data length is 2150 points. It means that the measurement time is 20 times the measurement points. The selected data in the paper come from reference [30], but they do not include the one of the first 5 days considering the instability of the previous data.

Several time-domain characteristic indexes of bearing A are shown in Figure 2. From RMS and waveform curves, it can be observed that at the beginning of experiments, the curves keep stable (the small change at the very beginning is considered to be caused by the instability of the running stage), and the main fluctuations are happened near to 1800 measurement points that means the early failure is happed in the bearing. This trend can also be found in kurtosis and peak index. While for mean value, it can be observed that its

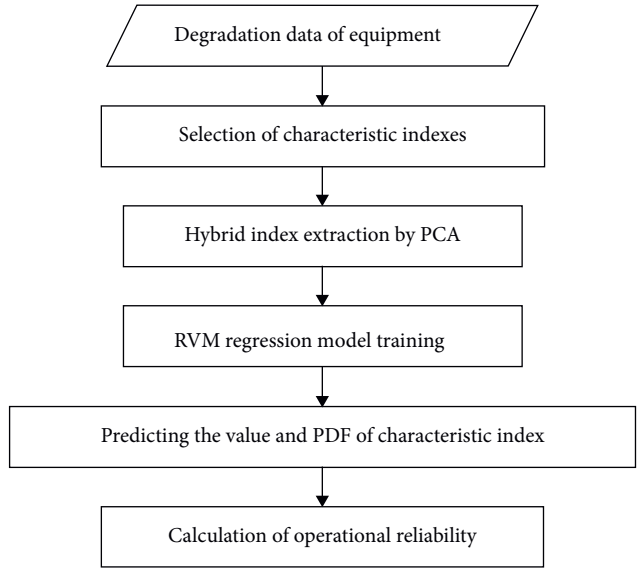


Figure 1: Operational reliability prediction process.

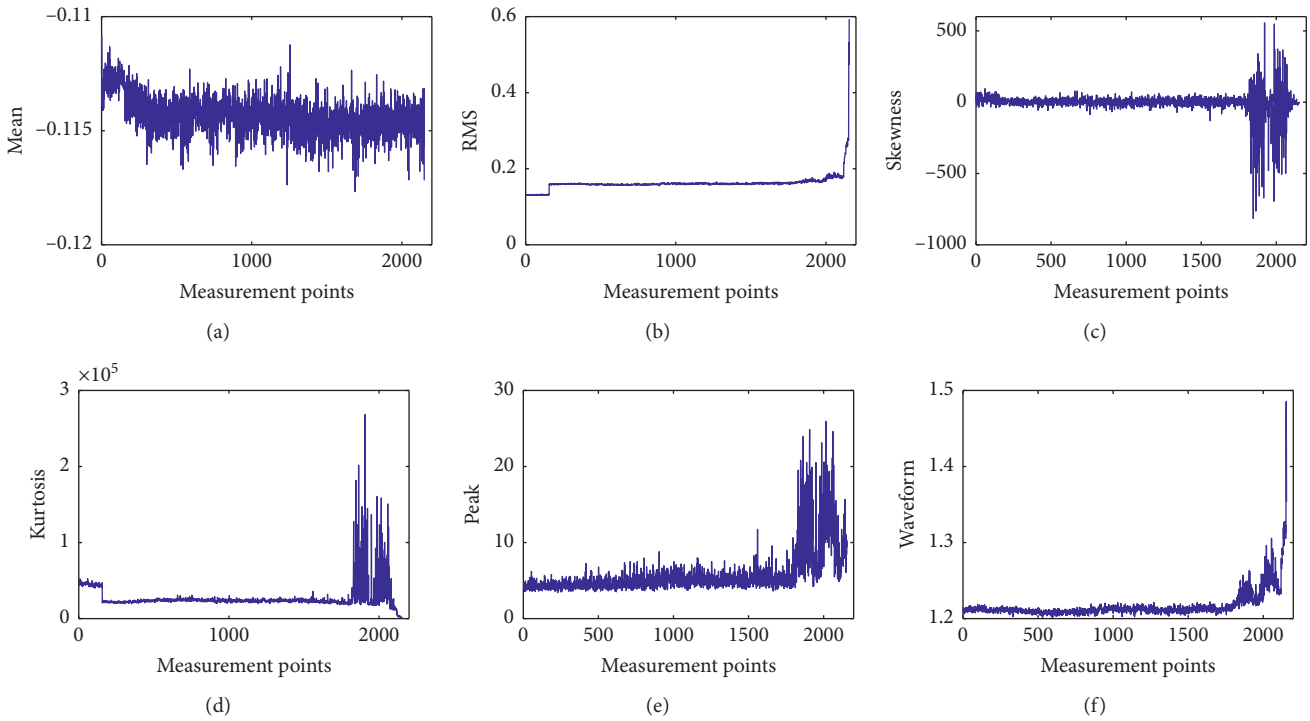


Figure 2: Vibration signal time-domain indexes: (a) mean, (b) RMS, (c) skewness, (d) kurtosis, (e) peak, and (f) waveform.

value decreases with the running of experiment and is not sensitive to the failure. The skewness index oscillates sharply when the early failure happens, which does not conform to our expectation of continuous increase or continuous reduction of the selected index in operational reliability evaluation.

From Figure 2, it is also can be found that different characteristic indexes have different sensitivity to the failure. Though the RMS index and waveform index rise sharply with the increase in the severity of the failure, the changes on the trend of the curves caused by early failure are relatively small compared with the peak index and kurtosis index. As for the kurtosis index, it increases at the early stage of failure, while decreases with the development of the failure. To make the characteristic index both sensitive and robust to the failure, RMS, waveform index, and peak value are selected and combined with the PCA method, and the obtained hybrid index of bearing A, B, and C is shown in Figure 3.

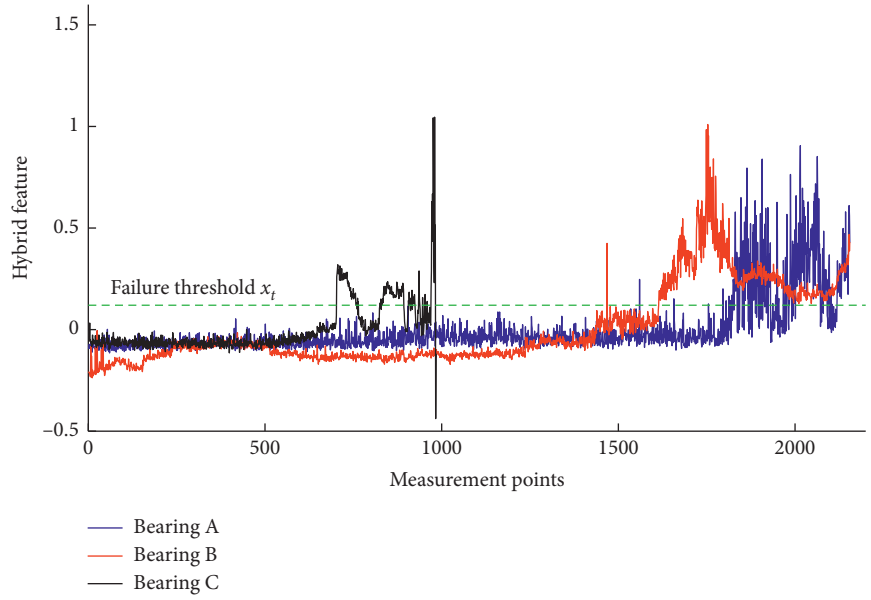


Figure 3: Hybrid index of three bearing.

An empirical threshold $x_t = 0.12$ is assumed and plotted in Figure 3 with dotted line. It can be observed that when the hybrid index is smaller than the threshold, the hybrid index changes more smoothly, and the bearing is supposed to work in normal condition. When the threshold is exceeded, the hybrid index increases rapidly and has a relatively violent fluctuation, which means the happening of failure in the bearing. The failure time of the three bearings determined by the failure threshold is 703, 1614, and 1821 points, respectively.

4. Results and Discussion

The hybrid index of the first 1500 points of bearing A, the first 1500 points of bearing B, and the first 650 points of bearing C are selected, respectively, to train the single-step RVM prediction model for each bearing. The embedding dimension is set to 20, the RVM kernel parameter optimization range is [0.1, 20], and the optimization objective fitness function is as follows [31]:

$$\text{fitness} = \frac{\left(\text{RMS}_{\text{train}} * n_1 + \text{RMS}_{\text{test}} * n_2\right) * \text{RV}}{n_1 + n_2}, \tag{17}$$

where $\text{RMS}_{\text{train}}$ and RMS_{test} are training errors and the test errors and are obtained with the 5-folds cross-validation method, respectively. n_1 and n_2 are the number of training samples and the number of test samples, and RV is the number of relevance vectors. Minimizing this fitness function makes the trade-off of kernel parameter between training and test errors and makes the trained model to have the best prediction accuracy. The obtained optimized kernel parameters for each model are given in Table 1, and the prediction accuracy is defined as follows:

$$\text{accuracy} = \left(1 - \frac{\left|t_a - t_p\right|}{t_a}\right) \times 100\%, \tag{18}$$

Table 1: Optimization results of kernel parameters for the RVM model.

RVM model	Kernel parameter	Fitness
Bearing A	16.8	0.0334
Bearing B	6.2	0.07
Bearing C	9.9	0.0203

where t_a is the actual failure time and t_p is the predicted failure time. It must be indicated that the failure time is 20 times the failure point for this case.

The prediction results on the hybrid index of three bearings are shown in Figure 4(a), Figure 5(a), and Figure 6(a), respectively. In each graph, the red points represent actual data and the blue points represent the predicted data. The results show that the change trend of the predicted value closely matches the actual value; it means RVM has acceptable accuracy in predicting the time series of the hybrid index. Figure 4(b), Figure 5(b), and Figure 6(b), respectively, show the prediction results on operational reliability of three bearings and corresponding enlarged diagram. It can be observed that at the early stage of bearing total lifetime, the bearings have relatively high operational reliability and relatively small degradation trend (from the start of operation to approximate 1823, 1512, and 704 measurement points of three bearings, respectively). After that, the operational reliabilities of three bearings show sharp drops, which means the initiation of a defect. At 1836 points of bearing A, 1623 points of bearing B, and 714 points of bearing C, the operational reliability is forecasted to fall close to zero, meaning that the bearings are closed to complete failure. This degradation process indicates that the occurrence and development of bearing fault is a rapid process and will happen in very short time, compared with the total service life of the bearing.

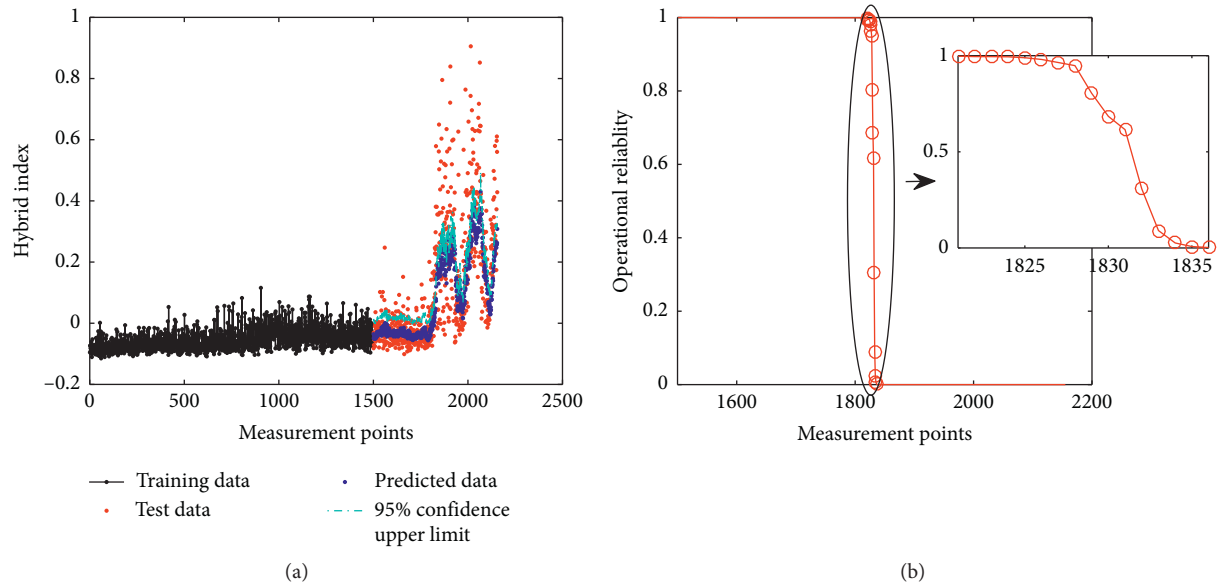


Figure 4: Prediction results of bearing A: (a) hybrid index and (b) operational reliability.

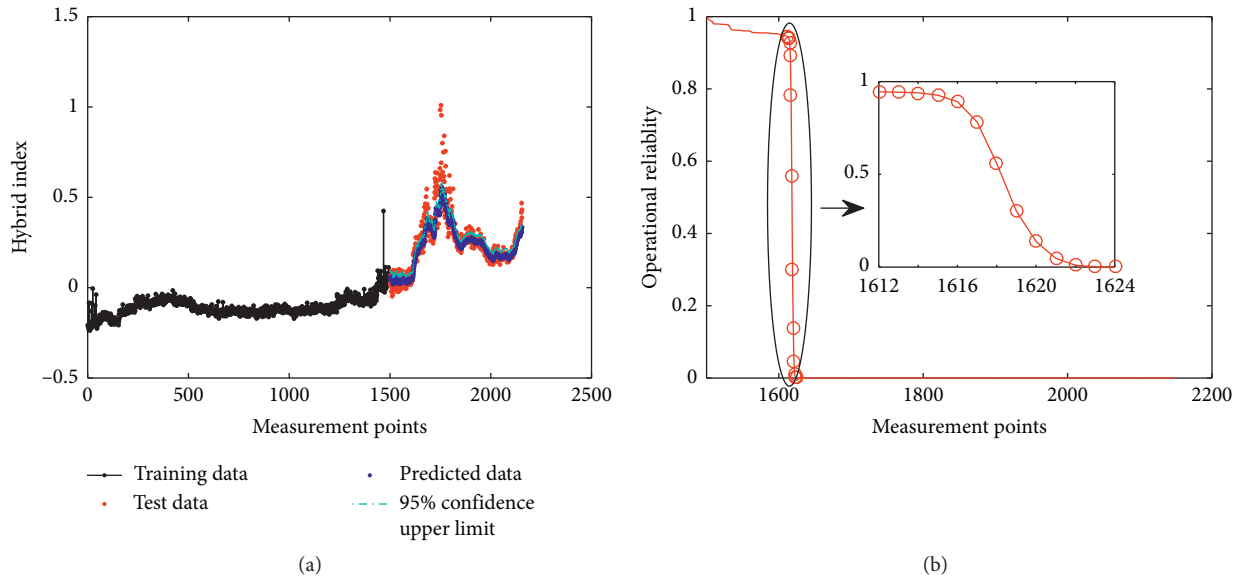


Figure 5: Prediction results of bearing B: (a) hybrid index and (b) operational reliability.

It is also noted that three bearings have experienced different time from initiation of the defect to final failure because of the difference of bearing fault. In fact, this failure mechanism and fault development process can also be observed in Figure 3, in which the similar evolutionary process of the hybrid index is shown; it proves that the RVM prediction models have learned the failure pattern of three bearings. The prediction accuracy obtained from the reliability curve with equation (18) is shown in Table 2. The results show that the prediction accuracy on failure time of three bearings is all above 98%. The results suggest that the operational reliability evaluation model captures the nonlinear relationship between the hybrid index and the actual health state of the monitored equipment.

In addition, in order to verify the validity of the hybrid index, this paper also compares the reliability evaluation results based on the single index with that on the hybrid index. Table 3 gives the results on bearing C as an example; it was shown that the prediction accuracy of the failure time obtained with the hybrid index is higher than that obtained with the single index, indicating that the hybrid index can synthetically consider a variety of information and obtain better evaluation results.

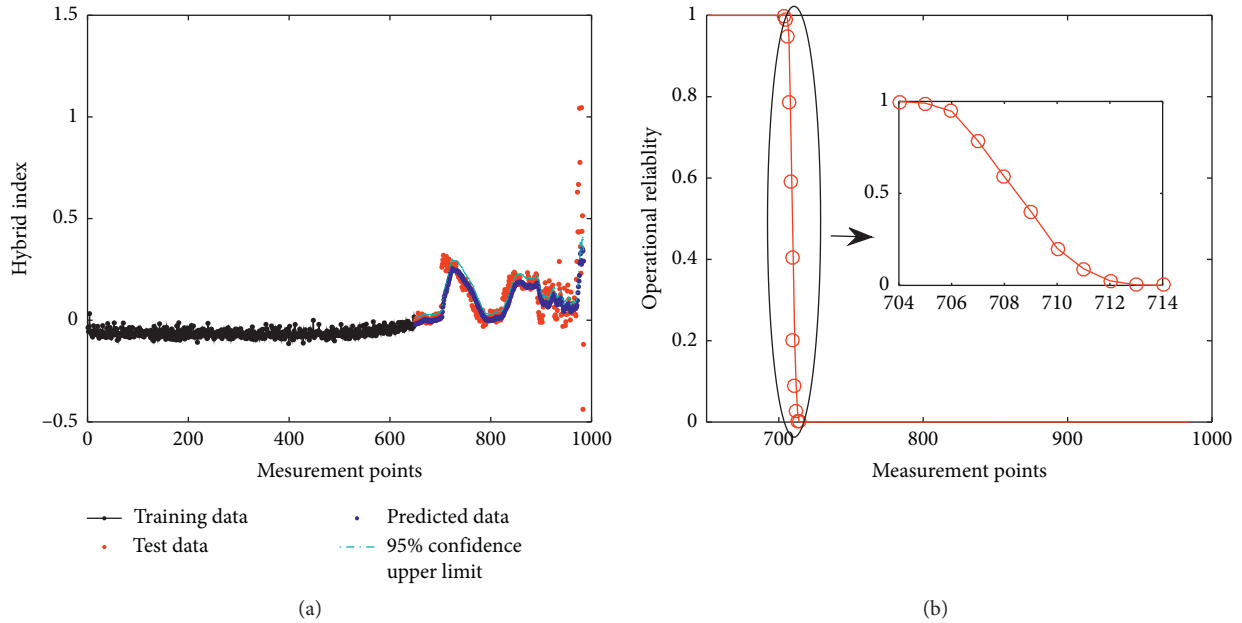


Figure 6: Prediction results of bearing C: (a) hybrid index and (b) operational reliability.

Table 2: Three bearing failure time prediction results.

Bearing	Actual failure time (min)	Predicted failure time (min)	Accuracy (%)
A	36420	36720	99.2
B	32280	32460	99.4
C	14060	14280	98.4

Table 3: Single index method and hybrid index method prediction results of bearing C.

Characteristic index	Actual failure time (min)	Predicted failure time (min)	Accuracy (%)
Waveform	14060	14380	97.7
Peak	14080	14360	98.0
Kurtosis	14080	14480	97.2
Hybrid index	14060	14280	98.4

5. Conclusions

This paper presents a method to evaluate the operational reliability for single or small patch equipment based on RVM and PCA. The PCA was used to fuse the features and obtain the hybrid index which can represent the degradation information of the equipment more robustly. The RVM was used to establish a single-step prediction model of the hybrid index and predict the future value of the hybrid index and the corresponding PDF at a certain moment. Based on PDF and the predicted value of the hybrid index, the operational reliability of the equipment is obtained with the interval integral defined by the failure threshold and the predicted value of the hybrid index. The performance of the proposed method is validated by predicting the failure time of aero-engine rotor bearings. This paper also compares the reliability evaluation results based on the hybrid index with the reliability evaluation results based on a single index. The results proves the plausibility and effectiveness of the proposed method.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (nos. 51805412 and 51635010), Science and Technology Major Project of Henan Province (no. 191110213300), National Key Research and Development Project (no. 2020YFB2007900), and China Postdoctoral Science Foundation (nos. 2018M631144 and 2019T120897).

References

[1] P. D. T. O'Connor, "Commentary: reliability-past, present, and future," *IEEE Transactions on Reliability*, vol. 49, no. 4, pp. 335–341, 2000.

[2] Z. He, H. Cao, Y. Zi, and B. Li, "Developments and thoughts on operational reliability assessment of mechanical equipment," *Journal of Mechanical Engineering*, vol. 50, no. 2, pp. 171–186, 2014.

[3] T.-H. Fan and C.-C. Chang, "A Bayesian zero-failure reliability demonstration test of high quality electro-explosive devices," *Quality and Reliability Engineering International*, vol. 25, no. 8, pp. 913–920, 2009.

[4] M. Balesdent, J. Morio, and J. Marzat, "Recommendations for the tuning of rare event probability estimators," *Reliability Engineering & System Safety*, vol. 133, pp. 68–78, 2015.

[5] H. Li, F. Chen, Z. Yang, Y. Kan, and L. Wang, *Bayesian Reliability Assessment Method for Single NC Machine Tool under Zero Failures*, pp. 291–302, Springer, Berlin, Heidelberg, 2015.

[6] J. C. Ferreira, M. A. Freitas, and E. A. Colosimo, "Degradation data analysis for samples under unequal operating conditions: a case study on train wheels," *Journal of Applied Statistics*, vol. 39, no. 12, pp. 2721–2739, 2012.

[7] D. Xu and W. Zhao, "Reliability prediction using multivariate degradation data," in *Proceedings of Conference on Reliability and Maintainability Symposium*, pp. 337–341, Alexandria, VA, USA, January 2005.

[8] M. S. Chang, Y. I. Kwon, and B. S. Kang, "Design of reliability qualification test for pneumatic cylinders based on performance degradation data," *Journal of Mechanical Science and Technology*, vol. 28, no. 12, pp. 4939–4945, 2014.

[9] A. Heng, A. C. C. Tan, J. Mathew, N. Montgomery, D. Banjevic, and A. K. S. Jardine, "Intelligent condition-based prediction of machinery reliability," *Mechanical Systems and Signal Processing*, vol. 23, no. 5, pp. 1600–1614, 2009.

[10] L. Chato, S. Tayeb, and S. Latifi, "A genetic algorithm to optimize the adaptive support vector regression model for forecasting the reliability of diesel engine systems," in *Proceedings of IEEE Conference on Computing and Communication Workshop and Conference*, pp. 1–7, Las Vegas, NV, USA, January 2017.

[11] A. Widodo and B.-S. Yang, "Application of relevance vector machine and survival probability to machine degradation assessment," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2592–2599, 2011.

[12] W. Caesarendra, A. Widodo, and B.-S. Yang, "Application of relevance vector machine and logistic regression for machine degradation assessment," *Mechanical Systems and Signal Processing*, vol. 24, no. 4, pp. 1161–1171, 2010.

[13] Y. Shaban, S. Yacout, and M. Aly, "Condition-based reliability prediction based on logical analysis of survival data," in *Proceedings of IEEE Conference on Reliability and Maintainability Symposium*, pp. 1–6, Orlando, FL, USA, January 2017.

[14] J. Zi, H. Liu, X. Jiang, and L. Liu, "Reliability assessment of electric spindle based on degradation values distribution," *China Mechanical Engineering*, vol. 25, pp. 807–812, 2014.

[15] H.-K. Li, Z.-X. Zhang, X.-G. Li, and Y.-J. Ren, "Reliability prediction method based on state space model for rolling element bearing," *Journal of Shanghai Jiaotong University (Science)*, vol. 20, no. 3, pp. 317–321, 2015.

[16] Y. Zheng, G. Sun, Z. Wei, F. Zhao, and Y. Sun, "A novel power system reliability predicting model based on PCA and RVM," *Mathematical Problems in Engineering*, vol. 2013, Article ID 648250, , 2013.

[17] W. Wang and D. Dragomir-Daescu, "Reliability quantification of induction motors-accelerated degradation testing approach," in *Proceedings of IEEE Conference on Reliability and Maintainability Symposium*, pp. 325–331, Seattle, WA, USA, February 2002.

[18] A. K. Schömig and Infineon, *On the Suitability of the Weibull Distribution for the Approximation of Machine Failures*, in Proceedings of IIE Annual Conference, pp. 1-7, Oregon, Portland, USA, May 2003.

[19] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[20] Y. He, A. Kusiak, T. Ouyang, and W. Teng, "Data-driven modeling of truck engine exhaust valve failures: a case study," *Journal of Mechanical Science and Technology*, vol. 31, no. 6, pp. 2747–2757, 2017.

[21] M. Fang, G. Zhou, Y. Cheng, and X. Lei, "Reliability prediction method for hydraulic system of CNC honing machine based on running status information," *Applied Science and Technology*, vol. 39, no. 6, pp. 30–33, 2012.

[22] B. Chen, X. Chen, Z. He, and B. Li, "Operating condition information-based reliability prediction of cutting tool," *Journal of Xi'an Jiaotong University*, vol. 44, no. 9, pp. 74–77, 2010.

[23] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[24] H. Li, D. Pan, and C. L. P. Chen, "Intelligent prognostics for battery health monitoring using the mean entropy and relevance vector machine," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 7, pp. 851–862, 2014.

[25] J. Lou, Y. Jiang, Q. Shen, Z. Shen, Z. Wang, and R. Wang, "Software reliability prediction via relevance vector regression," *Neurocomputing*, vol. 186, pp. 66–73, 2016.

[26] R. Mohebian, M. A. Riahi, and M. Afjeh, "Detection of the gas-bearing zone in a carbonate reservoir using multi-class relevance vector machines (RVM): comparison of its performance with SVM and PNN," *Carbonates and Evaporites*, vol. 33, no. 2-4, pp. 347–357, 2018.

[27] L. Cui, N. Wu, W. Wang, and C. Kang, "Sensor-based vibration signal feature extraction using an improved composite dictionary matching pursuit algorithm," *Sensors*, vol. 14, no. 9, pp. 16715–16739, 2014.

[28] F. Wang, J. Sun, D. Yan, S. Zhang, L. Cui, and Y. Xu, "A feature extraction method for fault classification of rolling bearing based on PCA," *Journal of Physics: Conference Series*, vol. 628, 2015.

[29] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016.

[30] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *Journal of Sound and Vibration*, vol. 289, no. 4-5, pp. 1066–1090, 2006.

[31] G. Li, G. Wang, and H. Xue, "Optimizing method to kernel function parameters of RVM," *Control Engineering of China*, vol. 17, pp. 335–337, 2010.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6624528, 12 pages
https://doi.org/10.1155/2021/6624528

Research Article

Interval Number-Based Safety Reasoning Method for Verification of Decentralized Power Systems in High-Speed Trains

Peng Wu,[1] **Ning Xiong**,[2] **Jiqiang Liu**,[1] **Liujia Huang**,[1,3] **Zhuoya Ju**,[1] **Yannan Ji**,[4] **and Jinzhao Wu**[3,5]

[1]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Innovation, Design and Engineering, Mälardalen University, 72123 Västerås, Sweden*
[3]*Institute of Artificial Intelligence, Guangxi University for Nationalities, Nanning 530006, China*
[4]*China Railway First Survey and Design Institute Group, Xi'an 710043, China*
[5]*School of Computer, Electronics and Information, Guangxi Universiy, Nanning 530004, China*

Correspondence should be addressed to Jinzhao Wu; wjzgxun@163.com

Received 6 November 2020; Revised 30 November 2020; Accepted 7 December 2020; Published 4 January 2021

Academic Editor: Yong Chen

Decentralized power systems are commonly used in high-speed trains. However, many parameters in decentralized power systems are uncertain and inevitably have errors. We present a reasoning method based on the interval numbers for decentralized power systems in high-speed trains. Uncertain parameters and their unavoidable errors are quantitatively described by interval numbers. We also define generalized linear equations with interval numbers (LAIs), which can be used to describe the movement of the train. Furthermore, it is proven that the zero sets of LAIs are convex. Therefore, the inside of the fault-tolerance area can be formed by their vertexes and edges and represented by linear inequalities. Consequently, we can judge whether the system is working properly by verifying that the current system state is in the fault-tolerance area. Finally, a fault-tolerance area is obtained, which can be determined by linear equations with an interval number, and we test the correctness of the fault-tolerance area through large-scale random test cases.

1. Introduction

Theorem proving is well established in formal verification [1, 2]. Unlike model checking [3, 4], the deductive reasoning method is used to verify the safety conditions or properties of the system. This method and model checking have complementary strengths and disadvantages [5, 6]. To verify certain properties of the system, labeled transition systems (LTSs) are widely used to describe the system behaviors in the field of system verification, such as communication protocols and hardware logic testing [7, 8], and a similar structure is used. Abstract labels of the labeled transition system (LTS) describe the system states by a set of logical assignments. For instance, the LTS model of a microwave oven, where "close" represents the state that the door of microwave oven is close, "~close" represents the state of "open door," "heat" represents the heating state, and "~heat" represents the nonheating state. Thus, (close, ~heat) indicates that the oven is closed and not heating. However, it is not completely adequate to describe the state of complex systems. For example, train motion must be described by algebraic equations Naturally, algebraic transition systems can be modeled in the above example, being labeled by polynomial algebraic equations [9, 10]. as the promotion of the logical labeled transition system. In recent years, algebraic polynomial-labeled transition systems or their similar structures are still largely involved with the verification of more complex systems [11]. Especially for verification of hybrid systems characterized by differential polynomial algebraic equations, many theories based on polynomial invariants have been put forward [12–16]. However, to find polynomial invariants, the symbolic calculation theory with high time complexity is involved, such as the Gröbner bases, cylindrical algebraic decomposition, and fixed points.

However, in some complex systems, some parameters may be uncertain. For example, in high-speed trains, some model parameters are uncertain, such as the changing weight with the number of passengers [17]. Even if it is assumed that the uncertain parameters change continuously, nonetheless, this may result in the discontinuity of the obtained Gröbner bases. Simply put, the Gröbner bases of two polynomial systems with very close coefficients may be completely different, which limits the application of the above theory to the verification of these systems. In contrast, for the design of systems with uncertain parameters, some scholars have performed system design based on fault-tolerant methods [18, 19]. The fault-tolerant method is mature and has been applied in many aspects [20–22]. The success of the fault-tolerant method in the design of complex systems implies that it may also be effective in the verification of complex systems [23–25]. Nevertheless, verification methods with fault-tolerance are rarely reported.

Furthermore, regardless of the uncertain parameters, even measurements cannot be completely accurate. For example, in a real system, it is impossible to accurately measure the temperature just to reach a specified value and often with a certain error. The measurement process is also accompanied by a certain error. Hence, verification with fault tolerance is significant in the industry. In addition, nonlinear problems can be approximated as linear problems in small parts of the system design space [26]. The generalized linear assertion also has a certain theoretical value.

In this paper, we present a new reasoning method with fault tolerance between generalized linear algebraic assertions to verify decentralized power systems in high-speed trains, and the method does not involve the methods in numerical calculations. Although the numerical calculation method is much faster than symbolic calculation to solve equations, the accumulation of errors during the reasoning process is inevitable and may lead to incorrect conclusions. In numerical calculations, the iterative algorithm for solving equations is terminated after reaching the termination condition. In fact, we still do not know the exact distance between the numerical solution and the unknown exact solution [27]. On the other hand, some scholars have studied fault detection in power systems, in which machine learning algorithms are involved [28, 29]. Their method is effective on nonlinear problems. However, there are still few reports about their methods in dealing with the system with uncertain parameters.

2. Problem Descriptions

Proper decentralized power can reduce the maintenance cost of high-speed trains and avoid unsafe speeds. Safe speed and decentralized power need to be considered.

2.1. Safe Speeds. The safe speed defines the safe speed range of high-speed trains. Excessive speed increases the risk of derailment, especially when the train is turning. For example, the derailment of a high-speed railway caused more than 80 deaths in Spain in 2013. When the train turns, excessive or insufficient speed increases the force between the train wheel flanges and the rails, which is a crucial cause of rail scratches. Moreover, rail scratches reduce the tolerance of the rail, which further increases the risk of derailment. Usually, when the train turns, there is an inclination angle in the rail to balance the centripetal force. Ideally, the centripetal force when the train turns is equal to the component of the train's gravity along the inclination. At this time, the force between the wheel flange and the rail is zero.

According to Newtonian mechanics, we have

$$m\left(\frac{v^2}{R} - g\tan\theta\right) = f_w, \tag{1}$$

where m denotes the mass of a carriage, g is the acceleration due to gravity, f_w denotes the combined force of all wheel flanges of the two carriages on their wheels, v denotes the speed of the train, R is the turning radius, and θ is the inclination angle of the rail. During train movement, some of the above parameters have inevitable errors. For example, m varies within a certain range depending on the passengers and their luggage.

2.2. Power Distribution. Decentralized power systems are often used in high-speed railways, compared with centralized power systems. High-speed trains usually consist of four or eight carriages. The power-decentralization problem of the two carriages can be considered first because the problem for sixteen carriages can be solved by the recursive 2-carriage problem. Among them, some parameters remain uncertain. For example, when the train is moving at a constant speed, the air resistance changes with the change in air pressure. In addition, the speed of the train cannot be held exactly constant, only within a very small range. The air resistance is proportional to the air pressure and the square of the speed. Hence, changing the air resistance will cause dynamic changes in the net power.

In the verification field based on the theorem proof, a reasoning method that fully considers the parameters with errors is necessary to verify the safety conditions and properties of the system.

3. Preliminary

In this section, we introduce some of the mathematical concepts that have been established and involved in our approach.

Definition 1 (Algebraic transition system). Let $A = <S, F, \Psi, \Lambda>$ be an algebraic transition system, where

S is the set of all states in the algebraic transition system

F is the set of transitions between states, $F \subseteq S \times S$

Ψ is the set of the algebraic assertions

$\Lambda$is the set of mapping relationships from F to Ψ and from S to Ψ. Each state or transition can be distributed into algebraic equations based on the mapping relationship

The algebraic transition system describes the system's transition relationship and the state itself of the system. The set F of an algebraic transition system describes the transition relations between states. Correspondingly, Λ provides each state satisfied algebraic equations or satisfied conditions of transitions between states.

In recent years, the algebraic transition system and its generalized structure have been well established in the field of verification [30, 31].

To establish a reasoning method with fault tolerance, we generalize linear algebraic assertions to quantitatively describe uncertain parameters. A quantitative description of these errors is necessary. Interval numbers have been widely used in the field of error estimation [32, 33]. As a result, we introduce interval numbers to describe the errors. The following is the definition and operation of interval numbers.

Definition 2 (Interval number). An interval number is a set of all real numbers in a closed interval.

Let $\overline{X}$ be an interval number. Then, $\overline{X} = [x^-, x^+]$, where $x^- \le x^+$; x^- is the lower bound of $\overline{X}$ and x^+ is the upper bound of $\overline{X}$. Thus, $\overline{X}$ can be any value in this closed interval. In particular, when $x^- = x^+$, the interval number becomes a normal real number.

Definition 3 (Interval number operation). The interval number operation includes the operations of addition, subtraction, multiplication, and division. Some operations are given as follows.

Let $\overline{a} = [a^-, a^+]$ and $\overline{b} = [b^-, b^+]$, where c is a constant.

Addition:

$$\begin{aligned} c + \overline{a} &= [a^- + c, a^+ + c], \\ \overline{a} + \overline{b} &= [a^- + b^-, a^+ + b^+]. \end{aligned} \tag{2}$$

Subtraction:

$$\begin{aligned} \overline{a} - \overline{b} &= [a^- - b^+, a^+ - b^-], \\ \overline{a} - c &= [a^- - c, a^+ - c]. \end{aligned} \tag{3}$$

Multiplication:

$$\overline{a}\overline{b} = [\min(a^-b^-, a^-b^+, a^+b^-, a^+b^+), \max(a^-b^-, a^-b^+, a^+b^-, a^+b^+)]. \tag{4}$$

Especially when $\overline{a} \ge 0$ and $\overline{b} \ge 0$, $ab = [a^-b^-, a^+b^+]$.

$$\begin{aligned} &\text{if } c > 0, c\overline{a} = [ca^-, ca^+], \\ &\text{if } c < 0, c\overline{a} = [ca^+, ca^-], \\ &\text{if } c = 0, c\overline{a} = [0, 0] = 0. \end{aligned} \tag{5}$$

Division:

$$\begin{aligned} &\text{if } c > 0, \frac{\overline{a}}{c} = \left[\frac{a^-}{c}, \frac{a^+}{c}\right], \\ &\text{if } c < 0, \frac{\overline{a}}{c} = \left[\frac{a^+}{c}, \frac{a^-}{c}\right]. \end{aligned} \tag{6}$$

In addition, there are some other definitions of interval number operations [34]. However, we do not elaborate here, as different definitions are irrelevant for the reasoning approach.

Unfortunately, although the errors can be described as any possible values over given intervals, the operation of interval number is not sufficient for reasoning between linear algebra assertions because it may lead to incorrect reasoning. For example, let $\overline{\varphi} = \begin{cases} \overline{X_2} - \overline{X_1} = [3, 4], \\ \overline{X_2} + \overline{X_1} = [5, 6]. \end{cases}$

According to the interval operation defined above, the linear equations can be solved as follows:

$$\begin{aligned} \overline{X_1} &= [0, 2], \\ \overline{X_2} &= [4, 5]. \end{aligned} \tag{7}$$

This is not a correct result. The correct result is in the blue diamond area in Figure 1.

Definition 4 (Polynomial). A polynomial is a mathematical expression consisting of a sum of terms, where each term includes one or more variables raised to a power and multiplied by a coefficient.

Let $V = \{x_1, \ldots, x_n\}$ be a set of variables. Let $\mathbb{R}[x_1, \ldots, x_n]$ be a set that comprising all polynomials with real coefficients on V. An example of a polynomial is as follows:

$$f_1(x_1, x_2, x_3) = 5x_1^3x_2x_3^2 + x_2^2 + 2x_3. \tag{8}$$

Definition 5 (Zero set of polynomials). $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$. The zero set of $f(x_1, \ldots, x_2)$ is the set as below, denoted by Zero(f):

$$\text{Zero}(f) = \{(x_1, \ldots, x_n) \in \mathbb{C}^n | f(x_1, \ldots, x_2) = 0\}.$$

Definition 6 (Linear algebraic assertions). A linear algebraic assertion consists of one or more linear equations. ψ is a linear algebraic assertion that contains the following equations:

$$\psi = \begin{cases} f_1(x_1, \ldots, x_n) = 0, \\ f_2(x_1, \ldots, x_n) = 0, \\ \ldots, \\ f_n(x_1, \ldots, x_n) = 0. \end{cases} \tag{9}$$

4. Implication and Equivalence Relations Based on Interval Numbers

In this section, we introduce the judgment rule of implication and equivalence relations based on interval numbers. Implication and equivalence relations are the most basic rules in any reasoning method. We first introduce the reasoning method (Definitions 7 and 8) involving implication and equivalence relations. Then, we introduce the definitions (Definitions 9–12).

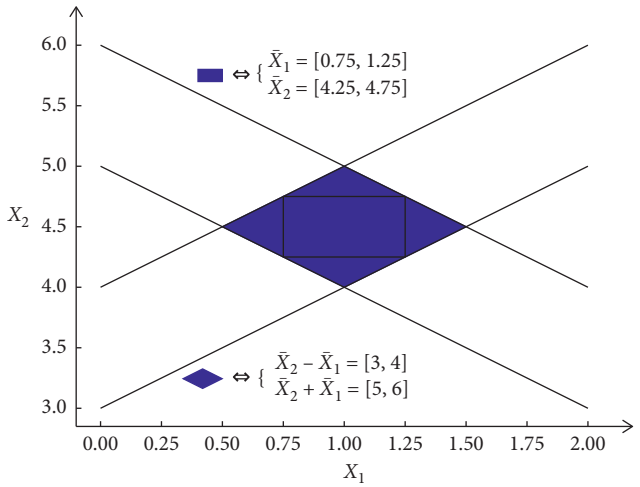


Figure 1: Inclusion relations between zero sets of two LAIs.

In the classic rules of reasoning, the implication relationship between algebraic assertions can be judged by the inclusion relationship of their zero set.

Definition 7 (Implication relations between algebraic assertions). Let φ_1 and φ_2 are two algebraic assertions. φ_1 implies φ_2, denoted as $\varphi_1| = \varphi_2$, iff $\text{Zero}(\varphi_1) \subseteq \text{Zero}(\varphi_2)$.

For example, $x - 1 = 0$ implies $x^2 + 2x - 3 = 0$ because $\{1\} \subseteq \{1, 2\}$.

Definition 8 (Equivalence relations between algebraic assertions). Let φ_1 and φ_2 be two algebraic assertions. φ_1 is equivalent to φ_2, denoted as $\varphi_1 \equiv \varphi_2$, iff $\text{Zero}(\varphi_1) = \text{Zero}(\varphi_2)$.

Definition 9 (LEI and LAI). An LEI is a linear algebraic equation whose variables and coefficients can be interval numbers. The LAI consists of one or more LEIs.

Definition 10 (Zero set of LEIs). Let $\overline{f}(x_1, \ldots, x_n) = \overline{a_0} + \overline{a_1}x_1 + \overline{a_2}x_2 + \cdots + \overline{a_n}x_n = 0$ be an LEI. $\overline{a_0}, \overline{a_1}, \overline{a_2}, \ldots, \overline{a_n}$ are given some interval numbers, as defined above. The zero set of $\overline{f}(x_1, \ldots, x_2)$ is the set as below and is denoted as $\text{Zero}(\overline{f})$:

$$\{(x_1, \ldots, x_n) | \forall a_0 \in \overline{a_0}, \forall a_1 \in \overline{a_1}, \forall a_2 \in \overline{a_2}, \ldots, \forall a_n \in \overline{a_n}, a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0\}. \tag{10}$$

Definition 11 (Implication relations between LAIs). Let $\overline{\varphi_1}$ and $\overline{\varphi_2}$ be two LAIs that have been defined above. $\overline{\varphi_1}$ implies $\overline{\varphi_2}$, denoted as $\overline{\varphi_1}| = \overline{\varphi_2}$, iff $\text{Zero}(\overline{\varphi_1}) \subseteq \text{Zero}(\overline{\varphi_2})$.

Definition 12 (Equivalence relations between LAIs). Let $\overline{\varphi_1}$ and $\overline{\varphi_2}$ be two LAIs. $\overline{\varphi_1}$ is equivalent to $\overline{\varphi_2}$, denoted as $\overline{\varphi_1} \equiv \overline{\varphi_2}$, iff $\text{Zero}(\overline{\varphi_1}) = \text{Zero}(\overline{\varphi_2})$.

The implication and equivalence relations are the two main reasoning rules. A simple example is shown in Figure 1 for reasoning between LAIs.

Let $\overline{\varphi_1} = \begin{cases} \overline{X_2} - \overline{X_1} = [3, 4], \\ \overline{X_2} + \overline{X_1} = [5, 6] \end{cases}$ and $\overline{\varphi_2} = \begin{cases} \overline{X_1} = [0.75, 1.25], \\ \overline{X_2} = [4.25, 4.75] \end{cases}$.

Obviously, $\text{Zero}(\overline{\varphi_2}) \subseteq \text{Zero}(\overline{\varphi_1})$, and we conclude that $\overline{\varphi_2}| = \overline{\varphi_1}$.

5. Reasoning Method between LAIs and Example

5.1. Reasoning Method between LAIs. In this section, we present a reasoning method to judge inclusion relations between zero sets of LAIs. Implication relations between LAIs can be judged by whether their zero set has an inclusion relation just like Definition 11 introduced above. The equivalence relations between LAIs can be judged by whether their zero sets have inclusion relations with each other. If the two sets are equal, the two sets contain each other. Before we introduce the reasoning method, we need to introduce the theorems (Lemma 1 and Theorem 1) and two basic mathematical definitions involving our reasoning method.

Definition 13 (Convex set). $D \subset \mathbb{R}^n$, and let arbitrary $x \in D$ and $y \in D$. D is the convex set, if for, $z = \lambda x + (1 - \lambda)y$ $\forall \lambda$ $0 \le \lambda \le 1$ then $z \in D$ is always true.

Definition 14 (Vertex set). V is a vertex set, if and only if for arbitrary, $z \in V$, and $z \notin \{z' | z' = \lambda x + (1 - \lambda)y, \forall x, \forall y \in V, \forall \lambda \in [0, 1]\}$.

Lemma 1. *The intersection of the zero set of LEI and the first quadrant is a convex set.*

Proof. Let $\overline{f}(x_1, \ldots, x_n)$ is LEI, and $\overline{f}(x_1, \ldots, x_n) = \overline{a_0} + \overline{a_1}x_1 + \overline{a_2}x_2 + \cdots + \overline{a_n}x_n = 0$, where $\overline{a_0} = [a_{0-}, a_{0+}]$, $\overline{a_1} = [a_{1-}, a_{1+}], \ldots, \overline{a_n} = [a_{n-}, a_{n+}]$.

Let, if arbitrary $\alpha \in \text{Zero}(\overline{f}), \beta \in \text{Zero}(\overline{f}), \alpha \ne \beta$,

$$\begin{aligned} \alpha &= \left[x_{\alpha 1}, x_{\alpha 2}, \ldots, x_{\alpha n}\right]^T, \\ \beta &= \left[x_{\beta 1}, x_{\beta 2}, \ldots, x_{\beta n}\right]^T, \end{aligned} \tag{11}$$

then

$$\begin{aligned} \exists a_\alpha, \exists a_\beta,\ a_\alpha = \left[a_{\alpha 1}, a_{\alpha 2}, \ldots., a_{\alpha n}\right]^T, a_\beta = \left[a_{\beta 1}, a_{\beta 2}, \ldots, a_{\beta n}\right]^T, \\ a_{\alpha 1} \in \overline{a_1}, a_{\alpha 2} \in \overline{a_2}, \ldots, a_{\alpha n} \in \overline{a_n}, \\ a_{\beta 1} \in \overline{a_1}, a_{\beta 2} \in \overline{a_2}, \ldots, a_{\beta n} \in \overline{a_n}. \end{aligned} \tag{12}$$

We have

$$\begin{aligned} a_{\alpha 0} + a_{\alpha 1}x_{\alpha 1} + \cdots + a_{\alpha n}x_{\alpha n} = 0, \\ a_{\beta 0} + a_{\beta 1}x_{\beta 1} + \cdots + a_{\beta n}x_{\beta n} = 0. \end{aligned} \tag{13}$$

For arbitrary $\lambda, 0 \le \lambda \le 1$, let $\lambda, 1-\lambda$. By multiplying the above two equations, we have

$$\begin{aligned} \lambda a_{\alpha 0} + \lambda a_{\alpha 1}x_{\alpha 1} + \cdots + \lambda a_{\alpha n}x_{\alpha n} = 0, \\ (1-\lambda)a_{\beta 0} + (1-\lambda)a_{\beta 1}x_{\beta 1} + \cdots + (1-\lambda)a_{\beta n}x_{\beta n} = 0. \end{aligned} \tag{14}$$

By adding the above two equations, we have

$$\begin{aligned} \left(\lambda a_{\alpha 0} + (1-\lambda)a_{\beta 0}\right) + \left(\lambda a_{\alpha 1}x_{\alpha 1} + (1-\lambda)a_{\beta 1}x_{\beta 1}\right) + \cdots \\ + \left(\lambda a_{\alpha n}x_{\alpha n} + (1-\lambda)a_{\beta n}x_{\beta n}\right) = 0. \end{aligned} \tag{15}$$

Let $(\lambda a_{\alpha 0} + (1-\lambda)a_{\beta 0}) = a_{z0}$.

Apparently, $a_{z0} = \lambda a_{\alpha 0} + (1-\lambda)a_{\beta 0} \in [\min(a_{\alpha 0}, a_{\beta 0}), \max(a_{\alpha 0}, a_{\beta 0})] \subseteq [a_{0-}, a_{0+}]$.

Let $z = \lambda\alpha + (1-\lambda)\beta$, following the definition of $\mathrm{Zero}(\overline{f})$.

$z \in \mathrm{Zero}(\overline{f})$, if and only if

$$\begin{aligned} \exists a_{z1} \in \overline{a_1}, \ldots, \exists a_{zn} \in \overline{a_n}, a_{z0} + a_{z1}\left(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1}\right) + \cdots \\ + a_{zn}\left(\lambda x_{\alpha n} + (1-\lambda)x_{\beta n}\right) = 0. \end{aligned} \tag{16}$$

Let us prove (16).

Take $a_{z1} = ((\lambda a_{\alpha 1}x_{\alpha 1} + (1-\lambda)a_{\beta 1}x_{\beta 1})/(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1})), \ldots, a_{zn} = ((\lambda a_{\alpha n}x_{\alpha n} + (1-\lambda)a_{\beta n}x_{\beta n})/(\lambda x_{\alpha n} + (1-\lambda)x_{\beta n}))$ according to (15). Apparently, we have

$$a_{z0} + a_{z1}\left(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1}\right) + \cdots + a_{zn}\left(\lambda x_{\alpha n} + (1-\lambda)x_{\beta n}\right) = 0. \tag{17}$$

Furthermore, assuming $a_{\alpha 1} \le a_{\beta 1}$ (the case of $a_{\alpha 1} > a_{\beta 1}$ is similar), we easily obtain

$$\begin{aligned} a_{\alpha 1}\left(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1}\right) \le \left(\lambda a_{\alpha 1}x_{\alpha 1} + (1-\lambda)a_{\beta 1}x_{\beta 1}\right) \\ \le a_{\beta 1}\left(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1}\right). \end{aligned} \tag{18}$$

Because both α and β belong to the first quadrant, we have

$$\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1} > 0. \tag{19}$$

By dividing (18) by $(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1})$, we obtain

$$a_{\alpha 1} \le \frac{\left(\lambda a_{\alpha 1}x_{\alpha 1} + (1-\lambda)a_{\beta 1}x_{\beta 1}\right)}{\left(\lambda x_{\alpha 1} + (1-\lambda)x_{\beta 1}\right)} \le a_{\beta 1}, \tag{20}$$

that is,

$$a_{\alpha 1} \le a_{z1} \le a_{\beta 1}. \tag{21}$$

So, $a_{z1} \in [a_{\alpha 1}, a_{\beta 1}] \subset [a_{1-}, a_{1+}] = \overline{a_1}$.

Similarly, $a_{zn} \in [a_{\alpha n}, a_{\beta n}] \subset [a_{n-}, a_{n+}] = \overline{a_n}$.

Theorem 1. *the intersection of the zero set of LAI and the first quadrant is a convex set.*

Proof. As the definition of LEI, an LAI consists of one LEI or much more LEI.

Let $\overline{\varphi}$ be an LAI,

$$\overline{\varphi} = \begin{cases} \overline{f_1}(x_1, \ldots, x_n), \\ \ldots, \\ \overline{f_n}(x_1, \ldots, x_n), \end{cases} \tag{22}$$

Let A_1 denote the first quadrant area.

Apparently, $\mathrm{Zero}(\overline{\varphi}) = \mathrm{Zero}(\overline{f_1}) \cap \cdots \cap \mathrm{Zero}(\overline{f_n})$, and $\mathrm{Zero}(\overline{f_1}) \cap A_1, \ldots, \mathrm{Zero}(\overline{f_n}) \cap A_1$ are convex sets, following the conclusion of Lemma 1 above.

According to one of the properties of a convex set [35], we find that

$(\mathrm{Zero}(\overline{f_1}) \cap A_1) \cap (\mathrm{Zero}(\overline{f_2}) \cap A_1) \cap \cdots \cap (\mathrm{Zero}(\overline{f_n}) \cap A_1)$ is also a convex set.

That is, $A_1 \cap \mathrm{Zero}(\overline{f_1}) \cap \mathrm{Zero}(\overline{f_2}) \cap \cdots \cap \mathrm{Zero}(\overline{f_n})$ is also a convex set.

For most engineering problems, only solutions in the first quadrant are meaningful. Although it may be meaningful that the solutions are negative to some problems, it is still possible to make the solution meaningful in the first quadrant through proper coordinate transformation.

For example, if c represents the temperature value in degrees Celsius, and represents c' represents the temperature in Kelvin. The coordinate transformation is $c' = c + 273.15$. Apparently, c' is meaningful only when it is positive. In this article, the zero set of LAIs to which we refer is its intersection with the first quadrant.

Because, if A and B are both convex sets and $p_1, p_2, \ldots, p_n$ are all vertexes of A, then $A \subseteq B$ iff $\forall p_i \in B$, $i = 1, 2, \ldots, n$. Therefore, we can judge whether all of its vertexes are contained by another zero set of LAI to determine whether there is an inclusion relationship between the two sets. We thus obtain the following reasoning method.

Let $\overline{\varphi_1}$ and $\overline{\varphi_2}$ be two linear algebraic assertions. Is $\overline{\varphi_1}| = \overline{\varphi_2}$ correct? A method for judging inclusion relations between the LAI can be given as follows:

Step 1. Calculate all vertexes of $\overline{\varphi_1}$.

Step 2. Determine the inequality equations, which are equivalent to the zero set of $\overline{\varphi_2}$.

Step 3. If all vertexes of $\overline{\varphi_1}$ satisfy the inequalities obtained in Step 2, we have that $\overline{\varphi_1}| = \overline{\varphi_2}$ is true; otherwise, it is not true. □

5.2. Reasoning Method Example with LAIs. In this section, we present a case that can be solved by linear algebraic reasoning rules with errors.

Let

$$\overline{\varphi_1} = \begin{cases} [0.9,1]\overline{X} + [0.8,1]\overline{Y} + [0.7,1]\overline{Z} = [0.8,1.2], \\ [0.9,1]\overline{X} + [0.7,1.1]\overline{Y} - [0.9,1]\overline{Z} = [0.3,0.4], \\ [0.9,1.1]\overline{X} - [0.7,1.1]\overline{Y} + [0.9,1]\overline{Z} = [0.5,0.6], \end{cases}$$
$$\overline{\varphi_2} = \begin{cases} \overline{X} = [0.35,0.55], \\ \overline{Y} = [0.25,0.4], \\ \overline{Z} = [0.2,0.3], \end{cases} \tag{23}$$

We want to know whether $\overline{\varphi_2}| = \overline{\varphi_1}$.

Proof. First, we can easily obtain the boundary equations as follows to find all vertexes of $\overline{\varphi_1}$:

$$\begin{cases} X + Y + Z = 0.8, \\ 0.9X + 0.8Y + 0.7Z = 1.2, \end{cases} \quad \begin{cases} 0.9X + 0.7Y - Z = 0.4, \\ X + 1.1Y - 0.9Z = 0.3, \end{cases} \quad \begin{cases} 1.1X - 0.7Y + Z = 0.5, \\ 0.9X - 1.1Y + 0.9Z = 0.6. \end{cases} \tag{24}$$

We obtain three equation systems, each of which contains two equations. An equation is selected arbitrarily from each equation system, so eight groups of equations can be formed. By solving these eight groups of equations, we obtain eight vertexes.

Calculate all vertexes of $\overline{\varphi_1}$: $p_1 = (9/20, 69/340, 5/34)$, $p_2 = (549/950, 3/50, 77/475)$, $p_3 = (9/19, 3/50, 253/950)$, $p_4 = (81/245, 48/245, 67/245)$, $p_5 = (284/1253, 7971253, 872/1253)$, $p_7 = (9/20, 1583/2580, 1121/2580)$, and $p_8 = (1934/2925, 118/325, 146/325)$.

The above eight vertexes of $\overline{\varphi_1}$ constitute a hexahedron in 3D space. To mathematically represent the interior region of the hexahedron (including boundary), we need to obtain a group of linear inequalities defining this region. According to the theory of cylindrical algebraic decomposition, all points in a linear closed region isolated by finite points satisfy the same inequalities. To find a point in the interior region of a hexahedron, we need to solve a group of equations, in which the coefficients can take any value within their interval according to Theorem 1, as proven above. If there is no solution or if the solution is not unique for arbitrarily selected coefficients, we can reselect a group of coefficients until the system of equations has a solution.

Find an interior point $P(X_P, Y_P, Z_P)$,

$$\begin{cases} 0.9X + 0.9Y + 0.9Z = 1, \\ X + Y - Z = 0.35, \\ X - Y + Z = 0.5, \end{cases} \Longleftrightarrow P(X_P, Y_P, Z_P) = \left(\frac{17}{40}, \frac{11}{36}, \frac{137}{360}\right). \tag{25}$$

$P(X_P, Y_P, Z_P)$ must be in the interior region of the hexahedron (including the boundary), as shown in Figures 2 and 3.

Substituting point $P(X_P, Y_P, Z_P)$ into formula (24), we obtain three groups of linear inequalities as follows:

$$\begin{cases} X_p + Y_p + Z_p \geq 0.8, \\ 0.9X_p + 0.8Y_p + 0.7Z_p \leq 1.2, \end{cases} \quad \begin{cases} 0.9X_p + 0.7Y_p - Z_p \leq 0.4, \\ X_p + 1.1Y_p - 0.9Z_p \geq 0.3, \end{cases} \quad \begin{cases} 1.1X_p - 0.7Y_p + Z_p \geq 0.5, \\ 0.9X_p - 1.1Y_p + 0.9Z_p \leq 0.6. \end{cases} \tag{26}$$

Calculate all vertexes of $\overline{\varphi_2}$:

$$\begin{aligned} q_1 &= (0.4, 0.3, 0.3), \\ q_2 &= (0.4, 0.3, 0.4), \\ q_3 &= (0.4, 0.35, 0.3), \\ q_4 &= (0.4, 0.35, 0.4), \\ q_5 &= (0.5, 0.3, 0.3), \\ q_6 &= (0.5, 0.3, 0.4), \\ q_7 &= (0.5, 0.35, 0.3), \\ q_8 &= (0.5, 0.35, 0.4). \end{aligned} \tag{27}$$

According to the properties of convex sets, if all vertexes of $\overline{\varphi_2}$ are in the interior region (including the boundary) of the zero set of $\overline{\varphi_1}$, then $\text{Zero}(\overline{\varphi_2}) \subseteq \text{Zero}(\overline{\varphi_1})$.

Take all eight vertexes of $\overline{\varphi_2}$ into formula (26) to verify all inequalities in formula (26). We find that only q_3 is not satisfied with one inequality of formula (26). For q_3, $1.1X - 0.7Y + Z \geq 0.5$ does not hold, as $1.1X_{q_3} - 0.7Y_{q_3} + Z_{q_3} = 0.495$.

So, $\text{Zero}(\overline{\varphi_2}) \not\subset \text{Zero}(\overline{\varphi_1})$.

In other words, $\overline{\varphi_2}| = \overline{\varphi_1}$ does not hold.

If only looking at Figures 4 and 5, it seems that all vertexes of $\overline{\varphi_2}$ are interior regions of the zero set of $\overline{\varphi_1}$. However, after the above calculations, q_3 is almost inside. So it dose not hold. □

6. Verification of Decentralized Power Systems during Turn

In this section, we present a case that can be solved by the reasoning method mentioned above in this article. The problem in four or eight carriages can be solved by a recursive 2-carriage problem. Hence, we mainly discuss the power decentralization of 2 carriages.

A simplified algebraic transition system for the train is shown in Figure 6. $\overline{g_1}$ and $\overline{g_2}$ represent the conditions satisfied by corresponding transitions between the states. $\overline{\varphi_1}$, $\overline{\varphi_2}$ and $\overline{\varphi_3}$ are the equations that need to be satisfied in the corresponding states. That is, when the train is in the acceleration state and if $\overline{\varphi_1}$ is not satisfied, there is a strong

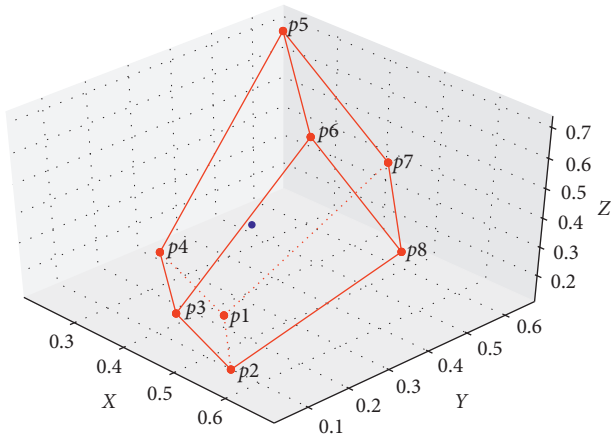


Figure 2: Blue point in the interior.

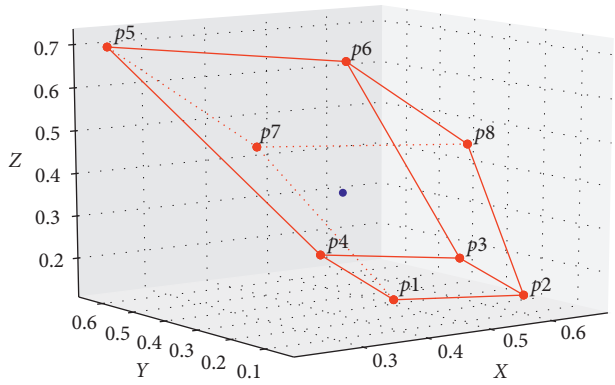


Figure 3: Another angle of Figure 2.

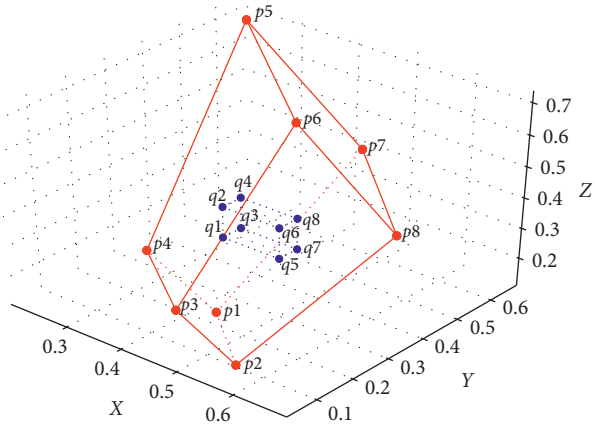


Figure 4: All vertexes in the area of $\overline{\varphi_1}$ and $\overline{\varphi_2}$.

possibility that the train is in an abnormal acceleration process, which requires timely troubleshooting.

The following case is when a constant-speed train is turning. Decentralized power systems are widely used in high-speed trains. The power source is scattered among the

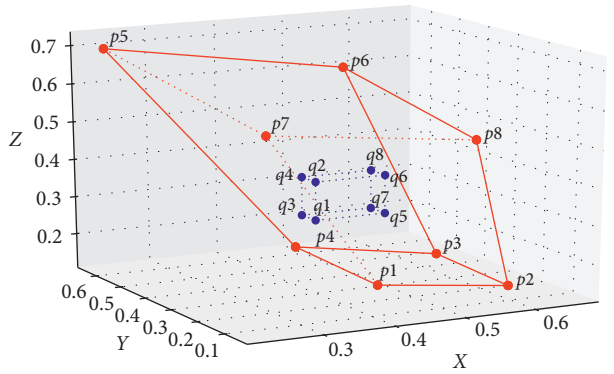


Figure 5: Another angle of Figure 4.

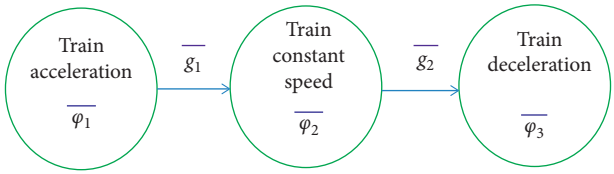


Figure 6: Algebraic transition system of a train.

engines of the carriages. The net traction power of each compartment in a train will vary randomly within a small range, caused by the movement of passengers, their luggage, and wind resistance. The role of the wheel flange of the train is to prevent derailment, especially when the train is turning. The force analysis during turning is shown in Figure 6.

The two self-powered carriages at constant speed are shown in Figure 7. Carriage 1 and carriage 2 may be two connected carriages or two groups of carriages. We have the following description: f_1 and f_2 denote the traction force of carriage 1 and carriage 2, respectively; m denotes the initial mass of each carriage; and Δm denotes the quality change in each carriage due to the variation of passengers and their luggage and is similar to the effect of mass change. When a train passes the inclined plane in turning, a part of the gravity caused by a certain inclination provides the centripetal force for the train to turn, which is also similar to the change in the mass of the train. μ stands for the friction coefficient; f_{12} denotes the force of carriage 1 on carriage 2; g is the acceleration due to gravity; and f_w denotes the combined force of all wheel flanges of the two carriages on their wheels. Wheel flange is a special device to reduce risks when turning. It is shown in Figure 8. When the train turns quickly, f_w may exceed the force limit of the wheel flange, which may cause the train to derail. Moreover, a larger f_w will increase the friction between the wheel flanges and the rail and cause injures on the wheel flange of both rail flats. Injured rails and wheels further increase the possibility of derailment when turning. Therefore, f_w should be within a certain range. ζ is related to the air density and the shape of

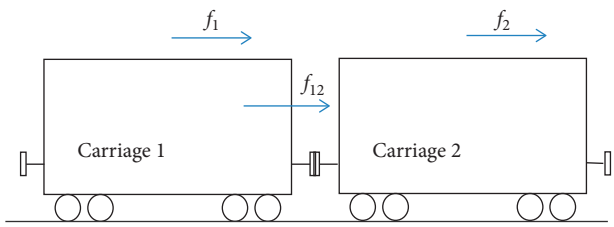


Figure 7: Two self-powered carriages turning.

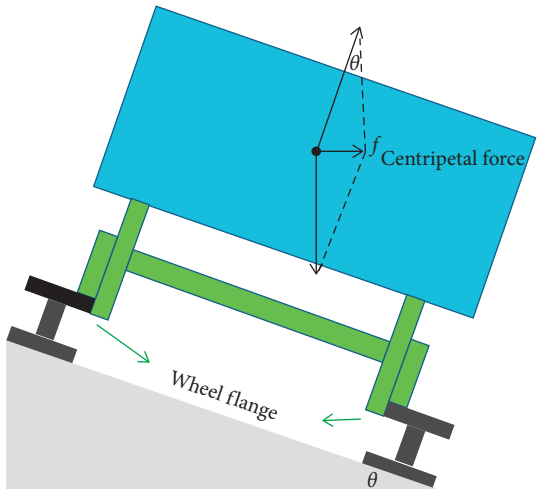


Figure 8: Train turning.

the train. Among them, the air density may change due to different altitudes. Therefore, ζ is also defined within a certain range.

It is assumed that the above parameters require the following values according to the design requirements of the train: $m = 50000\,\text{kg}$, $\Delta m = [0, 0.2\,\text{m}]$, $\mu = 0.01$, $f_{12} = [0, 2000]\,\text{N}$, $\theta = 10^\circ$, $g = 10\,\text{m}\cdot\text{s}^{-2}$, $\zeta = [1.5, 1.6]\,\text{kg}\cdot\text{m}^{-1}$, $R = 3000\,\text{m}$, and $f_w = [-10000, 10000]\,\text{N}$.

According to mechanics, we have

$$\overline{\varphi} = \begin{cases} f_1 + f_2 = 2(m + \Delta m)g\mu + \zeta v^2, \\ -f_{12} + f_1 = (m + \Delta m)g\mu + \zeta v^2, \\ (m + \Delta m)\left(\dfrac{v^2}{R} - g\tan\theta\right) = f_w. \end{cases} \tag{28}$$

From the third equation in (28), we obtain

$$\left(\frac{v^2}{R} - g\tan\theta\right) = \frac{f_w}{m[1, 1.2]} = [-0.2, 0.2], \tag{29}$$

$$v^2 = R[g\tan\theta - 0.2, g\tan\theta + 0.2].$$

Let $R(g\tan\theta - 0.2) = v_-^2$ and $R(g\tan\theta + 0.2) = v_+^2$. Then, $v^2 = [v_-^2, v_+^2]$. That is,

$$\overline{\varphi} = \begin{cases} f_1 + f_2 - \zeta v^2 = 2m[1, 1.2]g\mu, \\ f_1 - \zeta v^2 = m[1, 1.2]g\mu + f_{12}, \\ v^2 = \left[v_-^2, v_+^2\right]. \end{cases} \tag{30}$$

The boundary equation of each equation in $\overline{\varphi}$ can be obtained as follows, which can solve all vertexes of $\overline{\varphi}$:

$$\begin{gathered} \begin{cases} f_1 + f_2 - 1.5v^2 = 10000, \\ f_1 + f_2 - 1.6v^2 = 12000, \end{cases} \\ \begin{cases} f_1 - 1.5v^2 = 5000, \\ f_1 - 1.6v^2 = 8000, \end{cases} \\ \begin{cases} v^2 = v_-^2 = 3000\left(10\tan 10^\circ - 0.2\right), \\ v^2 = v_+^2 = 3000\left(10\tan 10^\circ + 0.2\right). \end{cases} \end{gathered} \tag{31}$$

We obtain three equation groups, each of which contains two equations. Select an equation arbitrarily from each equation system, so eight equation systems can be formed. Solving these eight equations, we can obtain eight vertexes.

Calculate all vertexes of $\overline{\varphi_1}$ (f_1, f_2, v^2):

$$\begin{aligned} p_1 &= \left(5000 + 1.5v_-^2, 5000, v_-^2\right) \approx (12034.7, 5000, 4689.8), \\ p_2 &= \left(5000 + 1.5v_+^2, 5000, v_+^2\right) \approx (13834.7, 5000, 5889.8), \\ p_3 &= \left(8000 + 1.6v_-^2, 2000 - 0.1v_-^2, v_-^2\right) \approx (15503.69, 1531, 4689.8), \\ p_4 &= \left(8000 + 1.6v_+^2, 2000 - 0.1v_+^2, v_+^2\right) \approx (17423.69, 1411, 5889.8), \\ p_5 &= \left(5000 + 1.5v_-^2, 7000 + 0.1v_-^2, v_-^2\right) \approx (12034.7, 7468.98, 4689.8), \\ p_6 &= \left(5000 + 1.5v_+^2, 7000 + 0.1v_+^2, v_+^2\right) \approx (13834.7, 7588.98, 5889.8), \\ p_7 &= \left(8000 + 1.6v_-^2, 4000, v_-^2\right) \approx (15503.69, 4000, 4689.8), \\ p_8 &= \left(8000 + 1.6v_+^2, 4000, v_+^2\right) \approx (17423.69, 4000, 5889.8). \end{aligned} \tag{32}$$

As before, the above eight vertexes of $\overline{\varphi}$ constitute a hexahedron in 3D space. To mathematically represent the interior region of the hexahedron (including boundary), we need to obtain a group of linear inequalities of this region. According to the theory of cylindrical algebraic decomposition, all points in a linear closed region isolated by finite points satisfy the same inequalities. To find a point in the interior region of a hexahedron, we just need to solve a group of equations, in which coefficients can take any value of their interval according to Theorem 1.

An interior point $P(f_1', f_2', v^{2'})$ can be solved by equations with certain coefficients in the allowable error range. Without losing generality, $P(f_1', f_2', v^{2'})$ can be solved by the following equations:

$$\begin{cases} f_1 + f_2 - 1.55v^2 = 1.1 \cdot 2mg\mu, \\ f_1 - 1.55v^2 = 1.1 \cdot mg\mu + 1000, \\ 1.1m\left(\dfrac{v^2}{R} - g\tan\theta\right) = 0. \end{cases} \Longleftrightarrow P\left(f_1', f_2', v^{2'}\right) \approx (14699.20, 4500, 5289.81), \tag{33}$$

By substituting point $P(f_1', f_2', v^{2'})$ into formula (31), we obtain the following three groups of linear inequalities:

$$\begin{cases} f_1 + f_2 - 1.5v^2 \geq 10000, \\ f_1 + f_2 - 1.6v^2 \leq 12000, \\ f_1 - 1.5v^2 \geq 5000, \\ f_1 - 1.6v^2 \leq 8000, \\ v^2 \geq v_-^2 = 3000\left(10\tan 10^\circ - 0.2\right), \\ v^2 \leq v_+^2 = 3000\left(10\tan 10^\circ + 0.2\right). \end{cases} \tag{34}$$

The inequality groups in formula (34) represent the fault-tolerance area of the system, which is the area where the system allows controllable errors. The fault-tolerance area is shown in Figures 9 and 10.

We can verify whether f_1, f_2, and v^2 satisfy the inequality group in formula (34) to judge whether the decentralized power systems and train speed are working properly. When f_1, f_2, and v^2 do not satisfy formula (34), the decentralized power system or train speed is probably working incorrectly and requires timely error detection.

In the fault-tolerance area, the interval number of v^2 can be transformed into an interval number of v. The equivalent fault-tolerance area will not be described again:

$$\left(v^2 \approx [4689.8, 5889.8]\right) \Leftrightarrow \left(v \approx [68.48, 76.74]\ \mathrm{m}\cdot\mathrm{s}^{-1}\right) \Longleftrightarrow \left(v \approx [246.52, 276.26]\ \mathrm{km}\cdot\mathrm{h}^{-1}\right). \tag{35}$$

7. Simulation and Comparison

7.1. Simulation. In this section, we test the fault-tolerance area in Section 6. Δm denotes the quality change; ζ is related to the air density and the shape of the train; f_w denotes the combined force of all wheel flanges of the two carriages on their wheels; and f_{12} denotes the force of carriage 1 on carriage 2. There are four parameters ($\Delta m \zeta f_w f_{12}$) with errors in (28). The meanings of these four uncertain parameters are the same as those in Section 6. Therefore, it will not be described in detail here. The four interval numbers are as follows:

$$\begin{aligned} \Delta m &= [0, 10000], \\ \zeta &= [1.5, 1.6], \\ f_w &= [-10000, 10000], \\ f_{12} &= [0, 2000]. \end{aligned} \tag{36}$$

An arbitrary test case refers to randomly assigned values for the four parameters. The N test cases can be described by the following formula:

$$\mathrm{case}_1 = [\Delta m_1, \zeta_1, (f_w)_1, (f_{12})_1], \ldots, \mathrm{case}_n = [\Delta m_n, \zeta_n, (f_w)_n, (f_{12})_n]. \tag{37}$$

By substituting formula (37) into formula (28), we can obtain the solutions of the corresponding test cases as the following formula:

$$\mathrm{solution}_1 = \left[(f_1)_1, (f_2)_1, v_1^2\right] \ldots, \mathrm{solution}_n = \left[(f_1)_n, (f_2)_n, v_n^2\right]. \tag{38}$$

By substituting formula (38) into formula (28), we can verify the correctness of the fault-tolerance area. After testing, the solutions of all test cases are inside the fault-tolerance area, including its boundary. Figures 11(a) and 11(b) show that the solutions of these 1000 test cases are all inside the fault-tolerance area. Figures 11(c) and 11(d) show that the solutions of these 10,000 and 100,000 test cases.

When the number of test cases is 10,000 and 100,000, the same conclusion is obtained, as shown in Figures 11(c) and 11(d), respectively.

The sensitivities of these four uncertain parameters are different. The change of mass and air coefficient is the most sensitive to safety conditions. For example, China's Fuxing high-speed railway has strict limits on the number of passengers.

7.2. Comparison. Previous reasoning methods based on algebraic polynomials have mainly concentrated on non-error systems [9, 12–15], whose coefficients and variables are accurately described. For systems with errors in coefficients

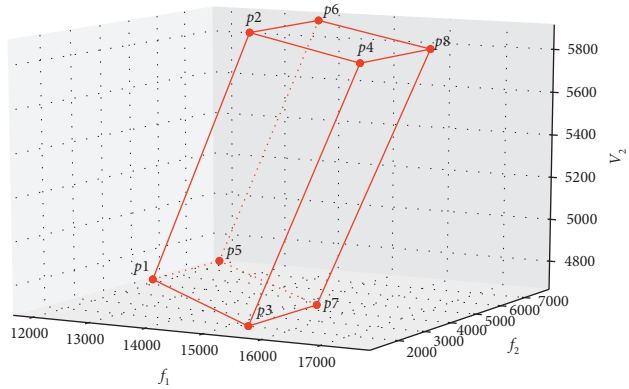


Figure 9: The fault-tolerance area of the system.

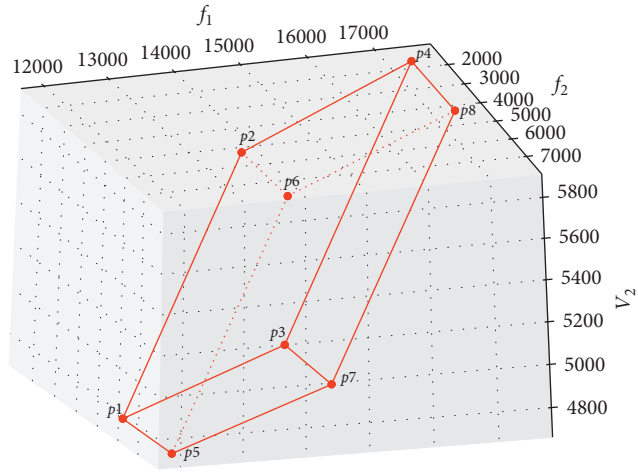


Figure 10: Another angle of Figure 9.

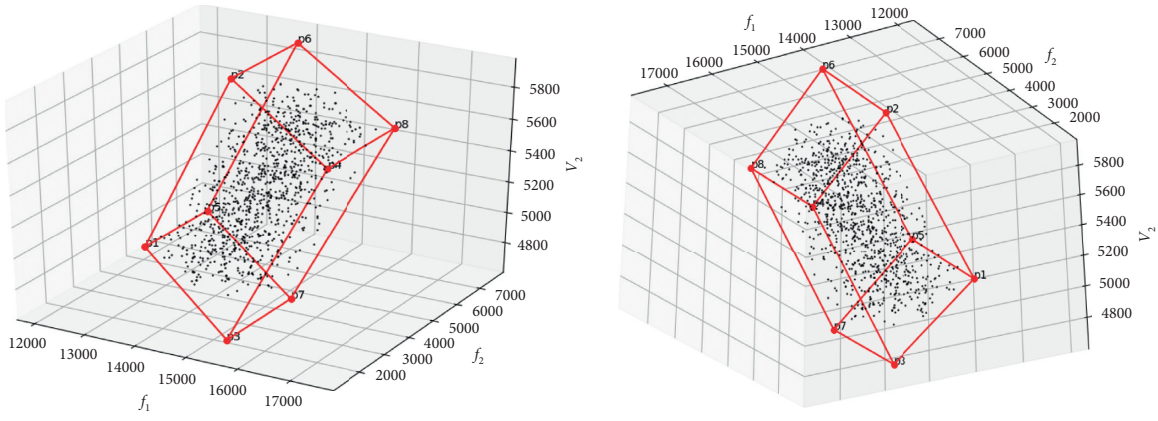
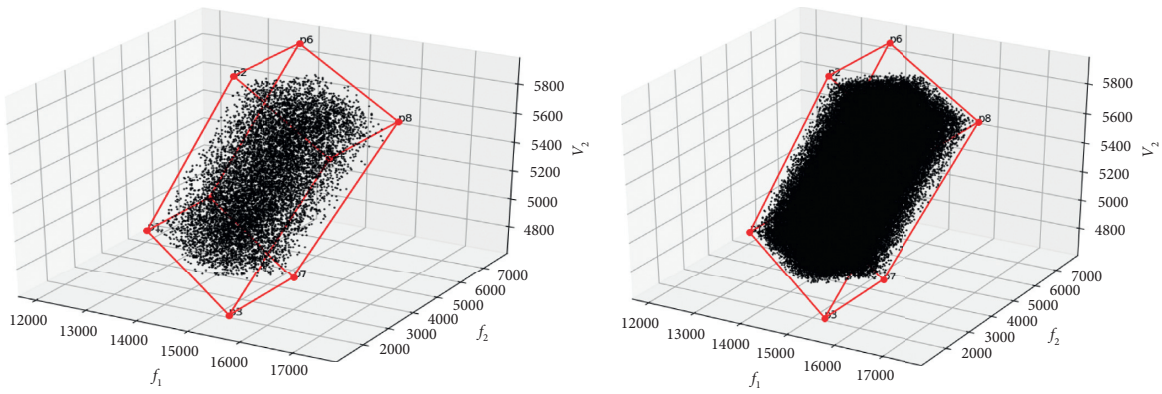


Figure 11: Continued.



(c) (d)

Figure 11: (a) 1000 test cases; (b) another angle of (a); (c) 10,000 test cases; (d) 100,000 test cases.

and variables, most previous methods are incompetent. However, the method of Reference [16] is very valuable in theory but is only effective for a single variable or coefficient with error not for multiple error variables or coefficients. Among the fault-tolerant methods, there are many similar fault-tolerant error analysis methods [23–25], but formal reasoning methods are rarely reported.

8. Conclusion

Our main contribution is to show that the reasoning method is reliable and the error controllable, even though errors exist in the coefficients and variables in the linear assertion. Furthermore, the method proposed in this paper is not limited only to the verification of decentralized power systems, as errors in many systems are common and unavoidable. This method is promising in systems described by linear equations with error parameters. In such systems, our method may remain effective by using linear equations to approximate the nonlinearity within a small time interval. Hence, the method in this study has a wide range of applications.

Nevertheless, if not to approximate the nonlinearity of the system by linear equations within a small time interval, our reasoning method may not be applicable to these nonlinear systems with errors. This is also the focus of our work in the future.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant no. 61772006, the Science and Technology Major Project of Guangxi under grant no. AA17204096, the Key Research and Development Project of Guangxi under grant no. AB17129012, and the Special Fund for Bagui Scholars of Guangxi.

References

[1] L. Lamport, "The temporal logic of actions," *ACM Transactions on Programming Languages and Systems*, vol. 16, no. 3, pp. 872–923, 1994.

[2] M. Fitting, "First-order logic and automated theorem proving," *Studia Logica*, vol. 61, no. 2, pp. 300–302, 1998.

[3] E. M. Clarke, O. Grumberg, and D. Peled, *Model Checking*, MIT Press, Cambridge, MA, USA, 1999.

[4] C. Baier and J.-P. Katoen, *Principles of Model Checking*, MIT Press, Cambridge, MA, USA, 2008.

[5] N. Shankar, "Combining theorem proving and model checking through symbolic analysis,"*in Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2000.

[6] T. E. Uribe, "Combinations of model checking and theorem proving,"*in Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2000.

[7] J.-P. Katoen, "Labelled transition systems,"*in Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2005.

[8] V. D'Silva, D. Kroening, and G. Weissenbacher, "A survey of automated techniques for formal software verification," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 7, pp. 1165–1178, 2008.

[9] J. Fu, J. Wu, and H. Tan, "A deductive approach towards reasoning about algebraic transition systems," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–12, Article ID 607013, 2015.

[10] S. Sankaranarayanan, H. B. Sipma, and Z. Manna, "Nonlinear loop invariant generation using Gröbner bases," *ACM SIGPLAN Notices*, vol. 39, no. 1, pp. 318–329, 2004.

[11] L. Doyen, G. Frehse, and G. J. Pappas, *Handbook of Model Checking*, Springer International Publishing, Manhattan, NY, USA, 2018.

[12] A. Platzer, *Logical Analysis of Hybrid Systems: Proving Theorems for Complex Dynamics*, Springer, Berlin, Germany, 2010.

[13] A. Platzer, "A differential operator approach to equational differential invariants," in *Lecture Notes in Computer Science Beringer*, A. P. Felty, Ed., Springer, Berlin, Germany, 2012.

[14] A. Platzer, *Logics of Dynamical Systems*, IEEE, Piscataway, NJ, USA, 2012.
[15] A. Platzer, "The structure of differential invariants and differential cut elimination," *Logical Methods in Computer Science*, vol. 8, no. 4, pp. 1–38, 2012.
[16] J. Liu, N. Zhan, and H. Zhao, "Computing semi-algebraic invariants for polynomial dynamical systems," in *EMSOFT Chakraborty*, A. Jerraya, S. K. Baruah, and S. Fischmeister, Eds., pp. 97–106, ACM, New York, NY, USA, 2011.
[17] S.-K. Li, L.-X. Yang, and K.-P. Li, "Robust output feedback cruise control for high-speed train movement with uncertain parameters," *Chinese Physics B*, vol. 24, no. 1, Article ID 010503, 2015.
[18] Y.-D. Song, S. Qi, and W.-C. Cai, "Fault-tolerant adaptive control of high-speed trains under traction/braking failures: a virtual parameter-based approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 737–748, 2014.
[19] Q. Song and Y. D. Song, "Data-based fault-tolerant control of high-speed trains with traction/braking notch nonlinearities and actuator failures," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2250–2261, 2011.
[20] X. Su, X. Liu, and Y.-D. Song, "Fault-tolerant control of multiarea power systems via a sliding-mode observer technique," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 38–47, 2018.
[21] Y. Zhang, C. Qin, W. Zhang, F. Liu, and X. Luo, "On the fault-tolerant performance for a class of robust image steganography," *Signal Processing*, vol. 146, pp. 99–111, 2018.
[22] Y. Chen and B. Guo, "Sliding mode fault tolerant tracking control for a single-link flexible joint manipulator system," *IEEE Access*, vol. 7, pp. 83046–83057, 2019.
[23] W. Liu and P. Li, "Disturbance observer-based fault-tolerant adaptive control for nonlinearly parameterized systems," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8681–8691, 2019.
[24] J. Wang, K. Liang, X. Huang, Z. Wang, and H. Shen, "Dissipative fault-tolerant control for nonlinear singular perturbed systems with markov jumping parameters based on slow state feedback," *Applied Mathematics and Computation*, vol. 328, pp. 247–262, 2018.
[25] Z. Qu, H. Wang, X. Peng, and Q. Wang, "Lineage chain mark fault-tolerant method for micro-batching monitoring data in distribution power network," *IEEE Access*, vol. 7, pp. 32949–32960, 2019.
[26] A. K. Louis, "Approximate inverse for linear and some nonlinear problems," *Inverse Problems*, vol. 12, no. 2, pp. 175–190, 1996.
[27] W. Kahan, "Numerical linear algebra," *Canadian Mathematical Bulletin*, vol. 9, no. 5, pp. 757–801, 1966.
[28] N. Markovic, T. Stoetzel, V. Staudt, and D. Kolossa, *Hybrid Fault Detection in Power Systems*, IEEE International Electric Machines & Drives Conference (IEMDC), San Diego, CA, USA, 2019.
[29] K. Moloi, Y. Hamam, and J. A. Jordaan, "Fault detection in power system integrated network with distribution generators using machine learning algorithms," in *Proceedings of the 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pp. 18–22, Johannesburg, South Africa, November 2019.
[30] H. Deng and J. Wu, "On approximate bisimulation equivalence for linear semi-algebraic transition systems," *Journal of Jilin University*, vol. 43, no. 4, pp. 1052–1058, 2013.
[31] J. Fu, J. Wu, H. Tan, and N. Zhou, "Quantitative specification of semi-algebraic transition systems with metrics," *Journal of Information and Computational Science*, vol. 12, no. 3, pp. 993–1000, 2015.
[32] F. Liu, W. Pedrycz, and W.-G. Zhang, "Limited rationality and its quantification through the interval number judgments with permutations," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4025–4037, 2017.
[33] Y. Wu, H. Xu, C. Xu, and K. Chen, "Uncertain multi-attributes decision making method based on interval number with probability distribution weighted operators and stochastic dominance degree," *Knowledge-Based Systems*, vol. 113, pp. 199–209, 2016.
[34] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, PA, USA, 2009.
[35] R. Fletcher, *Practical Methods of Optimization*, Wiley, Hoboken, NJ, USA, 2013.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6618168, 11 pages
https://doi.org/10.1155/2020/6618168

Research Article

Formal Verification on the Safety of Internet of Vehicles Based on TPN and *Z*

Yang Liu,[1] **Liyuan Huang**,[1] **and Jingwei Chen**[2]

[1]*Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[2]*Chongqing Key Laboratory of Automated Reasoning and Cognition, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China*

Correspondence should be addressed to Jingwei Chen; chenjingwei@cigit.ac.cn

Received 2 November 2020; Revised 30 November 2020; Accepted 3 December 2020; Published 29 December 2020

Academic Editor: Yong Chen

Nowadays, the Internet of Vehicles has become the focus of global technological innovation and transformation in the automotive industry. Its flow modelling appears to play a very important role for designing and controlling the transportation systems, since it is not only necessary for improving safety and transportation efficiency but also can yield a series of society, economy, and ecosystem environment problems. Considering the characteristics of the frame structure includes states and actions and discrete and continuous aspects of traffic flow dynamics, both petri net and *Z* have proved to be useful tools for modelling the Internet of Vehicles. It can formally describe the vehicle behavior accurately with petri net and more details with *Z* frame structure. A new integration formal method of time petri net and *Z* is presented in this paper for modelling the vehicle behaviors and traffic rules through taking into account state dependencies on external rules. Moreover, a case study in the Internet of Vehicles is proposed to deal with the accurate localization of events. It shows that this formal verification methods significantly improves the safety and intelligence of the Internet of Vehicles.

1. Introduction

With the development of communication technology, wireless sensing technology, automatics, artificial intelligence, and so on, the Internet of Vehicles techniques come out. It is the achievements combined with the latest technological of computers and the modern automobile industry. Because of the complex and dynamic environment when it is working, the control system becomes more and more complex. Since it is about life, the key safety factor, such as automotive engine, air bag control, brake system, sensor monitoring system, and traffic regulations, have very strict reliability requirements. Internet of Vehicles has made our life convenient; nevertheless, at the same time, accidents still happen often. Many researchers ensure the safety from different aspects [1–3] by different methods, such as control strategy, security factor, and intelligent platform. More and more experiences show that the formal method is very effective to ensure the safety of the Internet of Vehicles [4–7] systems.

In fact, the formal method is a good way to inspect the problems in system design or requirement design [8, 9]. The running environment of the Internet of Vehicles is very complex and changes dynamically. It is hard to describe the Internet of Vehicle using only one single formal language. The traditional process analysis methods, such as Petri nets [10], CCS (Calculus of Communicating Systems) [11, 12], and CSP (Communication Sequential Processes) [13, 14], can model different aspects of the system from different angles and abstractions, but the powers of description for functional and nonfunctional attribute and constraint condition are deficient. The traditional model languages such as V [15, 16], B [17], and Z [18, 20] are good at modelling description, but poor at describing system concurrency. At present, the integrated specification languages are a hot topic, which produced CSPZ [21], TCOZ [22], PZN [23, 24], and so on. However, it seems that these languages do not aim at the Internet of Vehicles. PZN has a good advantage in describing traditional systems, since specification Z has a good frame structure both in state description

and operation description, and Petri nets [25–28] are very suitable to express the behavior of the parallel and concurrent system model. So, the hybrid methodology which combines the advantages of both specification Z and Petri nets is very suitable for modelling and analyzing the Internet of Vehicles system. PZN has been used to model and analyze validity and accessibility of networked software. Experimental results showed that PZN is very suitable to apply in it. In the Internet of Vehicles circumstance, except states and operation, time constraint is also very important. It not only has continuous part time but also has discrete time. Some researchers used time Petri nets to model the requirements and software of system [29–34], but it lacked specific rule descriptions and state depictions.

Motivated by the previous experience in formal verification of requirements modelling and analyzing of networked software, in this paper, TPZN (integration Time Petri Net and Z) is proposed to formal modelling and verifying the Internet of Vehicles systems. It is able to describe the concurrent process and fore-and-aft states in systems at different times. TPZN consists of two parts TPZN-TPN and TPZN-Z. TPZN-TPN defines the data flow of the whole structure, order, and behavior of process at one moment, while, TPZN-Z depicts the abstract data frame, specific rule restriction, and time constraint. So, based on enhancing the abstraction of the data and refinements by Z, the number of states of the Time Petri Nets can be decreased effectively. A case study shows the modelling method in detail. This formal method is proved greatly by improving the safety and validity of the intelligent vehicle systems.

2. Background

In this section, we recall some preliminary backgrounds that are necessary for the rest of the paper.

2.1. Hybrid Petri Net Extension. Hybrid petri net extension for traffic road modelling is proposed by Riouali et al. in [7]. It brought discrete parts and continuous parts which include discrete and continuous places and transitions. The moving and evolution of the Internet of Vehicles depend on the state of places and are governed by various function, namely, creation, destruction, merging, and splitting; meanwhile, it defined the speed, maximum density, length, and maximum flow of the traffic road modelling.

A hybrid petri net consists of three kinds of objects: places, transitions, and directed arcs. However, unlike the traditional petri net, here places are divided into two kinds: discrete places and continuous places. Transitions as well as places also fall into discrete transitions and continuous transitions. Arcs still show the state dynamic from places to transitions or from transitions to places. Hybrid petri net extension is defined 6-tuplet $N = (P, T, \text{Pre}, \text{Post}, \Upsilon, \text{Time})$.

(1) P is a set of places, $P = \text{Pc} \cup \text{Pd}$, where Pc represents continuous places and Pd represents discrete places.

(2) T is a set of transitions.

(3) Pre is the backward incidence matrix $P \times T \longrightarrow N$.

(4) Post is the forward incidence matrix $T \times P \longrightarrow N$.

(5) γ represents the batch place function. It associates with each batch place 4-tuplet (Vi: speed; di: a maximum density; Si: length; $\Phi^{\max}$: a maximum flow).

(6) Time represents the firing delay in case of continuous or batch transitions.

Here, we consider the time factor, while the γ is more suitable to be used in more intelligent vehicle concurrent environment.

2.2. Z Frame Structure. Z is a good formalism for modelling and designing. Compared with Petri Net, Z has better abilities in type definition and data abstraction and model refining. Its basic frame contains states and operations as Figure 1. Every operation has relative states and constrain rules. However, it does not describe the dynamic behavior of the systems.

Although Ding et al. and Wei et al. proposed a method that models systems by both Z and Petri Nets in [27, 28] and the authors also showed that using PZN (Z and Petri Nets) to model the requirements of software is an effective and feasible way [9], it is still not good enough to model the Internet of Vehicles. The reason is that PZN does not have the ability to describe the real-time performance which is very important in vehicle systems. In transportation systems, time is a very important factor. So, all previous works have to be improved and time constraints will be added in PZN [9]. TPZN stands for the integration of PZN and time factor. In Section 3, we will introduce the modelling and analysis methods by TPZN.

3. Modelling with TPZN

For satisfying the real-time capability and dynamic evolution and data abstraction and type definition capabilities of the Internet of Vehicles, the integrated specification TPZN is presented in this paper. Based on enhancing the abstraction of the data and refinements by Z, the state-of-the-time Petri Nets can be decreased effectively. Compared with time petri nets, color petri nets, PZN, and CSPZ, TPZN is more suitable to define the intelligent vehicle systems.

3.1. TPZN

Definition 1. A TPZN is a tuple $< P, T, F, Zp, Z_T, S, C, M_0, \text{SI}>$, where

(1) P is a set of the states.

(2) T is a set of the transitions.

(3) F is a set of the arcs which links state and transition.

(4) $N = (P, T, F)$ is a SISO net.

(5) $\text{TPN} = (P, T, F, M_0, \text{SI})$ is like a traditional time petri net.

(6) $\text{PZN} = (P, T, F, Zp, Z_T, S, C)$ is a PZN as in [9, 19].

(7) Zp is a set of the state frame based on Z.

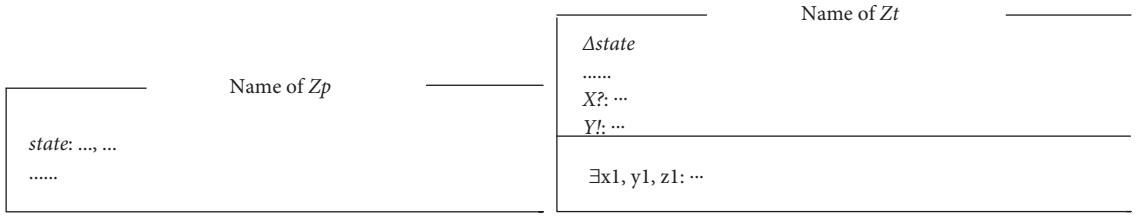


Figure 1: Frame structure of Z.

(8) Z_T is a set of the operation frame based on Z.

(9) S: $P \longrightarrow Zp$ is a set of the one-to-one map relationship between P and Zp.

(10) C: $T \longrightarrow Z_T$ is a set of the one-to-one map relationship between T and Z_T.

(11) M_0: is the initial mark, and $\exists t \in T, (p_0, t) \in F, M_0[t>$.

(12) $\exists \omega, \omega \in L(\text{TPN})$, φ_f $(\text{TPN}, \omega) = (M_f, D_f, \text{SI}_f)$, $M_0 = p_i + p_j + \cdots + p_k$, $D_0 = \{D_0(t_m), D_0(t_n), \ldots\}$, $\text{SI}_0 = [0, 0]$, $p_i, p_j, \ldots, p_k$ are all trigger states in the beginning and $t_m, t_{n,\ldots}$ are all trigger transitions. M_f represents the state of every node device in one time. D_f represents a set of the time interval of the next possible transition. SI_f represents the time interval of the system may need when it arrives M_f. φ_f represents the system's situation during time interval-SI_f. If M_f is the final state, $D_f = \varnothing$.

To ensure the compatibility and validity of the design, TPZN-Z frame is used to describe the sign, property, rules, and so on. The corresponding relation of TPN and Z is shown in Figure 2. The green dashed box is the precondition of transition. The rules and constraints are formally described by Z in Z_t. The purple dashed box represents the postcondition by Z.

3.2. Time Constrained in TPZN. This paper introduces global time and relative time for TPZN. The global time proves the standard system time, and the relative time supplies the time relative to previous status M_i. Here, it needs to define two variables. One is the earliest occurrence time, EAR(t), the other one is the latest occurrence time, $LAT(t)$. SI_i contains the earliest occurrence time EAR(t_i) and the latest occurrence time $\text{LAT}(t_i)$. $\text{SI}_i = [\text{EAR}(t_i), \text{LAT}(t_i)]$. $D_i(t)$ is the relative time to M_{i-1}, $M_{i-1}[t_i>$.

For example, in Figure 3, relative time is marked. For example, "$t7$ [15, 25]" means that $t7$ can be triggered in 15 seconds at least and 25 seconds at most. If it exceeds 25 seconds, the automatic delivery truck will stop working. Accordingly, the system will be warning. The global time is always synchronized with the time of the system.

3.3. Model Refining. The environment of the Internet of Vehicles running is always complex, dynamic, and unexpected so that model refining and topological evolution capability is to be very important. Suppose $\text{TPZN}_{11}$and TPZN_{12} are the subnet of TPZN_1:

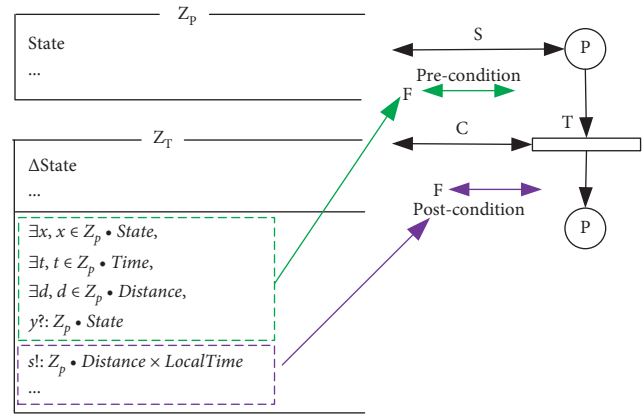


Figure 2: The relation between TPN and Z in TPZN.

$$\begin{aligned}\text{TPZN}_{11} &= \langle P_{11}, T_{11}, F_{11}, Zp_{11}, Z_{T11}, S_{11}, C_{11}, M_{011}, \text{SI}_{11}\rangle,\\ \text{TPZN}_{12} &= \langle P_{12}, T_{12}, F_{12}, Zp_{12}, Z_{T12}, S_{12}, C_{12}, M_{012}, \text{SI}_{12}\rangle.\end{aligned} \tag{1}$$

Then, $(\text{TPZN}_{11} \cap \text{TPZN}_{12}) \subset \text{TPZN}_1$. $\forall p_i, p_i \in P$, $P \in \text{TPZN}_{11}/(\text{TPZN}_{11} \cap \text{TPZN}_{12})$ are all the new additional virtual states which represent the possible states before or after the subnet TPZN_{11}. $\forall t_i, t_i \in T$, $T \in \text{TPZN}_{11}/$ $(\text{TPZN}_{11} \cap \text{TPZN}_{12})$ are all the new additional virtual transitions which represent the possible preconditions or postconditions. Of course, new Z frame structure Z_p' and Z_t' should be redefined by additional rules. In the similar way, a new TPZN′ can substitute a transition t_i, when the control structure change.

On the contrary, when one model is needed to be abstracted, it can be seen as a new transition t'; then adding its precondition and postcondition and reserving input and output are relative to the conterminal model.

Theorem 1. *If the global execution time of every transition sequence of the new refined TPZN model from the beginning to the end is equal to the execution time of the substituted t_i of the original TPZN, the new refined TPZN can maintain behavioral consistency with the original one.*

Because TPZN integrates TPN and Z, the refined TPN can maintain behavioral consistency with the original one and has been proved in [35–37].

4. Modelling Analysis

4.1. Accessibility. Traditionally speaking, there are two ways to analysis the accessibility of the model. One way is using

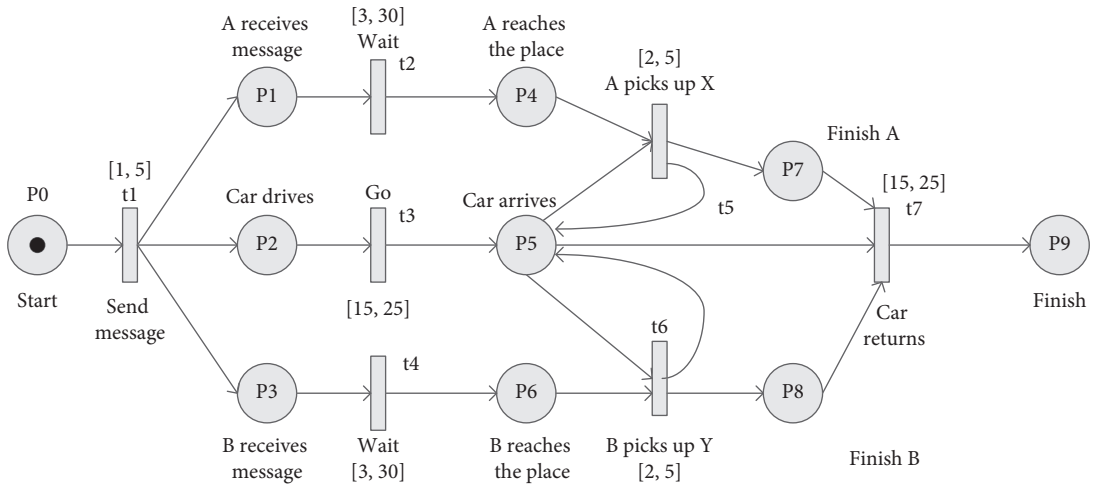


Figure 3: The TPZN of automatic delivery truck.

the reachability tree which is used to analysis the accessibility of model states. Because the accessibility of the TPZN involves limited time and there are lots of the state classes, some methods to reduce the state classes are necessary. For instance, Bourdil and Berthomieu have proposed some methods to reduce the state classes [28, 31]. Based on their work, we use Z frame to abstract the system so to reduce the state number. The layer can be subdivided into smaller layers. If the lowest layers can be verified to be correct, accessible, and safe, the whole upper layer will have the same character. The reachability tree can be built by φ_f based on TPZN. From φ_{fi} to φ_{fj}, the path from the node φ_{fi} of the tree to the node φ_{fj} shows the transition sequence (Figure 4).

The other way is using the incidence matrix marked $C(C = D^+ - D^-)$. Here, the output matrix-D^+ is defined as

$$D^+[i,j] = \begin{cases} 0, & \nexists f_k = (t_i, p_j), f_k \in T \times P, \\ n, & \exists f_k = (t_i, p_j), f_k \in T \times P \wedge \text{Token}_{Pj} = n, \end{cases} \tag{2}$$

where $D^+[i, j] = 0$ means there does not exist an arc from the t_i to p_j. While, $D^+[i, j] = n$ means that there is an arc from the t_i to p_j, and it will produce n same type elements with the transfer. The (i, j) entry of D^- is defined as

$$D^-[i,j] = \begin{cases} 0, & \nexists f_k = (p_i, t_j), f_k \in P \times T, \\ n, & \exists f_k = (p_i, t_j, f_k \in P \times T \wedge \text{Token}_{Pi} = n] > t_j, \end{cases} \tag{3}$$

where $D^-[i, j] = 0$ means there is not an arc from the p_i to t_j, while $D^-[i, j] = n$ means that there is an arc from the p_i to t_j and the transition can happen only if there is n same type elements in the p_j.

Supposing M_i is a marked state. From M_i to M_j, if there is an transition sequences $\sigma = t_i t_{i+1}, \ldots, t_j$ marked by X-vector quantity and it satisfies $M_j = M_i + X \bullet (D^+ - D^-)$, it proves that the Mi state is accessibility. However, in TPZN, it must consider the limited time. The time constrained rules are described by Z frame. In the automotive vehicles system,

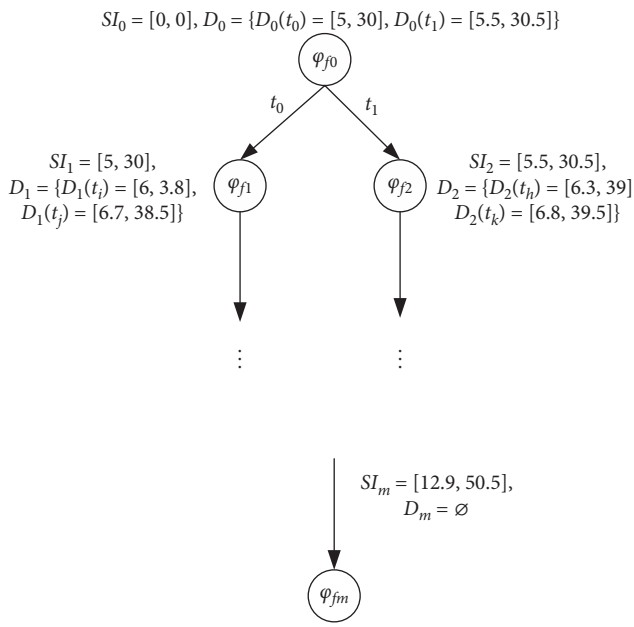


Figure 4: The TPZN of automatic delivery truck.

time constrained rules must be built strictly because subtle time change may cause serious traffic accident. So, modelling the vehicles' system, it needs to abstract the whole system, then subdivide the whole system into specific layers, and go on subdividing until it is subdivided into atom modules. By φ_f which represents the state class containing timestamp, we can get the possible behavior information of the system in certain time interval and then predict the next step. The algorithm of accessibility is designed as Algorithm 1 which shows the accessibility decision from Mi to Mj, and the case study explains how to use it in Figure 5.

4.2. Validity. The validity of the control structure can be analyzed by the transfer matrix L_{DP} of TPZN. From the L_{DP}, concurrent transition can be obtained by the same column and row. As the following in L_{DP1}, t_1 and t_2 can be trigger

simultaneously from p_0 to p_1 and p_e, while, if p_1 is arrived, t_1 and t_3 must be triggered:

$$L_{DP1} = \begin{array}{c} \\ p_0 \\ p_1 \\ \vdots \\ p_e \end{array} \begin{array}{c} \begin{array}{cccc} p_0 & p_1 & \cdots & p_e \end{array} \\ \begin{bmatrix} 0 & t_1 & \cdots & t_2 \\ 0 & t_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}. \quad (4)$$

So, the data flow structure can be mapped into the transfer matrix L_{DP}. If there exist several transitions in the same row p_i, it means when the system arrives into the state p_i, these transitions will be simultaneously triggered. While if there exist several transitions in the same column p_i, it means only under the condition that all the transitions are triggered, and p_i can be reached.

After getting the initial model and parameters, the sampled data or historical data can be used to correct the model and parameters. Of course, real time data also can be used to modify the model and parameters, but more often, it is used to predict possible state of the future.

The process of modelling the Internet of Vehicles with TPZN is as Figure 6. First, the node device information and traffic rules and evaluation indicators are obtained from the initial system model. Meanwhile, the data flow structure of the system should be obtained, and divide the initial system into subsystem. Second, the foregoing information is described by Z frame structures, and the latter is described by TPN. Third, the subsystem should be refined. Then, the whole system can be formally modelled by TPZN. Next, the related parameters such as L_{DP}, φ_f, D^+, and D^- can be obtained from the TPZN model. Combined with the current information of the system, the initial parameters are used to analyze the character. At last, the future behavior of the vehicle system can be predicted. If the prediction shows, it will be in danger, and some strategies can be adopted. If the danger is caused by some traffic rules, these rules will be modified.

4.3. Advantage. Compared with TPN, PZN, and Z, TPZN has better dynamic structure and more convenient time constraint which are very important to the Internet of Vehicles. Except these, TPZN has better frame structure which can abstract the system to reduce the number of the states to avoid the explosive growth usually happened in traditional Petri Net. So, the advantage of modelling with TPZN is shown very clearly in Table 1.

5. A Case Study

To verify effectiveness of our modelling methods to analyze our verification algorithms, in this section, a simple case study is offered. Suppose that an intelligent car has 4 lidars, 4 radars, 4 side vision, 1 full vision, image processing system, radar system, lidar system, brake system, and so on. It is running on the straight road, as shown in Figure 7.

For modelling the system, the first step is to obtain the Z frame structure of every node device. Here, parts of the system model's, such as Z_p and Z_t, are put forward as space is limited.

CAR

Number: *number*
Brand: *Volk*, *Ford*, *Benz*, *BMW*, ···
Fuel: *Gasoline*, *Electric*, ···
FuelState: *full*, *over*, *normal*
Lidar: *FrontLeftLi*, *FrontRightLi*, *FrontMiddleLi*, *BackLeftLi*, *BackRightLi*
Radar: *FrontLeftRa*, *FrontRightRa*, *BackLeftRa*, *BackRightRa*
Vision: *FrontLeftVi*, *FrontRightVi*, *BackLeftVi*, *BackRightVi*, *FullVi*
ProcessSystem: *RadarSystem*, *LidarSystem*, *VisionSystem*, *BrakeSystem*, ···
.....
State: *Start*, *Stop*, *Brake*, *Acceleration*, *Deceleration*, *Back*, *TurnLeft*, *TurnRight*, ···

The above frame is the same parts of one element of the Z_p, which is defined as one kind of state of the system. As the blue dashed box shows, it formally defines relative devices. The following one defines one node device of the system.

FrontLeftLi

Name: *Lidar*
Time: *GlobalTime*, *LocalTime*
Distance: *LongDistance*, *LimitDistance*, *SafeDistance*
Speed: *Distance* X *LocalTime*, *ConstrainSpeed*
*StateLi*1: *Work*, *Rest*
.....

The next frame is one element of the Z_t which defines one kind of possible transition of the system.

BEGIN

Δ CAR
Δ FrontLeftLi
Δ FrontRightLi
Δ FrontMiddleLi
Δ BackLeftLi
Δ BackRightLi
Δ FrontLeftRa
Δ Door
.....
x?: *CAR.State*
x1!: *FrontLeftLi.StateLi*1
x2!: *FrontRightLi.StateLi*2
x3!: *FrontMiddleLi.StateLi*3
x4!: *BackLeftLi.StateLi*4
x5!: *BackRightLi.StateLi*5
z!: *Door.StateDoor*
.....

∃*n*: *CAR.Number*, ∃*y*: *CAR.FuelState*···
((*n* ∈ *number*) ∧ (*x*? ∈ *Start*) ∧ *y* ∉ *over* ∧ ···)
→(*x*1! = 1) ∧ (*x*2! = 1) ∧ (*x*3! = 1) ∧ (*x*4! = 1) ∧ (*x*5! = 1) ∧ (*z*! = {1, 1, 1, 1})···
...

So, at the first step, every node device's Z frame structure and every transition can be defined. In second step, the TPN model of the Internet of Vehicles system will be built. Parts of the TPN model are shown in Figure 8.

Then, the TPZN of this case is <P, T, F, Zp, Z_T, S, C, M_0, SI>, where

(1) $P = \{p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}\}$.

(2) $T = \{t_0, t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}$.

Input: $\varphi_{\mathbf{f}} = \{\varphi_{\mathbf{f0}}, \varphi_{\mathbf{f1}}, \varphi_{\mathbf{f2}}, \ldots, \varphi_{\mathbf{fe}}\}$, M_i, M_j, D^+, D^-
Output: true (print the path); false
Find the X, $X = (\mathbf{M_j} - \mathbf{M_i}) \bullet (\mathbf{D}^+ - \mathbf{D}^-)^{-1}$
If X not exist, return false;
  Else
    For ($k = 0$; $k < n$; k++)
      $\sigma_{\mathbf{k}} = \mathbf{t_h}, \mathbf{t_{h+1}}, \ldots, \mathbf{t_{h+c}}$;//$\sigma_{\mathbf{k}}$ store the different value of X, n is the number of X.
$\varphi_{\mathbf{f0}}$ is the root node;//built the reachability tree
For ($k = 1$; $k \le e$; k++)
    {if ($\exists \mathbf{t_m}, \mathbf{t_m} \in \mathbf{T}$, $\mathbf{M_k}] > \mathbf{t_m}$) $\wedge$ ($[\mathbf{SI_k} \cdot \mathbf{EAR}(\mathbf{t_k}), \mathbf{SI_k} \cdot \mathbf{LAR}(\mathbf{t_k})] \subseteq \{\mathbf{system}(\mathbf{t}) + \mathbf{interval\ time}(\mathbf{t_k})\}$
    $\varphi_{\mathbf{fk}}$ is the child node of $\varphi_{\mathbf{f(k-1)}}$;
    }//test the time constrain
For($k = 0$; $k < n$; k++)
  {If ($\sigma_{\mathbf{k}} = \mathbf{t_h}, \mathbf{t_{h+1}}, \ldots, \mathbf{t_{h+c}}$) exist in one path of $\varphi_{\mathbf{fi}}$ to $\varphi_{\mathbf{fj}}$,
    Lookup(S, C); //find the relative $\mathbf{Z}'_{\mathbf{p}}$ and $\mathbf{Z}'_{\mathbf{T}}$, test the logical relationship
    If the logical relationship from $\mathbf{Z_{pa}}, \mathbf{Z_{pb}}, \ldots, \mathbf{Z_{p\,d}}$ $((\mathbf{M_i} = \mathbf{P_a} + \mathbf{P_b} + \cdots + \mathbf{P_d}), \mathbf{Z_{pa}}, \mathbf{Z_{pb}}, \ldots, \mathbf{Z_{p\,d}} \in \mathbf{Z}'_{\mathbf{p}})$ to $\mathbf{Z_{pe}}, \mathbf{Z_{pf}}, \ldots, \mathbf{Z_{pr}}$ $((\mathbf{M_j} = \mathbf{Z_{pe}} + \mathbf{Z_{pf}} + \cdots + \mathbf{Z_{pr}}), \mathbf{Z_{pe}}, \mathbf{Z_{pf}}, \ldots, \mathbf{Z_{pr}} \in \mathbf{Z}'_{\mathbf{p}})$ is reasoned to be correct.
      Print $\varphi_{\mathbf{fi}}, \mathbf{t_n}, \varphi_{\mathbf{fi+1}}, \mathbf{t_{n+1}}, \ldots, \mathbf{t_{n+c}}, \varphi_{\mathbf{fj}}$;
  }

Algorithm 1: The algorithm of accessibility. Accessibility decision from M_i to M_j.

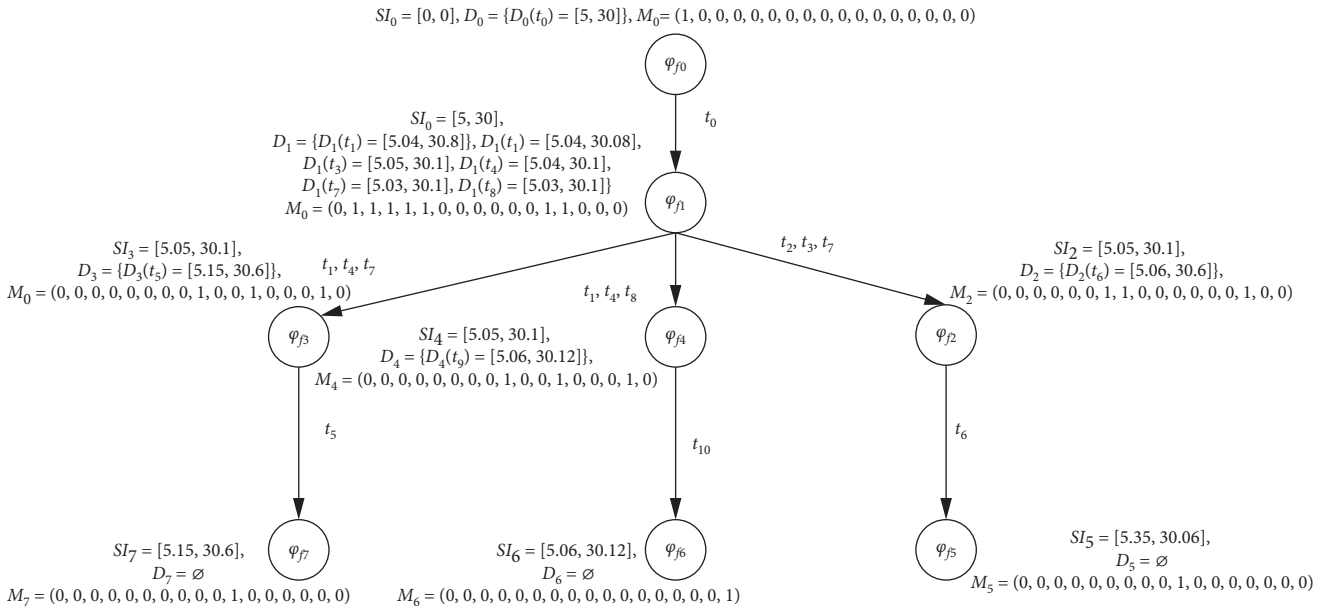


Figure 5: Reachability tree of the case study.

(3) F is the set of arcs in Figure 8. The elements are like the following form (p_0, t_0), (t_0, p_1), (t_0, p_2), (t_0, p_3), (t_0, p_4), (t_0, p_5), (p_1, t_1), ….

(4) Z_{pi} is the element of the set of Z_p, and it represents the state of Z frame of the node devices as CAR and FrontLeftLi.

(5) Z_{ti} is the element of the set of Z_T, and it represents the transition of Z frame of the system as BEGIN.

(6) S maps the relationship from state pi to Z frame of the state, as p_0- > CAR.

(7) C maps the relationship from transition ti to Z frame of the transition, as t_0- > BEGIN.

(8) M_0=(1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) represents the initial condition of the system.

SI_i is shown in Table 2, which represents the temporal interval under M_i. Some of the details of each p_i and t_i are shown as Table 3. Figure 7 shows parts of the case study, so, the p_9 and p_{10} are not the real final states. In fact, p_9 and p_{10} can turn into normal state by some steps.

From Figure 8, the final state classes are φ_{f5}, φ_{f6}, and φ_{f7}, where φ_{f5} is the emergency brake, $\varphi_{f7}$is slow

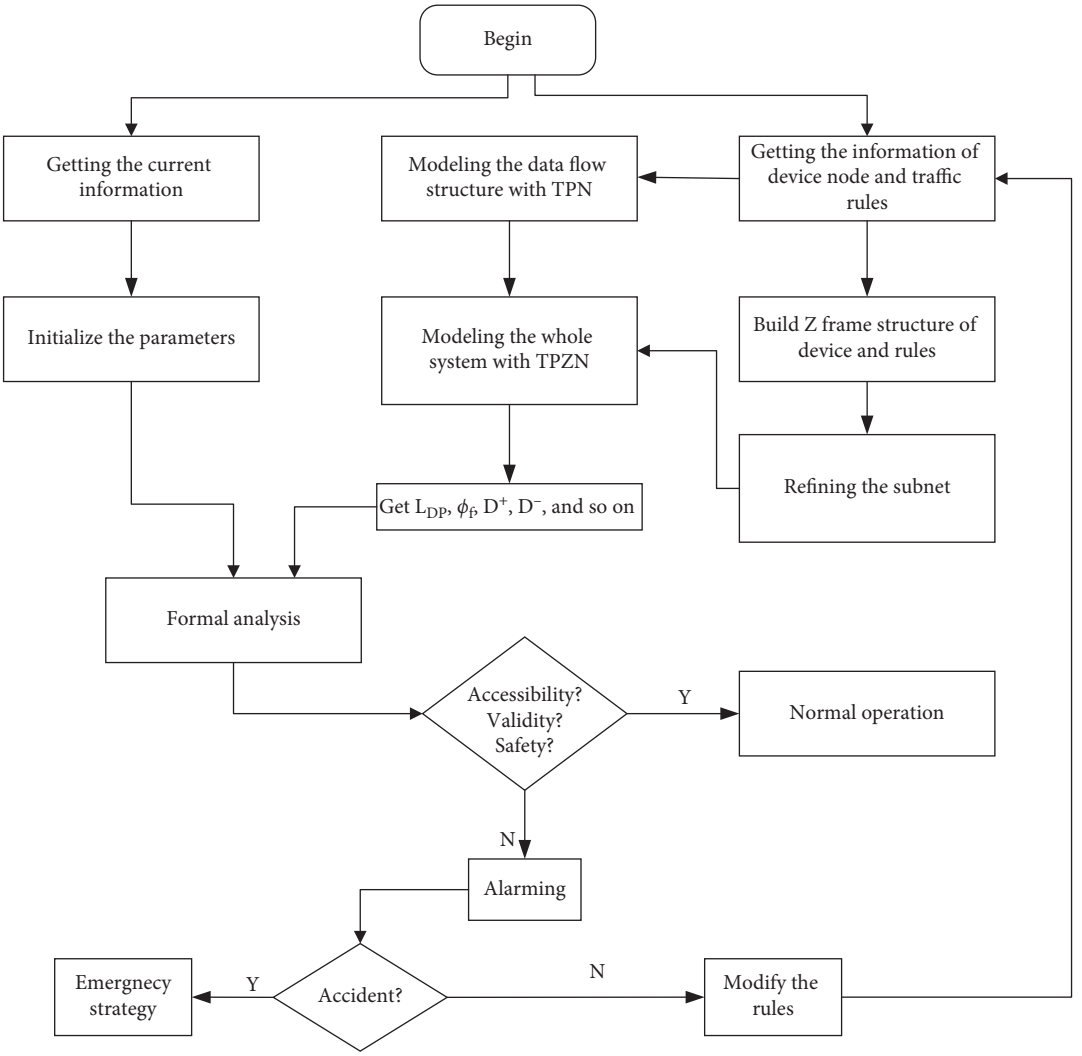


Figure 6: The flow diagram of modelling with TPZN.

Table 1: Compared TPZN, TPN, PZN, and *Z*.

	Dynamic structure	Frame structure	Number of states	Time constraint
TPZN	Good	Good	Abstract	Good
TPN	Good	Not good	Explosive growth	Good
PZN	Good	Good	Abstract	Not good
Z	Not good	Good	Abstract	Not good

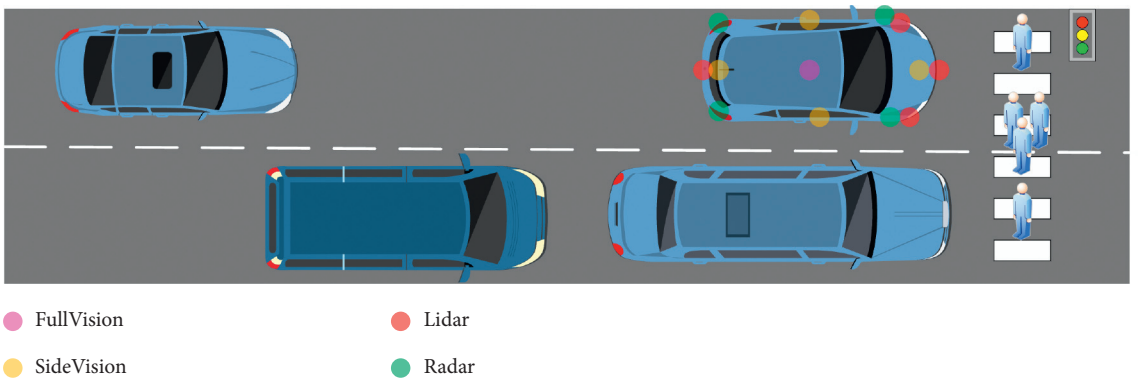


Figure 7: The environment of a case study.

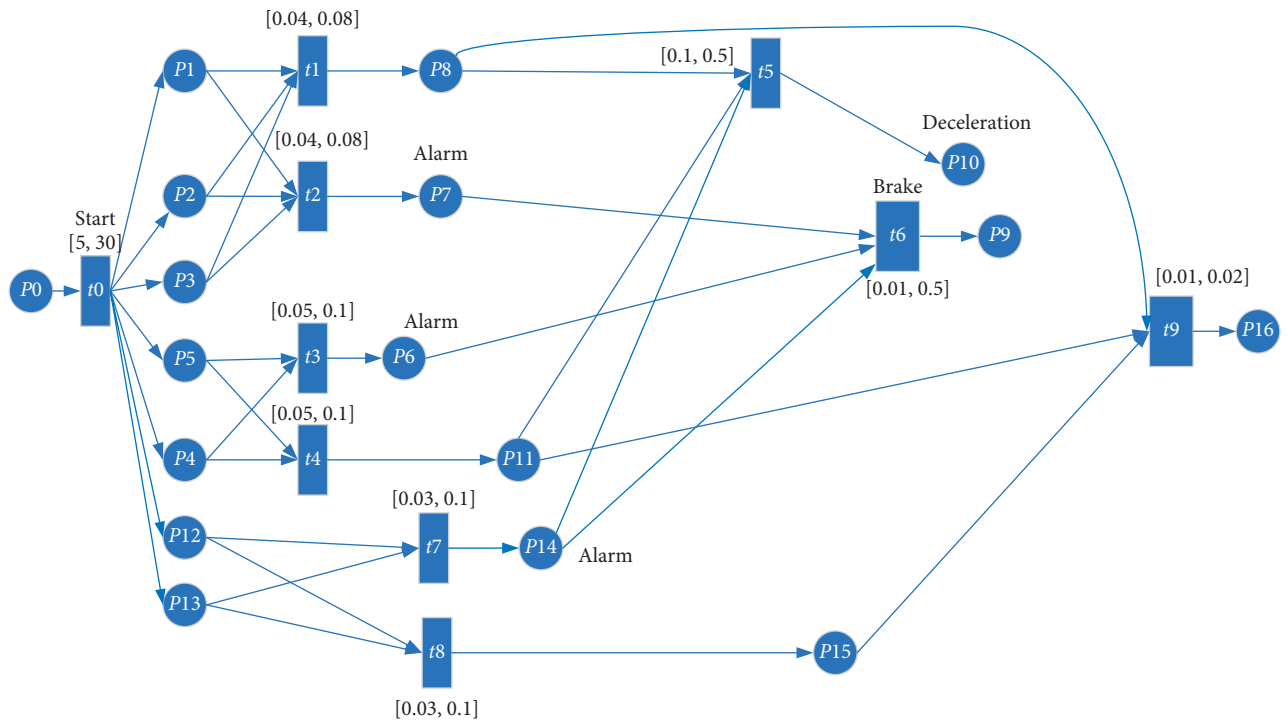


Figure 8: TPN model of the case study.

Table 2: The detail of SI.

I	φ_{fi}	$\mathbf{M_i}$	$\mathbf{D_i}$	$\mathbf{SI_i}$
$i=0$	$(M_0, D_0, \mathrm{SI}_0)$	P_0	$\{D_0(t_0) = [5, 30]\}$	$[0, 0]$
$i=1$	$(M_1, D_1, \mathrm{SI}_1)$	$P_1 + P_2 + P_3 + P_4 + P_5 + P_{12} + P_{13}$	$\{D_1(t_1) = [5.04, 30.08],$ $D_1(t_2) = [5.04, 30.08],$ $D_1(t_3) = [5.05, 30.1],$ $D_1(t_4) = [5.05, 30.1],$ $D_1(t_7) = [5.03, 30.1],$ $D_1(t_8) = [5.03, 30.1]\}$	$[5, 30]$
$i=2$	$(M_2, D_2, \mathrm{SI}_2)$	$P_7 + P_6 + P_{14}$	$\{D_2(t_6) = [5.06, 30.6]\}$	$[5.05, 30.01]$
$i=3$	$(M_3, D_3, \mathrm{SI}_3)$	$P_8 + P_{11} + P_{14}$	$\{D_3(t_5) = [5.15, 30.6]\}$	$[5.05, 30.1]$
$i=4$	$(M_4, D_4, \mathrm{SI}_4)$	$P_8 + P_{11} + P_{15}$	$\{D_4(t_9) = [5.06, 30.12]\}$	$[5.05, 30.1]$
$i=5$	$(M_5, D_5, \mathrm{SI}_5)$	P_9	$\varnothing$	$[5.35, 30.6]$
$i=6$	$(M_6, D_6, \mathrm{SI}_6)$	P_{16}	$\varnothing$	$[5.06, 30.12]$
$i=7$	$(M_7, D_7, \mathrm{SI}_7)$	P_{10}	$\varnothing$	$[5.15, 30.6]$

Table 3: The details of states and operations.

P_0	The start of intelligent car	P_{14}	Obstacles, traffic light, and so on
P_1	Lidar 1	P_{15}	Normal environment
P_2	Lidar 2	P_{16}	Keep running
P_3	Lidar 3	t_0	Start
P_4	Radar 1	t_1	Processed normal data by lidar system
P_5	Radar 2	t_2	Processed abnormal data by lidar system
P_6	Detected obstacles ahead by radar	t_3	Processed normal data by radar system
P_7	Detected obstacles ahead by lidar	t_4	Processed abnormal data by radar system
P_8	Detected normal environment by radar	t_5	Decelerating
P_9	Brake	t_6	Braking
P_{10}	Deceleration	t_7	Process by vision-front
P_{11}	Detected normal environment by lidar	t_8	Process by wide-angle
P_{12}	Vision-front	t_9	Check information
P_{13}	Wide-angle		

down, and φ_{f6} is running straight normally. The transfer matrix L_{DP}, D^+, and D^- of this case study is as follows:

$$
L_{DP} = \begin{array}{c} \\ p_0 \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \\ p_9 \\ p_{10} \\ p_{11} \\ p_{12} \\ p_{13} \\ p_{14} \\ p_{15} \\ p_{16} \end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccc} p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} & p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \end{array} \\
\begin{bmatrix}
0 & t_0 & t_0 & t_0 & t_0 & t_0 & 0 & 0 & 0 & 0 & 0 & 0 & t_0 & t_0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & t_2 & t_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & t_2 & t_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & t_2 & t_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & t_3 & 0 & 0 & 0 & 0 & t_4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & t_3 & 0 & 0 & 0 & 0 & t_4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_5 & 0 & 0 & 0 & 0 & 0 & t_9 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_5 & 0 & 0 & 0 & 0 & 0 & t_9 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_7 & t_8 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_7 & t_8 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_6 & t_5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & t_9 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array},
$$

$$
D^- = \begin{array}{c} \\ t_0 \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccc} p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} & p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \end{array} \\
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}
\end{array},
$$

$$
D^+ = \begin{array}{c} \\ t_0 \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccc} p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_{10} & p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \end{array} \\
\begin{bmatrix}
0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}.
$$

(5)

From the matrix L_{DP}, the concurrent behavior can be easily found. By the D^+, D^-, M_i, M_j, φ_{fi}, and φ_{fj}, the next behavior can be deduced exactly. The exact arrival time can also be obtained from SI_i and SI_j from the reachability tree as shown in Figure 5. The rules can be amended through the Z_p and Z_t with the new data coming as well. Every Z frame

structure can be coded by high-level programming language so to reason the logic relationship easily.

6. Conclusions

In this paper, we propose a new way that uses TPN and *Z* frame structure to formally model and analyze the safety and accessibility of the Internet of Vehicles. The method has been explained in detail by a case study. Although it promotes the efficiency of finding problem when the system goes wrong and can predict the future behavior, the multiple intelligent vehicles working cooperatively are not taken into account, which is an important and intriguing topic that we are working on.

Data Availability

The case study data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation under Grants 61903053 and 61703063, Science and Technology Research Project of Chongqing Municipal Education Commission of China under Grants Nos. KJZD-K201800701 and KJQN201900702, Chongqing Engineering and Technology Research Center for Big Data of Public Transportation Operation under Grant 2019JTDSJ-YB02, and Guizhou Science and Technology Program [2020] 4Y056.

References

[1] C. M. Martinez, M. Heucke, F. Y. Wang et al., "Driving style recognition for intelligent vehicle control and advanced driver assistance: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1–11, 2018.

[2] Y. Quan, H. Yang, and L. Yang, "Information security impacts future traffic safety of intelligent vehicle," in *Proceedings of the International Conference on Man-Machine-Environment System Engineering*, pp. 731–738, Beijing, China, 2018.

[3] L. B. Chen, H. Y. Li, W. J. Chang et al., "An intelligent vehicular telematics platform for vehicle driving safety supporting system," in *Proceedings of the International Conference on Connected Vehicles & Expo*, pp. 210-211, Shenzhen, China, October 2015.

[4] M. Kamali, L. A. Dennis, O. Mcaree, M. Fisher, and S. M. Veres, "Formal verification of autonomous vehicle platooning," *Science of Computer Programming*, vol. 148, pp. 88–106, 2017.

[5] Y. Teng, L. Qi, and Y. Du, "A logic petri net-based repair method of process models with incomplete choice and concurrent structures," *Computing and Informatics*, vol. 39, no. 1-2, pp. 264–297, 2020.

[6] A. Boucherit, L. M. Castro, A. Khababa, and O. Hasan, "Petri net and rewriting logic based formal analysis of multi-agent based safety-critical systems," *Multiagent and Grid Systems*, vol. 16, no. 1, pp. 47–66, 2020.

[7] Y. Riouali, L. Benhlima, and S. Bah, "Petri net extension for traffic road modelling," *Computer Systems & Applications*, vol. 7, no. 11, pp. 7–12, 2017.

[8] Y. Liu, J. Z. Wu, and R. Qiao, "Consistency verification between goal model and process model in requirements analysis of networked software," *Journal of Computational and Theoretical Nanoscience*, vol. 11, no. 5, pp. 1248–1261, 2014.

[9] Y. Liu, J. Z. Wu, and R. Qiao, "Dynamic evolution of requirements process model deployed on networked environment with PZN," *Journal of Computational Information Systems*, vol. 9, no. 8, pp. 3329–3336, 2013.

[10] C. Liu, Q. Zeng, H. Duan et al., "Petri net based data-flow error detection and correction strategy for business processes," *IEEE Access*, vol. 8, pp. 43265–43276, 2020.

[11] R. Bruni and U. Montanari, "CCS, the calculus of communicating systems," *Models of Computation*, pp. 221–270, Springer, Berlin, Germany, 2017.

[12] R. C. Bhushan and D. K. Yadav, "Modelling a safety-critical system through CCS," *International Journal of Applied Engineering Research*, vol. 12, no. 21, pp. 11213–11217, 2017.

[13] J. Whitney, C. Gifford, and M. Pantoja, "Distributed execution of communicating sequential process-style concurrency: golang case study," *The Journal of Supercomputing*, vol. 75, no. 3, pp. 1396–1409, 2019.

[14] M. Hatzel, C. Wagner, K. Peters, and U. Nestmann, "Encoding CSP into CCS," *Electronic Proceedings in Theoretical Computer Science*, vol. 190, pp. 61–75, 2015.

[15] V. Bandur V, P. W. V. Tran-Jørgensen, M. Hasanagic et al., "Code-generating VDM for embedded devices," in *Proceedings of the 15th Overture Workshop*, London, UK, October 2017.

[16] M. Hasanagić, T. Fabbri, P. G. Larsen et al., "Code generation for distributed embedded systems with VDM-RT," *Design Automation for Embedded Systems*, vol. 23, no. 3-4, pp. 153–177, 2019.

[17] D. Sabatier, "Using formal proof and B method at system level for industrial projects," *Reliability, Safety, and Security of Railway Systems*, pp. 20–31, Springer, Berlin, Germany, 2016.

[18] P. Saratha, G. V. Uma, and B. Santhosh, "Formal specification for online food ordering system using Z language," in *Proceedings of the 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, pp. 343–348, IEEE, Tindivanam, India, February 2017.

[19] G. O'Regan, "Z formal specification language," *Concise Guide to Formal Methods*, pp. 155–171, Springer, Berlin, Germany, 2017.

[20] Z. H. Muhamad, D. A. Abdulmonim, and B. Alathari, "An integration of uml use case diagram and activity diagram with Z language for formalization of library management system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, p. 3069, 2019.

[21] T. Gouasmi, A. Regayeg, and A. H. Kacem, "Automatic generation of an operational CSP-Z specification from an abstract temporalZ specification," in *Proceedings of the 2012 IEEE 36th Annual Computer Software & Applications Conference Workshops*, pp. 248–253, Izmir, Turkey, July 2012.

[22] B. Mahony and S. D. Jin, "Blending object-Z and timed CSP: the semantics of TCOZ," in *Proceedings of the ICSE'98 Proceedings of the 20th International Conference on Software Engineering*, pp. 95–104, Kyoto, Japan, April 1998.

[23] Y. Liu, J. Z. Wu, R. Zhao et al., "Formal verification of process layer with petri nets and Z," *Advances in Information Sciences and Service Sciences*, vol. 5, no. 1, pp. 68–77, 2013.
[24] F. Peschanski and D. Julien, "When concurrent control meets functional requirements or Z+Petri nets," *ZB 2003: Formal Specification and Development in Z and B*, pp. 79–97, Springer, Berlin, Germany, 2003.
[25] T. Yin, Z. Li, C. Seatzu et al., "Verification of state-based opacity using petri nets," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2823–2837, 2017.
[26] X. Wu, S. Tian, and L. Zhang, "The Internet of Things enabled shop floor scheduling and process control method based on Petri nets," *IEEE Access*, vol. 7, pp. 27432–27442, 2019.
[27] Z. Ding, Y. Zhou, and M. C. Zhou, "Modeling self-adaptive software systems with learning petri nets," *IEEE Transactions on Systems Man & Cybernetics Systems*, vol. 46, no. 4, pp. 483–498, 2017.
[28] L. Wei, W. Lu, Y. Du et al., "Deadlock property analysis of concurrent programs based on petri net structure," *International Journal of Parallel Programming*, vol. 45, no. 4, pp. 1–20, 2016.
[29] M. Gaied, A. M'halla, D. Lefebvre, and K. Ben Othmen, "Robust control for railway transport networks based on stochastic P-timed Petri net models," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 233, no. 7, pp. 830–846, 2019.
[30] E. Kučera, O. Haffner, P. Drahoš et al., "New software tool for modeling and control of discrete-event and hybrid systems using timed interpreted petri nets," *Applied Sciences*, vol. 10, no. 15, p. 5027, 2020.
[31] H. B. Attia, L. Kahloul, S. Benhazrallah et al., "Using hierarchical timed coloured petri nets in the formal study of TRBAC security policies," *International Journal of Information Security*, vol. 19, no. 2, pp. 163–187, 2020.
[32] B. Aman, P. Battyányi, G. Ciobanu et al., "Local time membrane systems and time petri nets," *Theoretical Computer Science*, vol. 805, pp. 175–192, 2018.
[33] R. Cao, L. Hao, F. Wang et al., "Modelling and analysis of hybrid stochastic timed Petri net," *Journal of Control and Decision*, vol. 6, no. 3, pp. 1–21, 2018.
[34] P.-A. Bourdil, B. Berthomieu, S. Dal Zilio, and F. Vernadat, "Symmetry reduction for time Petri net state classes," *Science of Computer Programming*, vol. 132, pp. 209–225, 2016.
[35] B. Berthomieu, D. L. Botlan, and S. D. Zilio, *Petri Net Reductions for Counting Markings*, pp. 1–20, Springer, Berlin, Germany, 2018.
[36] C. J. Jiang and Z. J. Ding, *Petri Net Refinement Based System Modeling and Analysis*, pp. 91–105, Tongji University Press, Shanghai, China, 2017.
[37] H. Duan, C. Liu, Q. Zeng et al., "Refinement-based hierarchical modeling and correctness verification of cross-organization collaborative emergency response processes," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 8, pp. 2845–2859, 2018.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6643167, 13 pages
https://doi.org/10.1155/2020/6643167

Research Article

A Fault Diagnosis Method of Rolling Bearing Integrated with Cooperative Energy Feature Extraction and Improved Least-Squares Support Vector Machine

Zhang Xu, Darong Huang, Tang Min, and Yunhui Ou

Chongqing Jiaotong University, School of Information Science and Engineering, Chongqing 400074, China

Correspondence should be addressed to Darong Huang; drhuang@cqjtu.edu.cn

Received 5 November 2020; Revised 23 November 2020; Accepted 10 December 2020; Published 24 December 2020

Academic Editor: Yong Chen

To solve the problem that the bearing fault of variable working conditions is challenging to identify and classify in the industrial field, this paper proposes a new method based on optimization of multidimension fault energy characteristics and integrates with an improved least-squares support vector machine (LSSVM). First, because the traditional wavelet energy feature is difficult to effectively reflect the characteristics of rolling bearing under different working conditions, based on analyzing the wavelet energy feature extraction in detail, a collaborative method of multidimension fault energy feature extraction combined with the method of Transfer Component Analysis (TCA) is constructed, which improves the discrimination between the different features and the compactness between the same features of rolling bearing faults. Then, for solving the problem of the local optimal of particle swarm optimization (PSO) in fault diagnosis and recognition of rolling bearing, an improved LSSVM based on particle swarm optimization and wavelet mutation optimization is established to realize the collaborative optimization and adjustment of LSSVM dynamic parameters. Based on the improved LSSVM and optimization of multidimensional energy characteristics, a new method for fault diagnosis of rolling bearing is designed. Finally, the simulation and analysis of the proposed algorithm are verified by the experimental data of different working conditions. The experimental results show that this method can effectively extract the multidimensional fault characteristics under variable working conditions and has a high fault recognition rate.

1. Introduction

In industrial equipment, the rolling bearing is an essential part of high-speed rating machinery. During actual operation, once the rolling bearings have failures (such as internal cracking, abrasion, external cracking, et al.), the safety and reliability of the entire system will be directly affected. In this situation, the condition monitoring and fault diagnosis of rolling bearings have become a hot research topic in prognostic and health management (PHM) of industrial systems.

Over the past decade, the accretion data, which reflect the running performance of rolling bearings, are usually used to analyze and test the fault characteristics of the rolling bearings. However, because the working operation of rolling bearings is influenced by various kinds of dynamic factors, the fault characteristics of the acceleration data may be quickly submerged by the ambient noise. Thus, these all cause a huge difficulty in diagnosing the fault of rolling bearings. And fortunately, when the fault characteristic of rolling bearings is located in the blind areas, the energy waveform of the original signal shows the characteristics of high efficiency and low amplitude. So, some researchers have tried to extract the energy feature from the acceleration data to accomplish the target of fault diagnosis for the rolling bearings. Unfortunately, because the fault data signal of rolling is nonlinear and nonstationary vibration response in the industrial environment, the energy characteristics extracted by based-wavelet packet and improved methods cannot effectively distinguish the differences of different features and the tightness of the similar characteristics for the complex industrial environment.

To extract useful information of energy features, Wavelet Theory is a usual method to analyze the vibration data of rolling bearing in previous work [1]. The key reason is that

the wavelet packets can adaptively be selected according to the characteristics of the signal and may divide a frequency band into multiple frequency bands. Based on this, some scholars had presented some improved extracted method (such as wavelet packet transform (WPT), the fuzzy mutual information of wavelet packet transform (FMIWPT), dual-tree complex wavelet packet transform, support vector machine based- WPT, complex wavelet packet energy moment entropy and maximal overlap wavelet packet transform, et al.) of energy feature in [2–7]. These methods may not only implement the initial enhancement of the fault feature but also extract multiple permutation entropy features in the real application.

To address this problem, some scholars have tried to decompose the fault data signal into different frequency bands by using the wavelet packet and reconstruct the nodes in the frequency band in [8]. The advantage of the method is that characteristic frequency points may be located as quickly as possible in an industrial environment. Meanwhile, to treat the irregular vibration signal, the fault features may be extracted in time domain by using wavelet transform (see [9]). Besides, the optimization of structure of the energy characteristics has been also discussed briefly using Transfer Component Analysis (TCA) in [10]. By using the algorithm, the data properties may be preserved and the data distributions in different domains may converge to a stable scale. However, the running state of rolling bearings is affected by the endogenous factors; the different decomposition depths of the energy features are a very key problem in our working condition. Due to the diversity and variability of the actual fault diagnosis distribution, some methods (such as optimized transfer learning (TL) algorithm and regularization terms of multilayer) are aimed at solving the domain adaptation and reducing the distribution discrepancy and the among-class distance of the learned transferable features in [11, 12]. These methods can optimize the structure of feature sets better. At the same time, for getting better effectiveness in fault diagnosis under variable working conditions, some improved methods based on transfer learning (such as high-order Kullback–Leibler, parameter transfer, improved joint distribution adaptation, et al.) were also presented in [13–16]. So, it is essential to find out a new method to further optimize the structure of energy characteristics of rolling bearings in the real application. This problem is the first key core of this paper.

Additionally, on the one hand, the purpose of extracting the energy features is to implement the accurate diagnosis of fault state in industrial scenes. For this reason, the diagnosis method needs to be also simultaneously concerned while the multifeatures of fault signal are extracted. In the light of this, the support vector machine (SVM), which has the preferable ability of the classification, is usually used to implement the classification and recognition of fault in running processing of rolling bearing. However, the algorithm does not suit the situation of large amounts of data. Thus, some researchers presented improved algorithms such as binary SVM or based-HV SVM to identify the multifault types of the rolling bearing [17]. Further, the least-squares support vector machine (LSSVM) was constructed to reduce the difficulty of calculation and improve the recognition speed. The algorithm and model have solved the inequality constraint in SVM. But how does equality constraint substitute the inequality constraint is very difficult in practice. To overcome this problem, some optimized methods (such as multiclass LSSVM, trend analysis based-LSSVM, et al.) have been presented to diagnose the fault state in [18–22]. These models can better identify the fault state in a complex industrial scenario. Meanwhile, to get the better classification performance of fault states, some integrated intelligent diagnosis methods and models (such as based-SVM neural network, based-LSSVM neural network, et al.) were also established in [23]. The result showed that the improved algorithms have the classification performance for the rolling bearing in industrial systems.

However, although these improved models and algorithms may achieve the desired goal in fault diagnosis of rolling bearing, there are two crucial parameters of LSSVM worth noting, i.e., the penalty factor and the kernel function parameter. Because the penalty factor trades off between misclassification samples and interface simplicity and the kernel function defines the size of the impact of the single training sample, the accuracy of fault diagnosis is decided by them to a great extent. At present, the optimization of the two parameters has not yet been resolved. From the point of practical engineering application, there are few methods to synergistically adjust the structure of the feature to make it better for practical fault diagnosis. To overcome the problem, some optimized algorithms (such as multimode PSO, the PSO based on the Mahalanobis distance (MD), implements mutation based on PSO, et al.) was proposed to adjust significant parameters in [24–26]. So, the second area that we were focusing on in this paper is considering the interactive impact between the optimization selecting of the energy feature and the accuracy of fault diagnosis [27–31].

According to the statement, the method in this paper uses three effective methods to construct a bearing fault diagnosis model. First, wavelet transform and energy features are used to represent the characteristics of the bearing signal,d while the eight-dimension energy feature set cannot distinguish the difference between the five bearing states. Then, TCA is introduced to optimize the distribution of the energy feature set. Because the TCA can both reduce the distribution between different bearing states and increase the distance of the learned transferable features, the optimized feature set is beneficial to the improvement of diagnostic accuracy. At last, the improved PSO aims to find the optimal parameters of LSSVM.

According to these two points, a new fault diagnosis method of rolling bearing was presented by integration with cooperative energy feature extraction and improved LSSVM to extract the multidimensional feature set and enhance the accuracy of fault diagnosis. The rest of the paper is arranged as follows. In Section 2, the cooperative energy feature extraction rule has been discussed in detail combined with TCA and WP. In Section 3, we have established an improved LSSVM algorithm with dynamic parameter adjustment-based Particle Swarm Optimization (PSO) and Wavelet Mutation Optimization (WMO). In Section 4, the fault data

coming from the laboratory of the Guangdong Institute of Petrochemical Technology was used to verify the effectiveness of the model algorithm. Finally, some promising applications of the model have been discussed in detail in Section 5.

2. Cooperative Energy Feature Extraction Model and Algorithm for Vibration Signal of Rolling Bearings

In general, the extraction of the reasonable feature from original signal data is a universal method in an industrial scenario. But, as we all know, the original signal data of the bearing is large and complicated in the real industrial scene. In addition, the original data set is disturbed by complex noises. Therefore, in this situation, how to extract the energy features of the original signal data is very important to exactly represent the running state of the rolling bearing. Based on this, the Wavelet Theory and Transfer Component Analysis are introduced to construct the cooperative energy feature extraction model. The advantage of this processing method has the following two points: the first point is that the primary signal components with different frequency bands may be in detail depicted by wavelet packet because the wavelet packet may provide satisfactory localization properties in both time and frequency domains; the second point is that the structure of energy feature can be optimized by the TCA. Also, this cooperative processing method can get the internal form of the energy feature. Next, the energy feature extraction model based on wavelet packet shall first be expounded.

2.1. Energy Feature Extraction Model Based on Wavelet Packet. To address the above first problem, the vibration data may be divided into multiple frequency bands by wavelet packet. In a real application, the internal characteristics of the signal can be adaptively selected. To better understand the idea, the detailed algorithm shall be in detail described by using a wavelet packet in the next step.

To further analyze the data resource, let $L^2(R) = \oplus_{j\in Z} W_j$ indicate the fact that multiresolution analysis is based on different scale factors of j. In the multifrequency analysis, $L^2(R)$ is decomposed into a series of subspaces of the orthogonal sum of $W_j\,(j \in Z)$, where W_j is the subspace of the wavelet function. In our work, the wavelet space of W_j is refined in binary mode to achieve the goal of increasing the frequency resolution. To ensure the mapped performance between the scale-space V_j and wavelet subspace of W_j in a new subspace U_j^n , the iteration formula was defined as follows:

$$\begin{cases} U_j^0 = V_j, \\ U_j^1 = W_j, \\ j = Z, \end{cases} \tag{1}$$

where the subspace of U_j^n is the closure space of the function $\omega_n(t)$ and U_j^{2n} is the closure space of function $\omega_{2n}(t)$; the following two-scale equations should be also satisfied:

$$\begin{cases} \omega_{2n}(t) = \sqrt{2}\sum h(k)\omega_n(2t-k), \\ \omega_{2n+1}(t) = \sqrt{2}\sum g(k)\omega_n(2t-k), \end{cases} \tag{2}$$

where $g(k) = (-1)^k h(1-k)$; the sequence of $\{\omega_n(t), n \in Z\}$ is the basis function. And then the sequence constructed is determined by the basis function $\omega_0(t) = \phi(t)$ and is called the orthogonal wavelet packet; $\omega_0(t)$ and $\omega_1(t)$ are the scaling function of $\phi(t)$ and the wavelet basis function $\psi(t)$, respectively.

In addition, the normalized orthogonal basis of $L^2(R)$ is composed of $\omega_n(t-k), \omega_n(t-l) = \delta_{kl}$ and $\{\omega_n(t), n \in Z\}$. The wavelet packet series of $h(k)$ is described as $\{\omega_n(t), n \in Z\}$.

Further, for an arbitrary $c_j^n(t) \in U_j^n$, $c_j^n(t)$ can be expressed as follows:

$$c_j^n(t) = \sum_t d_l^{k,n}\omega_n\left(2^j t - l\right), \tag{3}$$

where $U_j^{2n} \perp U_j^{2n+1}, U_{j+1}^n = U_j^{2n} \oplus U_j^{2n+1}$. And then, $c_{j+1}^n(t)$ can be decomposed into $c_j^{2n}(t)$ and $c_j^{2n+1}(t)$ by using wavelet decomposition.

In addition, $\left\{d_l^{i+1,n}\right\}$ is used to obtain the equations for $\left\{d_l^{j,2n}\right\}and\left\{d_l^{j,2n+1}\right\}$ according to the following formula:

$$\begin{cases} d_l^{j,2n} = \sum_k h_{k-2l} d_k^{j+1,n}, \\ d_l^{j,2n+1} = \sum_k g_{k-2l} d_k^{j+1,n}. \end{cases} \tag{4}$$

In conventional approaches, the three-layer wavelet packet decomposition structure is shown in Figure 1.

From Figure 1, for an arbitrary signal S at which the frequency range is in$[0, f]$, it may be decomposed into a high-frequency part D_1 and a low-frequency part A_1. After the first layer in the multiresolution analysis framework, the frequency range of the high-frequency part is $[f/2, f]$, and the frequency range of the low-frequency part signal is $[0, f/2]$. Once the first layer was ended, the decomposition in the second layer starts to perform; i.e., the low-frequency part AA_2 and the high-frequency part DA_2 are obtained from decomposing the low-frequency part A_1. The high-frequency part D_1 is also decomposed to obtain the low-frequency component AD_2 and the high-frequency componentDD_2. This means that the four frequency ranges may be indicated as $[0, f/4]$, $[f/4, f/2]$, $[f/2, 3f/4]$, and $[3f/4, f]$. Analogously, the signal data set may be implemented to decompose layer by layer. The decomposition relationship for signal S may be formulated as follows:

$$\begin{aligned} S = {} & \mathrm{AAA}_3 + \mathrm{DAA}_3 + \mathrm{ADA}_3 + \mathrm{DDA}_3 + \mathrm{AAD}_3 \\ & + \mathrm{DAD}_3 + \mathrm{ADD}_3 + \mathrm{DDD}_3. \end{aligned} \tag{5}$$

Through the above algorithm, the different orthogonal wavelet spaces of U_j^n have different time–frequency resolution spaces, and all U_j^n can cover the entire bandwidth of signal S. Obviously, the time–frequency domain analysis can adaptively project the spectral components of the signal onto the

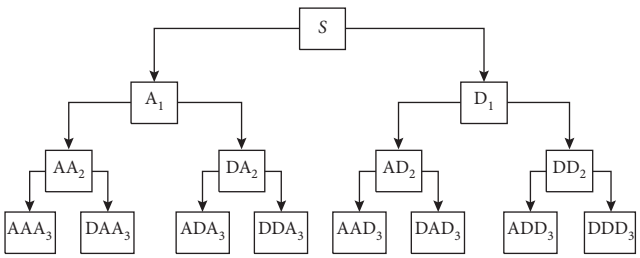


Figure 1: Schematic diagram of the wavelet packet decomposition structure.

orthogonal wavelet packet space of the corresponding frequency band. In engineering of energy feature extraction of rolling bearing, because the components of the original signal at each decomposition level represent the signal information in the corresponding local time–frequency area, the information of the component signal may be always intact. Of course, the energy of the signal distribution has been calculated at a certain decomposition level, and the energy in the orthogonal wavelet packet space at a certain decomposition level can be calculated. Then, the frequency indices of energy wavelet packets are arranged to form the eigenvectors of the original signals.

To better characterize the energy feature, suppose that the calculation formula of the wavelet packet energy is as follows:

$$E(j,n) = \sum_{k\in Z} \left[\omega_n(2t-k)\right]^2, \tag{6}$$

where $\omega_n(2t-k)$ is the wavelet packet transform coefficient.

To further understand and analyze the distribution of energy features, the statistical distribution of the energy is calculated according to the decomposed signals at different frequency bands.

Unfortunately, when the energy feature with different working conditions is input into the classifier, the result of training accuracy is 97.5%, and the test accuracy is only 87.2%. The energy characteristics cannot fully depict the differences among different states of the bearing, which results in low accuracy for bearing fault diagnosis. Thus, to find a method to make up for the shortage of wavelet packet is necessary. To reduce the data dimension and optimize the data distribution, the TCA theory was used to further optimize the feature sets in our research. Next, an improved cooperative energy feature extraction shall be established to solve the problem of combining with wavelet packet and TCA.

3. An Improved Cooperative Energy Feature Extraction Method Based on the Transfer Component Analysis Algorithm

In the real operation of extracted energy feature for signal data of rolling bearing, how to accurately distinguish the differences among the states in variable working conditions is very crucial. To ensure that the energy feature set of rolling bearing has the characteristics such as stronger class and compact inner class, the TCA was introduced to reduce the distribution discrepancy and among-class distance of the learned transferable features. The main role of TCA is to optimize the structure of energy characteristics gotten in the above section. To accommodate more flexible modeling, based on introducing the basic concept and approach of transfer feature, this section would design and implement a cooperative energy feature extraction method by using the TCA algorithm.

3.1. Basic Concepts of Transfer Feature. Notice that the TCA algorithm can adjust the edge distribution probability of the data set, and the edge distribution probability represents the probability distribution of the data set. The distribution of the bearing feature set is insufficient to meet the accuracy requirements of fault diagnosis. To reduce the distance in the same feature set and expand the gap of different feature sets, the method can reduce the distribution between the source domain and target domain data. The transfer feature mapping process is designed in Figure 2.

The circle and triangle represent source domain and target domain, respectively. A and B mean different data. Before the common mapping process is implemented, the edge probability distribution between the feature set of the source domain differs from the feature set of the target domain. The mapping relationship from the source domain to the target domain should be depicted.

For simplicity of further analysis, assume that the source domain is $D_S = \{X_S, X_T\}$, the target domain is $D_T = \{X_T\}$, X_T is the feature set of the source domain, Y_S is the label set, and X_T is the feature set of the target domain. And then, $P(X_S) \neq P(X_T)$.

In fact, after feature mapping by using the TCA algorithm, the edge probabilities of $M(X_S)$ and $M(X_T)$ are as similar as possible, and the following relationship should be satisfied:

$$P\left[M(X_S)\right] \approx P\left[M(X_T)\right]. \tag{7}$$

Once the above formula is correct, the source domain feature sample set and the target domain feature sample set are mapped to the shared subspace, and the knowledge of the feature sample transfer process can be fully utilized to improve the cross-domain learning ability.

4. Energy Feature Extraction Method Based on the Transfer Component Analysis Algorithm

To ensure that the difference between the source domain and the target domain should be reduced by finding the common points, the distance between the transfer method and retaining the original features of the two data sets is defined as follows:

$$D_S = \{X_S, X_T\} = \left\{x_{S_i}, y_{S_i}\right\}_{i=1}^{n_s}, \tag{8}$$

where n_S is the number of labeled source domain training samples data and n_t is the number of unlabeled data in the target domain for$X_T = \left\{x_{T_i}\right\}_{j=1}^{n_t}$.

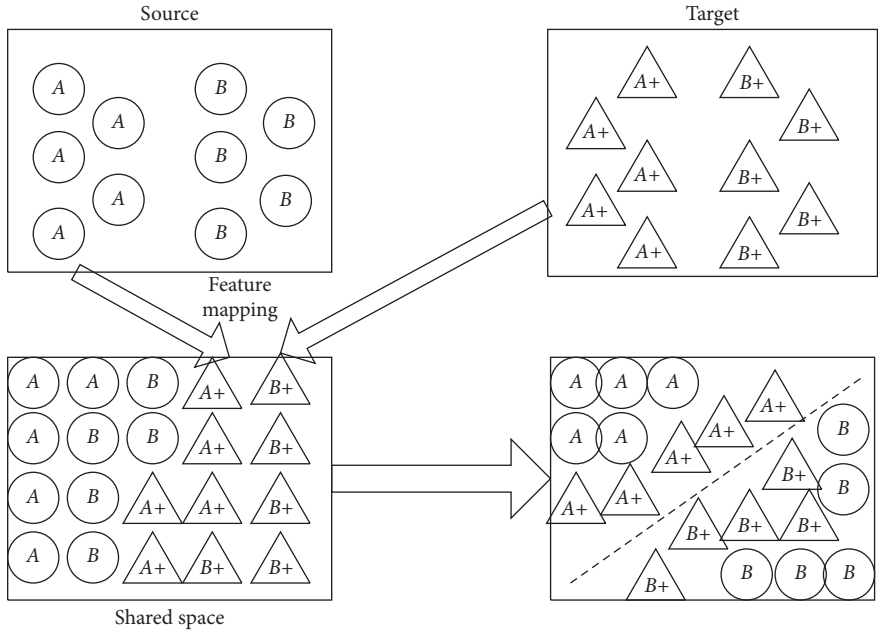


Figure 2: Transfer learning feature mapping diagram.

In this situation, the goal is to predict the sample label of y_{T_i}. At the same time, the data mapping function ϕ between the source domain and the target domain is defined as follows:

$$\begin{aligned} X_S &\longrightarrow \phi(X_S) = X_S^*, \\ X_T &\longrightarrow \phi(X_T) = X_T^*. \end{aligned} \tag{9}$$

The objective of this process is to reduce the difference between the edge probability distributions $P(X_S)$ and $P(X_T)$ so that $P(X_S^*) \approx P(X_T^*)$.

Similarly, for a given source domain data set X_S and associate target domain data set X_T, the distance function MMD between the two data sets can be expressed as follows:

$$\mathrm{MMD}(X_S, X_T) = \frac{1}{n_s}\sum_{i=1}^{n_s}\phi\left(x_{S_i}\right) - \frac{1}{n_t}\sum_{j=1}^{n}\phi\left(x_{T_j}\right)_H^2, \tag{10}$$

where $\phi(x_{T_j})_H^2$ is the squared standard operation performed in the regenerative kernel Hilbert space. The source and target domain data are mapped into a shared low-dimensional potential space through the nonlinear mapping, and then the kernel functions can be solved as follows:

$$K = \begin{pmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{pmatrix} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}. \tag{11}$$

In equation (11), $K_{S,S}$, $K_{S,T}$, $K_{T,S}$, and $K_{T,T}$ are the corresponding kernel functions obtained from the source domain, the target domain, and the hybrid domain, respectively.

Further, the formula can be rewritten as

$$\mathrm{MMD}(X_S, X_T) = Tr(\mathrm{KL}), \tag{12}$$

where Tr represents the trace of a matrix

For simplicity, $L_{i,j}$ maybe expressed as follows:

$$L_{i,j} = \begin{cases} \dfrac{1}{n_s^2}, & x_i, x_j \in X_S, \\ \dfrac{1}{n_t^2}, & x_i, x_j \in X_T, \\ -\dfrac{1}{(n_x n_t)}, & \text{others.} \end{cases} \tag{13}$$

Through the above analysis, the distribution among different data may be reduced, and shared feature representation of the two domains is realized. The representation may also maintain the data feature attributes of the two domains. Also, the method may achieve this goal and extract the data components for the transfer of data from different but related fields. The main purpose of the algorithm is twofold. First, the distance between $\phi(X_S)$ and $\phi(X_T)$ is minimized; second, the main feature attributes of the raw data sets X_s and X_T are preserved.

For the whole mapped samples, we can find an embedded matrix $W \in R^{(n_x+n_t)\times m}$ ($m \ll n_x + n_t$), s.t.

$$\widetilde{K} = \left(\mathrm{KK}^{-1/2}\widetilde{K}\right)\mathrm{KWW}^T K. \tag{14}$$

Based on equation (14), equation (11) may be rewritten as follows:

$$\mathrm{MMD}(X_S^*, X_T^*) = Tr\left(\left(\mathrm{KWW}^T K\right)L\right) = Tr\left(W^T \mathrm{KLKW}\right). \tag{15}$$

Once the covariance matrix W may be found, the largest variance of the energy feature can be maintained into the

newly created subspace. The concrete kernel matrix formula can be indicated as $\widetilde{\sum}$, i.e.,

$$\widetilde{\sum} = W^T \mathrm{KHKW}, \tag{16}$$

where $H = I - (1/(n_s + n_t)1^T) \in R^{(n_s+n_t)\times(n_s+n_t)}$ indicates the center matrix.

Therefore, the problem may be transformed into the optimal problem of $\widetilde{\sum} = I_m$, and $I_m \in R^{m\times m}$ is a unit matrix. The final core learning problem can be established as follows:

$$\begin{aligned} &\min_W Tr\left(W^T \mathrm{KLKW}\right) + \mu Tr\left(W^T W\right), \\ &\text{s.t.} \quad W^T \mathrm{KHKW} = I_m, \end{aligned} \tag{17}$$

where μ is the trade-off parameters and $\mu > 0$.

Next, the optimization problem can be transformed into a maximum mapping matrix W, which can be obtained by matrix decomposition. First, we need to calculate the matrix $(\mathrm{KLK} + \mu I)^{-1}\,\mathrm{KHK}$ to obtain W.

So far in this discussion, the core energy features can be selected by integration with the above model. On the other hand, because the distance between the same-state features becomes increasingly similar, the separability of energy feature becomes increasingly clear for different states. All in all, the compactness of features has been greatly improved after integrating with TCA. It is convenient to use classifiers to improve the fault diagnosis accuracy of bearings.

4.1. Design and Analysis of a Cooperative Energy Feature Extraction Algorithm. According to the above theoretical analysis, an improved cooperative energy feature extraction algorithm may be designed as follows.

(i) *Step 1*. Original signals in different working conditions of the bearing are input to the wavelet packets for three-layer decomposition.

(ii) *Step 2*. According to the signal component, the energy of every component is calculated, and the bearing feature set is constructed.

(iii) *Step 3*. The training sample set of the source domain is built based on the energy characteristics with explicit working conditions. Moreover, feature sets under unknown working conditions are constructed to collect test samples in the target field.

(iv) *Step 4*. The source domain feature set and the target domain feature set are mapped into the kernel space together. The maximum mean distance between the source domain feature samples and the target feature samples is measured in space. The calculated maximum mean distance is used as a criterion for judging the source domain data.

(v) *Step 5*. The data are input into the optimized LSSVM, training is performed with the source domain data, and the target domain data are used to test the training result. Finally, the classification results are obtained and accuracy is assessed. The detailed flowchart is shown in Figure 3.

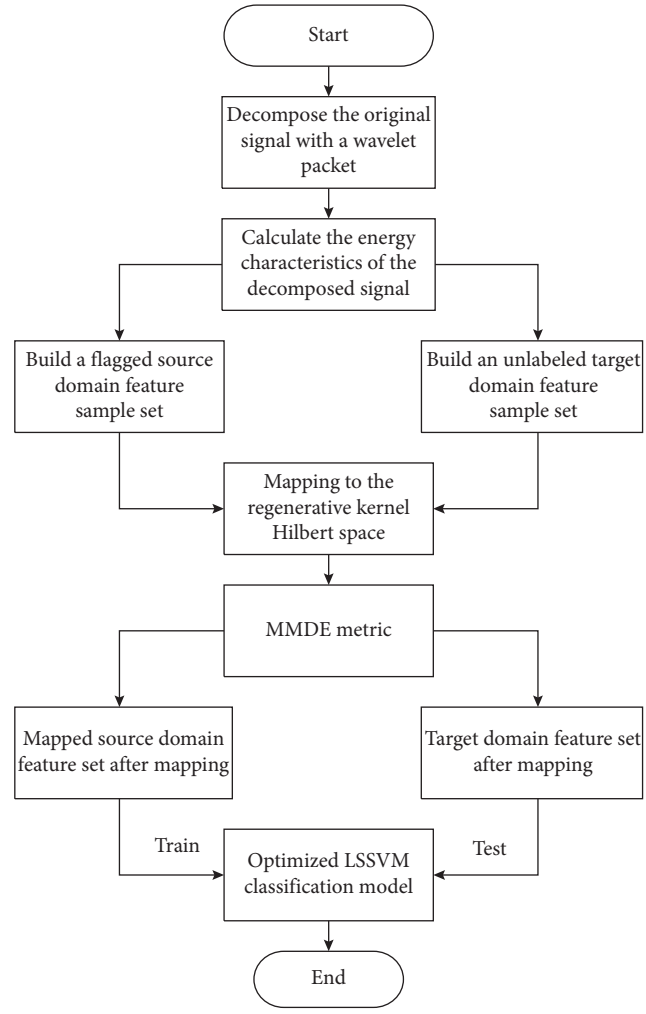


Figure 3: Flowchart of the bearing fault diagnosis method.

Whether the extraction mechanism of the energy feature is improved, the final goal of the energy feature is to improve the accuracy of the fault recognition of rolling bearing. Of course, good differentiation among different states and a high correlation among the same states will bring some gains in diagnostic accuracy. That is to say, it is also convenient to use classifiers to improve the fault diagnosis accuracy of rolling bearings. In the next step, the fault diagnosis method of rolling bearing shall be established to solve the goal.

5. Classification Process of Improved LSSVM with Dynamic Parameter Adjustment

According to the above analysis, an improved fault diagnosis method combining with improved LSSVM with dynamic parameter adjustment is listed in Figure 4.

(i) *Step 1*. Input the extracted data features into the improved LSSVM model and train the two parameters that need to be optimized.

(ii) *Step 2*. Initialize the parameter in particle swarm, such as evolutionary algebra, the learning factor, the

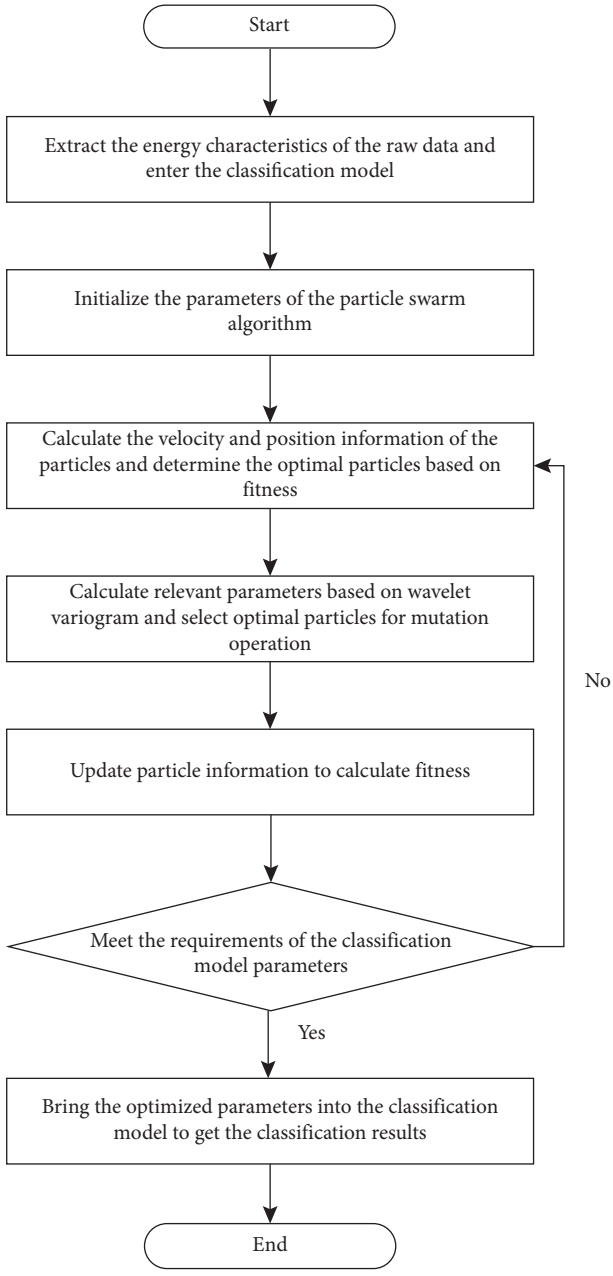


Figure 4: Schematic diagram of the optimized LSSVM classification model.

initial position x_{id} of each particle, the initial velocity v_{id}, et al.

(iii) *Step 3*. The best position is set as the initial position of each particle. The optimal fitness equals the best position of each particle. The speed and position of each particle are calculated according to the formula.

(iv) *Step 4*. Calculate the scale parameter and wavelet function value from the wavelet variogram. The mutation operation is performed on the current optimal particle according to the wavelet function formula.

(v) *Step 5*. Update p_{best} and g_{best} according to the fitness value of the particle. Then, update the velocity and position information of the particle at the same time.

(vi) *Step 6*. Determine whether the results of the algorithm reach the optimal condition. The training classification accuracy of the classification model is defined as the fitness degree of the PSO. If the fitness value calculated in the current cycle is the best, the current particle is saved as the best particle. If the fitness is not the best, the optimal parameters from the end of the previous cycle are used. The optimal particle search continues until the end of the cycle. The punishment coefficient C and Gaussian radial kernel function R are saved to construct the LSSVM classification model.

Through this algorithm, the cooperative energy feature extraction model and algorithm for the vibration signal of a rolling bearing are used to build a multidimensional feature set. And the fault diagnosis may also be implemented. The special flowchart of fault diagnosis is designed in Figure 4.

6. Experiments and Discussions

To verify the effectiveness of the proposed fault diagnosis method, the experimental acceleration data of bearings are used for fault diagnosis. The experimental data were obtained from the multifault diagnosis equipment of the rotary unit in the State Key Laboratory of Bearings, Guangdong University of Petrochemical Technology. Figure 5 shows the single-stage centrifugal fan fault diagnosis unit. Figure 6 shows the schematic diagram of the inner and outer cracks of bearings.

With this experimental platform, the data for each fault can be acquired under five states: normal, external cracking, internal cracking, missing bearings, and wearing bearings.

6.1. Signal Processing and Feature Set Construction. In our testing rig, the acceleration signals of five different conditions in the normal, external cracking, internal cracking, wearing, and missing states of the bearing during operation are used as the original signals for fault diagnosis.

The five different working conditions are in A of length 1150 mm, speed 2870 r/min; in B of length 1730 mm, speed 2980 r/min; in C of length 1800 mm, speed 2970 r/min; in D of length 2450 mm, speed 2980 r/min; and in E of length 2200 mm, speed 4800 r/min. The 300 characteristics of the first three working conditions are used as the source domain set; the 200 characteristics of the latter two working conditions are used as the target domain set. The original signals for the original states of the bearing are shown in Figure 7.

To verify the effectiveness of our algorithm, a sample data set containing 10240 sampling points in a sample period is used to extract the energy feature. To facilitate signal processing and extraction, the original signals are divided into 1024∗10 groups. Figure 7 shows the acceleration signal of the bearing under normal conditions. From Figure 7, the original data set has been divided into 1024∗10 groups. And

Figure 5: Single-stage centrifugal fan fault diagnosis unit.

the signal in each group is decomposed into 8 frequency bands by wavelet packet algorithms as shown in Figure 8.

The signal has a large volume due to high sampling frequency, and it is difficult to distinguish faults from these signals. First, the original signal is decomposed into three layers of wavelet packets to obtain signal components with eight different frequency bands. Figure 8 is a diagram showing the signal components obtained by decomposing the original signals for the original states of the bearing. The frequency band of the original signal is divided into multiple bands. According to the characteristics of the signal, the corresponding frequency band is adaptively selected to match the signal frequency, thereby improving the resolution of the signal frequency. Figure 7 only shows that the signal is decomposed into signal components of different frequency bands, and it does not reflect obvious fault characteristics. Therefore, the next step is to further extract the energy characteristics of the signal components.

After the original signal in five different states is decomposed by the wavelet packet, the characteristic histogram is obtained by calculating the energy of the node in the component signal. The energies described the multidimensional feature set of the bearing and the energy features extracted from one group of the normal signal. From the using point of view, constructing the feature set of signals is reasonable by using them in a sample period. As shown in Figure 9, the distribution of energy is different under different bearing states. The energy characteristics can initially show the difference, and then the energy feature values are extracted to construct the energy feature table. So, these features may be used to constitute a complete feature set for structure processing and fault diagnosis. Based on this, the energy feature values extracted from the 1024∗10 group of the normal signal are listed in Table 1.

In our experiment, the data sets for the five bearing states include 10∗10240 groups, and each group of signals is divided into ten groups for signal decomposition. We can obtain 8 different frequency bands from the original signal. The energy characteristics of the nodes are used to construct a multidimensional energy feature set for the bearing. Table 1 shows the multidimensional energy feature data sets for the original states; it is obvious that different bearing state has different energy features. Then, the table of energy features is input into the classifier.

The energy features extracted from the bearing fault vibration signal constitute a feature set that has been normalized. The labeled source domain data sample set and the unlabeled target domain data sample set are mapped to the regenerative Hilbert kernel space. Between the source domain and target domain, the difference in the total maximum mean value reflects the difference in the distribution. The smaller the maximum mean difference is between the source domain and target domain, the stronger the source domain to target domain mobility. It is beneficial to select source domain data with high similarity to the target domain data.

Unfortunately, when the energy feature with different working conditions is input into the classifier, the result of training accuracy is 97.5%, and the test accuracy is only 87.2%. The energy characteristics cannot fully depict the differences among different states of the bearing, which results in low accuracy for bearing fault diagnosis. Thus, to find a method to make up for the shortage of wavelet packet is necessary. To reduce the data dimension and optimize the data distribution, the TCA theory was used to further optimize the feature sets in our research. Next, an improved cooperative energy feature extraction shall be established to solve the problem of combining with wavelet packet and TCA.

Energy feature is recalculated from each component by the improved cooperative energy feature extraction algorithm. In our simulation experiment, the feature set in A is inputted into TCA which is used to optimize the distribution of the feature set. In this hidden subspace, a classifier can be trained using the tagged samples from the mapped source domain, and the classifier is used to test the target domain data in the hidden space. The simulation results are shown in Figures 10 and 11.

Figure 10 shows the original energy distributions of the bearing. The five state characteristics of the bearing (normal, outer crack, inner crack, wear, and missing steel ball in bearing) are not distinct, and a poor energy distribution leads to low classification accuracy. Obviously, after inputting the energy characteristics into the TCA algorithm, the energy distribution of the bearing is shown in Figure 11. For indeed, the distance between same-state features becomes increasingly similar and the energy features possessed the advantage of the time–space concentricity. That has shown that our model and algorithm are effective.

Whether the extraction mechanism of the energy feature is improved, the final goal of the energy feature is to improve the accuracy of the fault recognition of rolling bearing. Of course, good differentiation among different states and a high correlation among the same states will bring some gains in diagnostic accuracy. That is to say, it is also convenient to use classifiers to improve the fault diagnosis accuracy of rolling bearings. In the next step, the fault diagnosis method of rolling bearing shall be established to solve the goal.

6.2. Comparative Experimental Analysis. As a classifier, the optimized LSSVM is used for random cross-validation experiments. The data set of the source domain is used as a training set, and the data set of the target domain is used as a test set.

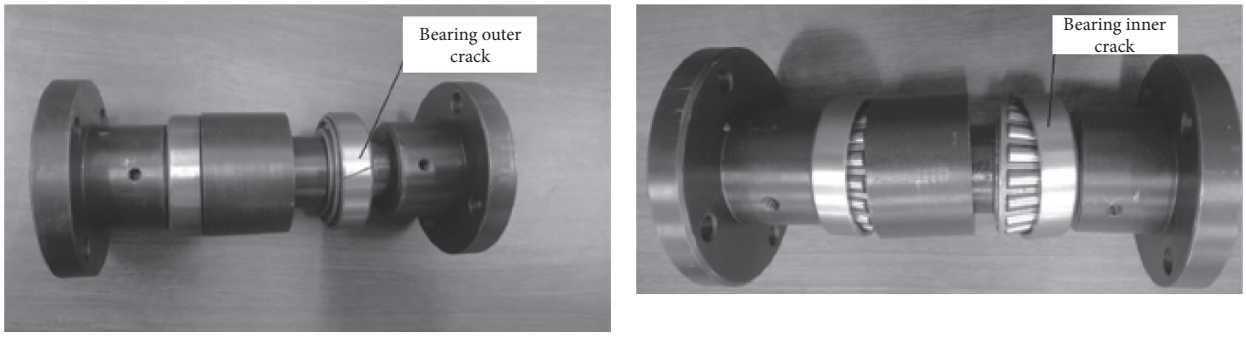


Figure 6: (a) Bearing outer crack. (b) Bearing inner crack.

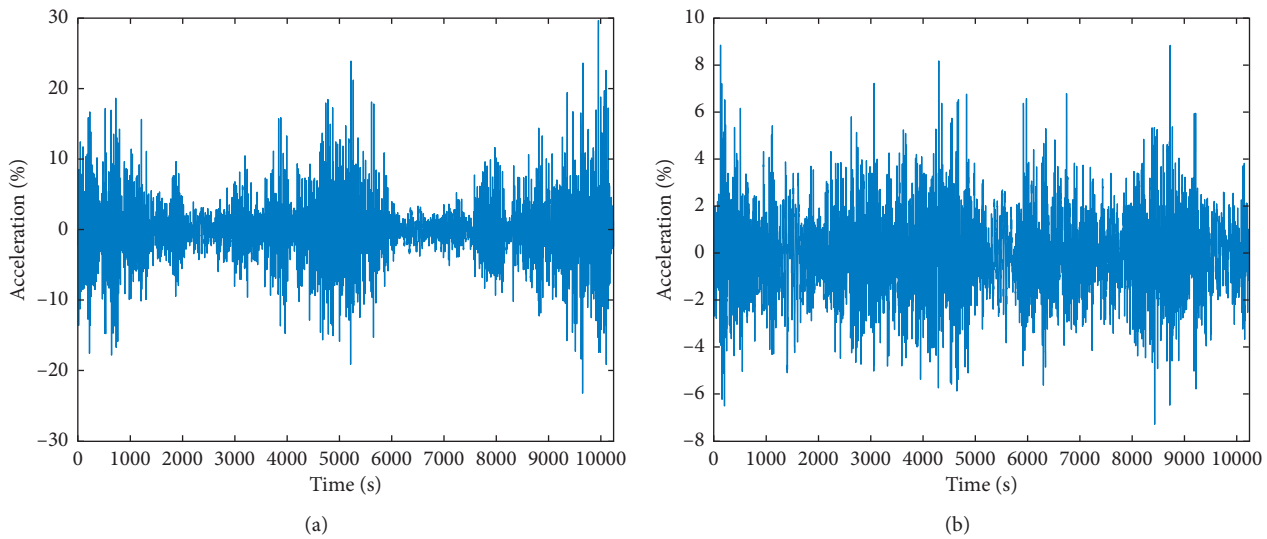


Figure 7: (a) Normal original signal diagram. (b) Outer crack original signal diagram.

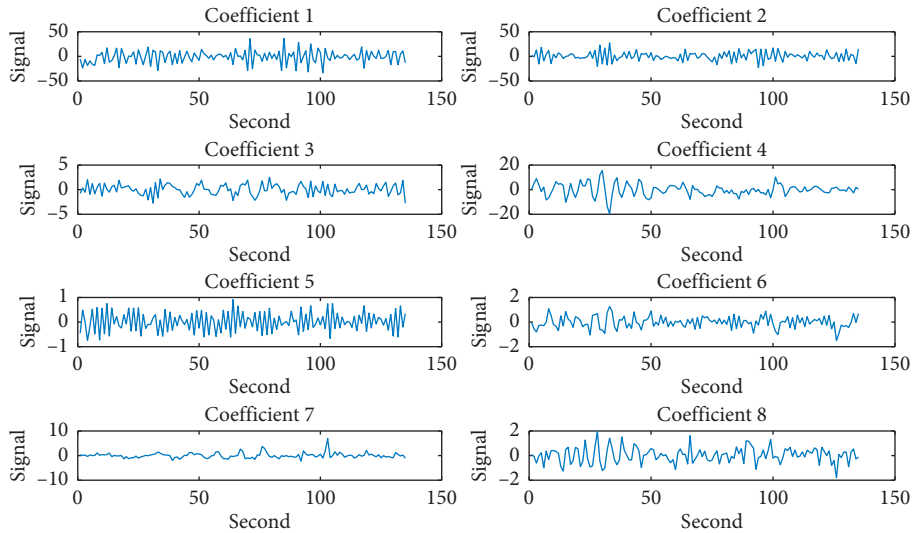


Figure 8: Normal signal component.

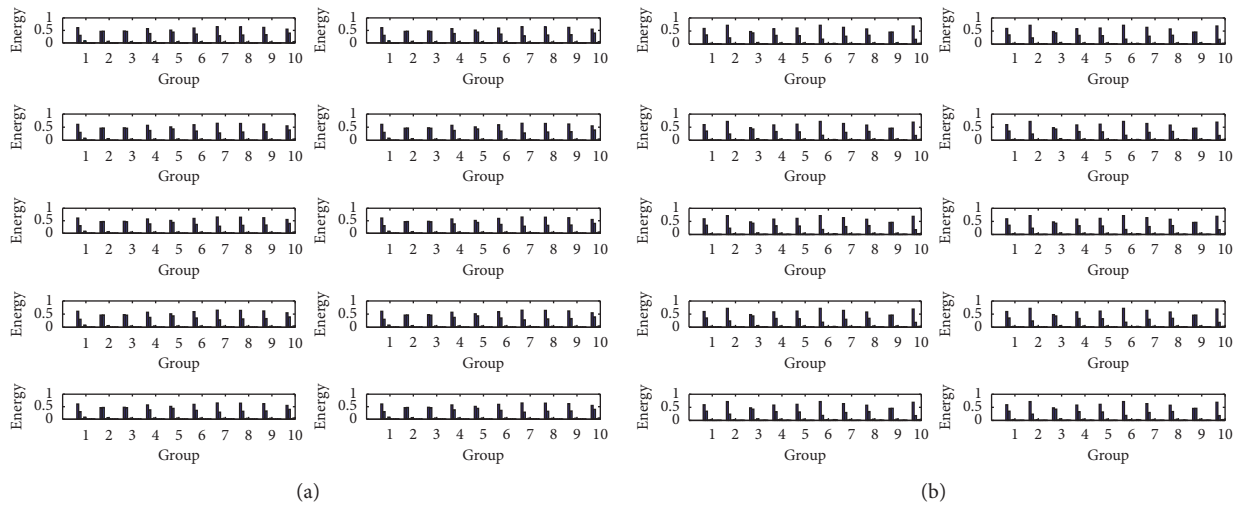


Figure 9: (a) Normal signal energy characteristics. (b) Outer crack signal energy characteristics.

Table 1: Normal signal energy characteristics.

	$E1$	$E2$	$E3$	$E4$	$E5$	$E6$	$E7$	$E8$
1	0.61	0.30	0.00	0.08	0.00	0.00	0.00	0.00
2	0.46	0.47	0.00	0.06	0.00	0.00	0.00	0.00
3	0.48	0.46	0.01	0.05	0.00	0.00	0.00	0.00
4	0.57	0.37	0.00	0.05	0.00	0.00	0.00	0.00
5	0.51	0.43	0.01	0.05	0.00	0.00	0.00	0.00
6	0.59	0.35	0.00	0.05	0.00	0.00	0.00	0.00
7	0.65	0.28	0.01	0.04	0.01	0.01	0.01	0.00
8	0.64	0.32	0.00	0.04	0.00	0.00	0.00	0.00
9	0.62	0.33	0.01	0.04	0.00	0.00	0.00	0.00
10	0.55	0.39	0.00	0.05	0.00	0.00	0.00	0.00

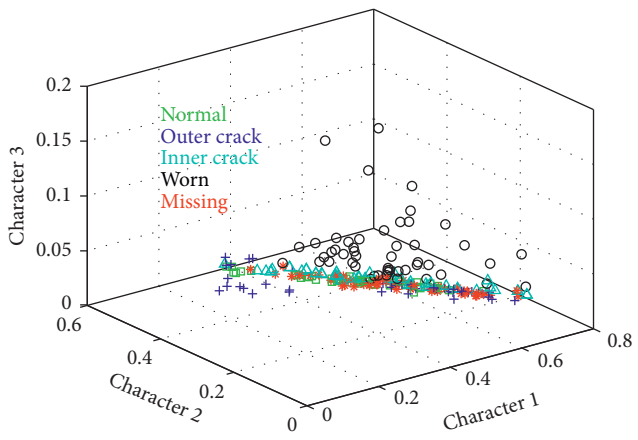


Figure 10: Original energy distributions.

In the TCA algorithm, the kernel function maps the data from the source domain and the target domain to the high-dimensional space. Therefore, the choice of the kernel function is related to the data mapping process of the source domain and the target domain. Four different kernel functions, namely, primal, RBF, linear, and SAM, are used to conduct comparative experiments. Under different kernel functions in TCA, the ability to analyze the corresponding

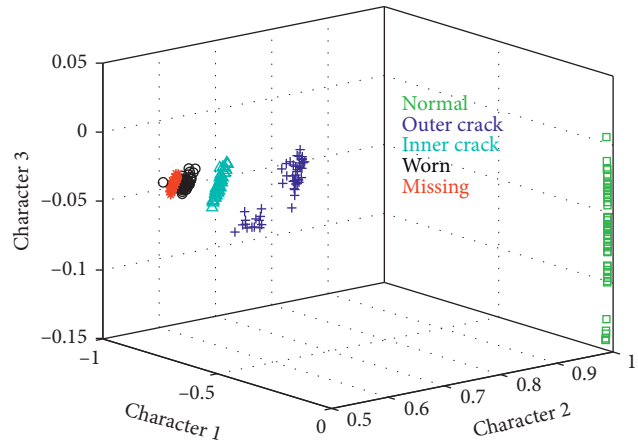


Figure 11: Energy distributions after transfer.

energy characteristics is tested. The training accuracy and test accuracy are calculated. Because TCA is a data dimensionality reduction algorithm, the dimension of data reduction is related to the classification accuracy. In this paper, the original data dimension of the energy feature data set is 8, and the dimensionality reduction is varied from 1 to 8 to test the diagnostic accuracy of the fault diagnosis method. Combining the results from Table 2 and Figure 12, the diagnostic accuracy of the RBF kernel function is relatively high and stable. Therefore, the RBF kernel function is used for bearing fault diagnosis analysis.

Based on the above fault diagnosis classification model, each group uses 100 sets of data features for fault identification. The simulation results for the training phase and the test phase are shown in Figure 12(a).

As shown in Figure 13(a), there are 210 training values for the 5 states ((1) normal, (2) internal cracking, (3) outer cracking, (4) wear, and (5) missing). The training accuracy is 100%.

According to the test data in Figure 13(b), there are 120 groups of test data for the 5 states ((1) normal, (2) internal cracking, (3) outer cracking, (4) wear, and (5) missing).

Table 2: Fault diagnosis accuracy under different kernel functions.

Different kernel functions	Training accuracy	Test accuracy	Statistical accuracy (%)
Primal	100	98.4	98.7
Linear	99.8	97.6	98
RBF	100	99.6	99
SAM	99.6	96.8	97.6

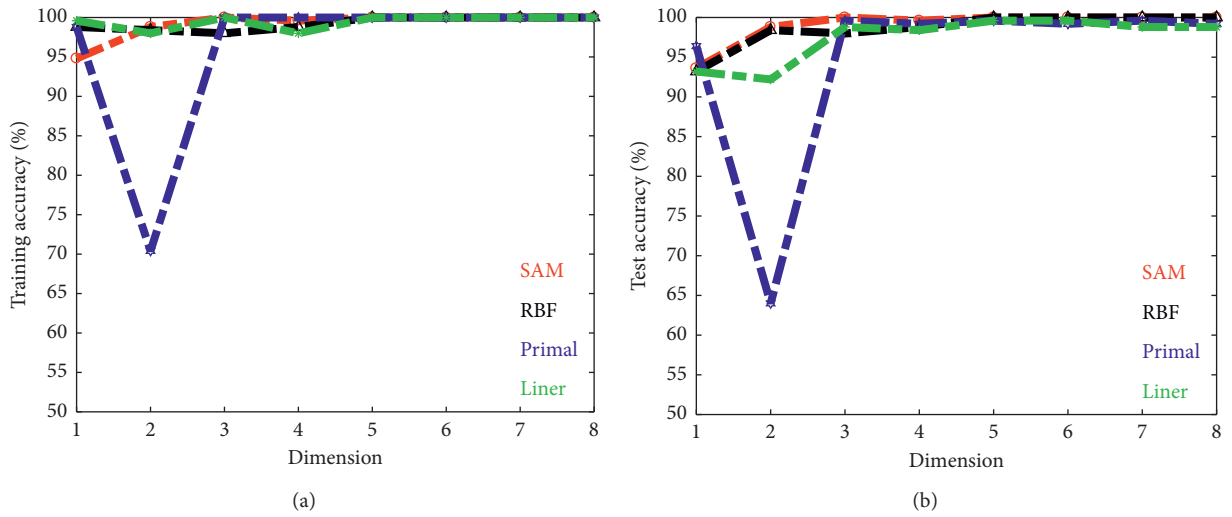


Figure 12: (a) Relationship between the feature dimension and training accuracy after mapping. (b) Relationship between the feature dimension and test accuracy after mapping.

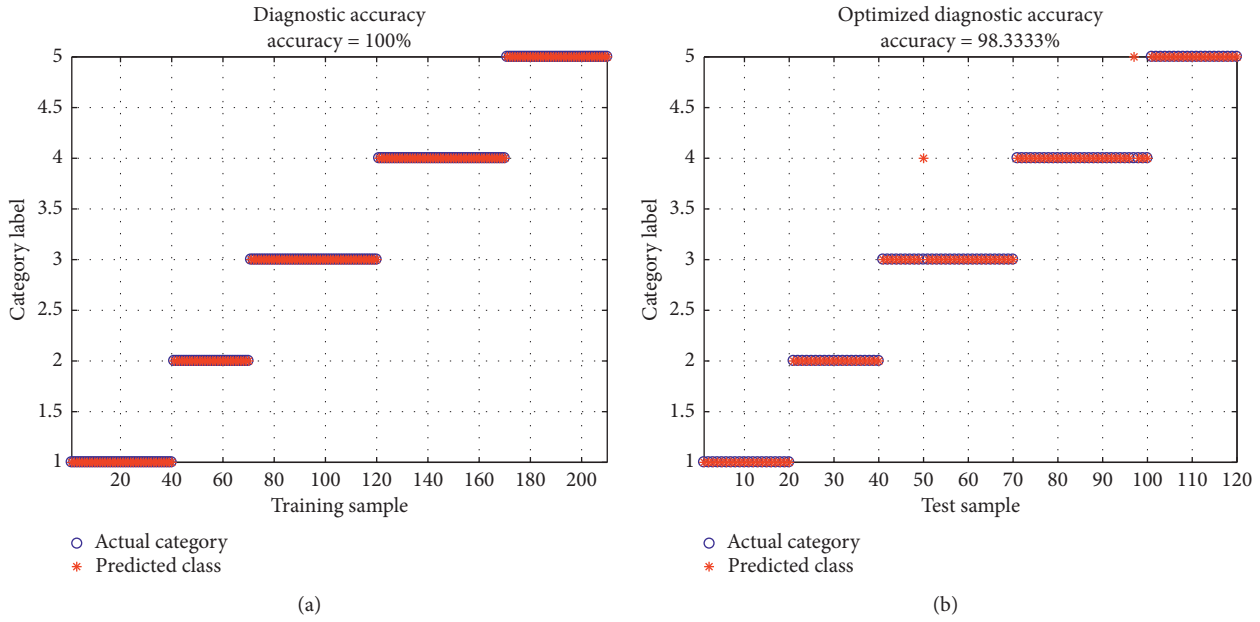


Figure 13: (a) Training accuracy. (b) Test accuracy.

Among them, two values are incorrectly classified, so the test accuracy is 98.3%.

To verify the validity and superiority of the algorithm presented in this paper, we compare different unoptimized algorithms with the optimized classification algorithm proposed.

Four different methods are compared under the same experimental environment and the same experimental data. Table 3 shows that the correct rate can reach 100% during the training process using the method developed in this paper.

Additionally, the correct rate can reach 99.8% during the test process. The fault diagnosis accuracy is better

Table 3: Accuracy of fault diagnosis in the comparative experiments.

Diagnosis method	Training accuracy (%)	Test accuracy (%)
The method proposed in this paper	100	99.8
EN-TCA-PSO-LSSVM	100	97.5
EN-PSO-LSSVM	98.8	88.4
EN-LSSVM	97.5	87.2

than that of the other three methods. The comparison shows that the TCA algorithm is effective in analyzing the energy characteristics of wavelet packets. Moreover, the optimized classification algorithm is superior to the traditional single classification algorithm and has a better diagnostic ability.

7. Conclusions

In this paper, to improve the accuracy of identifying and classifying fault in variable working conditions, a new method based on optimization of multidimension fault energy characteristics and integrate with an improved least-squares support vector machine (LSSVM). The main contributions of this paper are as follows.

(1) The method of wavelet packet is used to reduce the surrounding noise and decompose the original signal with eight different frequency bands. The energy of every component is calculated to construct a feature set for bearing.

(2) Because the TCA can amend the distribution of the energy feature, the The distribution of the feature set is optimized, and the data dimension is much closer than before. The optimized feature structure could improve the accuracy of bearing fault diagnosis.

(3) Particle swarm and the wavelet mutation were integrated to optimize two parameters of LSSVM. Through the real data of bearing, the training accuracy of the proposed method is 100%, and the test accuracy 99.8%. The experiment result shows that the proposed method is effective in the low-precision problem of fault diagnosis for complex bearings in the equipment.

(4) Unfortunately, there are still two problems to be solved in the next research. First, the complex noise of the original signal brings interference to the fault diagnosis of bearing. Second, the kernel function selected in the TCA algorithm is very single. Therefore, the next step is to focus on signal denoising and TCA construction of multicore kernel functions to further improve the fault accuracy.

Data Availability

The data used to support the findings of this study are included within the article. The data sets are provided by the Guangdong University of Petrochemical Technology for experimental verification.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank the Guangdong University of Petrochemical Technology for providing datasets for experimental verification. This work was supported by the National Natural Science Foundation of China under Grant no. 61304104; Chongqing Technology Innovation and Application Special Key Project under Grant no. cstc2019jscx-mbdxX0015; the Innovation Foundation of Chongqing Postgraduate Education under Grant no. CYS20282; and the Science and Technology Project of Power Science Research Institute of State Grid Xinjiang Electric Power Co., Ltd., under Grant no. SGXJDK00JLJS1800161.

References

[1] R. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 121–131, 2010.

[2] C. Shen, D. Wang, F. Kong, and P. W. Tse, "Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier," *Measurement*, vol. 46, no. 4, pp. 1551–1564, 2013.

[3] J. Qu, C. Shi, F. Ding, and W. Wang, "A novel aging state recognition method of a viscoelastic sandwich structure based on permutation entropy of dual-tree complex wavelet packet transform and generalized chebyshev support vector machine," *Structural Health Monitoring*, vol. 19, no. 1, pp. 156–172, 2020.

[4] Z. Ling, D. Jing, D. Huang, M. Bo, K. Lan, and Y. Liu, "The incipient fault feature enhancement method of the gear box based on the wavelet packet and the minimum entropy deconvolution," *Systems Science & Control Engineering*, vol. 6, no. 3, pp. 235–241, 2018.

[5] X. Li, S. Wu, X. Li, H. Yuan, and D. Zhao, "Particle swarm optimization-support vector machine model for machinery fault diagnoses in high-voltage circuit breakers," *Chinese Journal of Mechanical Engineering*, vol. 33, no. 1, pp. 1–10, 2020.

[6] H. Shao, J. Cheng, H. Jiang, Y. Yang, and Z. Wu, "Enhanced deep gated recurrent unit and complex wavelet packet energy moment entropy for early fault prognosis of bearing," *Knowledge-Based Systems*, vol. 188, p. 105022, 2020.

[7] A. C. Jahagirdar and K. K. Gupta, "Fractional envelope to enhance spectral features of rolling element bearing faults," *Journal of Mechanical Science and Technology*, vol. 34, no. 2, pp. 573–579, 2020.

[8] H. Jin, A. Titus, Y. Liu, Y. Wang, and a. Z. Han, "Fault diagnosis of rotary parts of a heavy-duty horizontal lathe based on wavelet packet transform and support vector machine," *Sensors*, vol. 19, no. 19, p. 4069, 2019.

[9] S. Gao, Y. Wu, and Z. Jiang, "Static and dynamic rubbing positions identification of cryocooler based on wavelet packet analysis and support vector machine," *Journal of Infrared and Millimeter Waves*, vol. 38, no. 5, pp. 627–632, 2019.

[10] S. Pan, I. Tsang, J. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks a Publication of the IEEE Neural Networks Council*, vol. 22, no. 2, pp. 199–210, 2010.

[11] S. Zhou, S. Qian, W. Chang, Y. Xiao, and Y. Cheng, "A novel bearing multi-fault diagnosis approach based on weighted permutation entropy and an improved SVM ensemble classifier," *Sensors*, vol. 18, no. 6, p. 1934, 2018.

[12] Y. Zhang, Y. Qin, Z.-y. Xing, L.-m. Jia, and X.-q. Cheng, "Roller bearing safety region estimation and state identification based on LMD-PCA-LSSVM," *Measurement*, vol. 46, no. 3, pp. 1315–1324, 2013.

[13] X. Gu, F. Deng, X. Gao, and R. Zhou, "An improved sensor fault diagnosis scheme based on TA-LSSVM and ECOC-SVM," *Journal of Systems Science and Complexity*, vol. 31, no. 2, pp. 372–384, 2018.

[14] X. Jiang, S. Li, and Y. Wang, "A novel method for self-adaptive feature extraction using scaling crossover characteristics of signals and combining with LS-SVM for multi-fault diagnosis of gearbox," *Journal of Vibroengineering*, vol. 17, no. 4, pp. 1861–1878, 2015.

[15] W. Fu, K. Wang, C. Zhang, and J. Tan, "A hybrid approach for measuring the vibrational trend of hydroelectric unit with enhanced multi-scale chaotic series analysis and optimized least squares support vector machine," *Transactions of the Institute of Measurement and Control*, vol. 41, no. 15, pp. 4436–4449, 2019.

[16] A. Moosavian, S. Jafari, M. Khazaee, and H. Ahmadi, "A comparison between ANN, SVM and least squares SVM: application in multi-fault diagnosis of rolling element bearing," *International Journal of Acoustics & Vibration*, vol. 23, no. 4, pp. 432–440, 2018.

[17] M. Sun, H. Wang, P. Liu, S. Huang, and P. Fan, "A sparse stacked denoising autoencoder with optimized transfer learning applied to the fault diagnosis of rolling bearings," *Measurement*, vol. 46, no. 4, pp. 1551–1564, 2019.

[18] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mechanical Systems and Signal Processing*, vol. 122, pp. 692–706, 2019.

[19] W. Qian, S. Li, and J. Wang, "A new transfer learning method and its application on rotating machine fault diagnosis under variant working conditions," *IEEE ACCESS*, vol. 6, pp. 69907–69917, 2019.

[20] H. Kim and B. D. Youn, "A new parameter repurposing method for parameter transfer with small dataset and its application in fault diagnosis of rolling element bearings," *IEEE ACCESS*, vol. 7, pp. 46917–46930, 2019.

[21] W. Qian, S. Li, P. Yi, and K. Zhang, "A novel transfer learning method for robust fault diagnosis of rotating machines under variable working conditions," *Measurement*, vol. 138, pp. 514–525, 2019.

[22] J. Wang, L. Qiao, Y. Ye, and Y. Chen, "Fractional envelope analysis for rolling element bearing weak fault feature extraction," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 2, pp. 353–360, 2016.

[23] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2019.

[24] Y. Zhang, P. Zhou, and G. Cui, "Multi-model based PSO method for burden distribution matrix optimization with expected burden distribution output behaviors," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1506–1512, 2019.

[25] B. Long, W. Xian, M. Li, and H. Wang, "Improved diagnostics for the incipient faults in analog circuits using LSSVM based on PSO algorithm with Mahalanobis distance," *Neurocomputing*, vol. 133, pp. 174–278, 2014.

[26] Z. Lv, L. Wang, Z. Han, J. Zhao, and W. Wang, "Surrogate-assisted particle swarm optimization algorithm with pareto active learning for expensive multi-objective optimization," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 838–849, 2019.

[27] K. Zhang, K. Peng, S. X. Ding, Z. Chen, and X. Yang, "A correlation-based distributed fault detection method and its application to a hot tandem rolling mill process," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 3, pp. 2380–2390, 2020.

[28] K. Zhang, H. Hao, Z. Chen, S. X. Ding, and K. Peng, "A comparison and evaluation of key performance indicator-based multivariate statistics process monitoring approaches," *Journal of Process Control*, vol. 33, pp. 112–126, 2015.

[29] H. Luo, S. Yin, T. Liu, and A. Q. Khan, "A data-driven realization of the control-performance-oriented process monitoring system," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 1, pp. 521–530, 2020.

[30] H. Luo, X. Yang, M. Krueger, S. X. Ding, and K. Peng, "A plug-and-play monitoring and control architecture for disturbance compensation in rolling mills," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 200–210, 2020.

[31] H. Luo, K. Li, O. Kaynak, S. Yin, M. Huo, and H. Zhao, "A robust data-driven fault detection approach for rolling mills with unknown roll eccentricity," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2641–2648, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5303658, 23 pages
https://doi.org/10.1155/2020/5303658

Research Article

An Integrated Health Condition Detection Method for Rotating Machinery Using Refined Composite Multivariate Multiscale Amplitude-Aware Permutation Entropy

Fuming Zhou,[1] **Wuqiang Liu**,[1] **Ke Feng**,[1] **Jinxing Shen**,[1] **and Peiping Gong**[2]

[1]*Field Engineering College of Army Engineering University, Nanjing 210007, China*
[2]*Training Base of Army Engineering University, Xuzhou 210004, China*

Correspondence should be addressed to Ke Feng; 2020659722@qq.com and Jinxing Shen; 565423803@qq.com

Received 20 August 2020; Revised 26 November 2020; Accepted 1 December 2020; Published 22 December 2020

Academic Editor: Yong Chen

With a view to realizing the fault diagnosis of rotating machinery effectively, an integrated health condition detection approach for rotating machinery based on refined composite multivariate multiscale amplitude-aware permutation entropy (RCmvMAAPE), max-relevance and min-redundancy (mRmR), and whale optimization algorithm-based kernel extreme learning machine (WOA-KELM) is presented in this paper. The approach contains two crucial parts: health detection and fault recognition. In health detection stage, multivariate amplitude-aware permutation entropy (mvAAPE) is proposed to detect whether there is a fault in rotating machinery. Afterward, if it is detected that there is a fault, RCmvMAAPE is employed to extract the initial fault features that represent the fault states from the multivariate vibration signals. Based on the multivariate expansion and multiscale expansion of amplitude-aware permutation entropy, RCmvMAAPE enjoys the ability to effectively extract state information on multiple scales from multichannel series, thereby overcoming the defect of information loss in traditional methods. Then, mRmR is adopted to screen the sensitive features so as to form sensitive feature vectors, which are input into the WOA-KELM classifier for fault classification. Two typical rotating machinery cases are conducted to prove the effectiveness of the raised approach. The experimental results demonstrate that mvAAPE shows excellent performance in fault detection and can effectively detect the fault of rotating machinery. Meanwhile, the feature extraction method based on RCmvMAAPE and mRmR, as well as the classifier based on WOA-KELM, shows superior performance in feature extraction and fault recognition, respectively. Compared with other fault identification methods, the raised method enjoys better performance and the average fault recognition accuracy of the two typical cases in this paper can all reach above 98%.

1. Introduction

As one of the widely applied mechanical equipment, rotating machinery plays a vital role in industrial production. Nevertheless, it usually operates in harsh environments such as heavy load and high speed, which greatly increases the risk of faults. These faults may result in equipment shutdown and even casualties cause if they are not dealt with in time [1, 2]. Due to the particularity of industrial machinery, direct disassembly overhaul will affect normal production. Hence, research on nondisassembly health condition detection technology of rotating machinery has always been a hotspot. When encountering faults, some changes will occur in the internal structure of rotating machinery, which affects the frequency and amplitude of vibration signals. It indicates that the vibration signals contain a wealth of information related to the operating states of rotating machinery [3, 4]. Consequently, analyzing vibration signals is a feasible method for fault diagnosis [5].

The essences of vibration signals-based fault diagnosis are the fault feature extraction and pattern recognition issues. Among which, how to extract the features which can represent the working states from the vibration signals is the key in fault diagnosis. In the past decades, time-frequency analysis is widely applied in feature extraction of vibration signals. Many time-frequency analysis methods such as

empirical mode decomposition (EMD) [6], local mean decomposition (LMD) [7], wavelet packet transform (WPT) [8], and variational mode decomposition (VMD) [9] are applied to fault diagnosis of rotating machinery. Unfortunately, the vibration signals of rotating machinery usually exhibit nonlinear and nonstationary characteristics, which cause the above methods to have some defects in practical applications. For instance, WPT needs to choose the suitable wavelet kernel function [8] and VMD need to set the penalty factor α and the number of intrinsic mode functions (IMFs) K before processing the vibration signals [10], thereby the self-adaptive capacity of them is poor. EMD enjoys good adaptability, but it has defects such as mode mixing and end effect. In addition, the application of time-frequency analysis methods alone requires the operators to have a certain knowledge reserve, which limits the efficiency and application scope of these methods. Therefore, developing an efficient and accurate fault feature extraction tool is urgent and necessary.

Recently, the entropy-based theory has been widely adopted as feature extraction tool in the field of fault diagnosis due to its excellent performance in measuring the nonlinear complexity of time series [11]. Entropy methods that are commonly applied include approximate entropy (AE) [12], sample entropy (SE) [13], fuzzy entropy (FE) [14], and permutation entropy (PE) [15]. Among them, AE is highly dependent on the data length and is prone to undefined entropy value. SE and FE are time-consuming, so they are not suitable for processing signals with a large amount of data, while PE is favored by many scholars because of its high computational efficiency and strong antinoise ability. Zhang et al. [16] adopted PE to detect bearing faults and proposed a bearing fault diagnosis model based on PE, ensemble empirical mode decomposition, and optimized SVM. Kuai et al. [17] proposed a fault diagnosis method for planetary gears based on PE, CEEMDAN, and ANFIS.CEEMDAN is applied to decompose the vibration signal of planetary gears, and PE is used to extract the characteristics of the obtained IMFs. Finally, ANFIS is used as a classifier to complete fault identification. Nevertheless, PE also exists some inherent defects. For example, it loses sight of the influence of amplitude information of signals on the entropy value, which may lose the crucial information. To address this problem, Azami et al. [18] presented the amplitude-aware permutation entropy (AAPE), which is not only sensitive to the frequency but also sensitive to the amplitude of signals. The excellent performance of AAPE has been verified through the simulation and biological signals experiments.

However, AAPE also possesses some shortcomings that cannot be ignored. Firstly, AAPE only measures the complexity of the measured signal on one temporal scale, thereby cannot capture the long correlation of the signal [19]. To address this question, based on multiscale entropy theory [19], multiscale amplitude-aware permutation entropy (MAAPE) was proposed to extract the fault information of rolling bearings [20]. Unfortunately, MAAPE enjoys poor stability, especially for short-time series. The defect will cause MAPPE to produce unreliable entropy values on high scales. Secondly, AAPE cannot extract fault features from multichannel vibration signals, which limits its ability to extract fault information for large equipment. For large equipment, the long transmission path will reduce the vibration impulse to a certain extent. In other words, the fault information will be lost. Therefore, the vibration signal collected by single channel is usually not enough to provide enough fault information to identify the fault type [21]. It is necessary to improve AAPE so that it can extract fault features from multichannel vibration signals synchronously.

With a view to solving the aforementioned defects, refined composite multivariate multiscale amplitude-aware permutation entropy (RCmvMAAPE) is presented in this paper. Compared with the existing AAPE methods, the proposed RCmvMAAPE possesses two main improvements. Firstly, refined composite multiscale method is employed to substitute the traditional multiscale method in MAAPE to overcome the entropy instability problem [22]. In addition, on the basis of multidimensional embedding reconstruction theory [23], AAPE is expanded to multivariate AAPE (mvAAPE) to measure the complexity of multichannel vibration signals. Based on the above improvements, RCmvMAAPE overcomes the abovementioned defects and can stably measure the complexity of multichannel signals on multiple scales. The performance of RCmvMAAPE is comprehensively tested utilizing a variety of synthetic signals in this paper, and the results indicate that RCmvMAAPE can availably measure the complexity of multivariate signals. In view of the advantages of RCmvMAAPE, this paper employs it to extract the fault features of multichannel vibration signals of rotating machinery.

As we know, the fault features distributed on multiple scales extracted by RCmvMAAPE are a high-dimensional feature vector. Among which, some sensitive features can effectively represent the fault information, but some redundant features not only affect the accuracy of subsequent fault classification but also reduce the diagnosis efficiency. For this reason, it is necessary to compress the high-dimensional fault features to improve the fault recognition rate. The max-relevance and min-redundancy (mRmR) is a typical features selection method based on spatial search, which uses mutual information to measure the relevance and redundancy of features [24]. The maximum correlation indicates that the feature has a large correlation with the sample category, that is, it can reflect the sample category information to the greatest extent. Minimal redundancy means that the correlation between features is the smallest, that is, the redundancy of features is the smallest. This paper adopts mRmR to select the sensitive features to form sensitive features vectors that represent the fault state of rotating machinery.

Afterward, different fault states of rotating machinery will be identified according to the sensitive feature vectors, namely, pattern recognition. At this stage, a classifier with high computational efficiency and good generalization performance is needed. Kernel extreme learning machine (KELM) [25] is a machine learning method that combines ELM and kernel function. While retaining the high calculation efficiency of ELM, the introduction of kernel function

enables KELM to enjoy stronger generalization ability compared with commonly used classifiers such as BP neural network (BP) [26], support vector machine (SVM) [27], and extreme learning machine (ELM) [28] when dealing with linear inseparable problems; meanwhile, KELM is sensitive to parameter setting due to the existence of kernel function. To choose the best parameters, we need to employ a suitable optimization algorithm to determine the best parameters of KELM. Commonly used optimization algorithms consist of particle swarm optimization (PSO) [29], ant colony optimization (ACO) [30], and whale optimization algorithm (WOA) [31]. Among which, WOA has attracted more and more attention due to its uncomplicated operation, less adjustment parameters, and strong capability to jump out of local optimum. Therefore, WOA is utilized to iteratively select the optimal parameter of KELM to build a classifier based on WOA-KELM. The low-dimensional sensitive feature vectors are input into WOA-KELM so as to judge the fault type of the rotating machinery.

Consequently, a new integrated health detection method for rotating machinery is proposed, which includes two parts: fault detection and fault identification. In the fault detection stage, mvAAPE is employed to extract the features of the vibration signals to determine whether the rotating machinery is malfunctioning. By introducing the key link of fault detection, the unnecessary disassembly and maintenance of the equipment can be avoided, and the damage to the equipment can be reduced. In the fault identification stage, the presented method based on RCmvMAAPE, mRmR, and WOA-KELM is applied to diagnose different fault types and fault severity of rotating machinery. Two examples are conducted to prove the performance of the proposed method and its superiority compared to other existing methods.

The rest of the paper is arranged as follows: in Sections 2 and 3, the basic theory of RCmvMAAPE and WOA-KELM is introduced in detail; Section 4 displays the steps of the proposed approach; two typical cases are adopted for experiments to verify the excellent performance of the proposed approach in Section 5; finally, this paper is summarized in Section 6.

2. The Basic Theory of RCmvMAAPE

2.1. Multivariate Amplitude-Aware Permutation Entropy

2.1.1. AAPE. AAPE is a method based on PE, which is a powerful tool for analyzing nonlinear time series. Therefore, it is necessary to introduce the concept of PE firstly. The original theory of PE is reviewed in [15].

For a given time series $X = \{x_i\}$, $i = 1, 2, \ldots, N$, at any time point t, the m dimensional reconstruction vector can be obtained as

$$X_t^{m,d} = \left[x_t, x_{t+d}, \ldots, x_{t+(m-2)d}, x_{t+(m-1)d}\right],\ t = 1, 2, \ldots, N-(m-1)d, \tag{1}$$

where m denotes the embedding dimension and d denotes the time delay.

For each reconstruction vector, in accordance with the size of the elements in ascending order, the permutation $\pi_{r_0, r_1, \ldots, r_{m-1}}$ can be acquired, which fulfills that

$$\left[x_{t+(j_1-1)d}, x_{t+(j_2-1)d}, \ldots, x_{t+(j_{m-1}-2)d}, x_{t+(j_m-1)d}\right], \tag{2}$$

where j_* represents the index of the column of each element in the reconstructed component. Accordingly, there are $m!$ possible permutation patterns, of which the i-th permutation is marked as π_i.

The relative frequency of π_i can be expressed as

$$p(\pi_i) = \frac{g(\pi_i)}{N-(m-1)d}, \tag{3}$$

where $g(\pi_i)$ represents the function that counts the number of π_i in $X_t^{m,d}$. The value of $g(\pi_i)$ will increase by 1 if the permutation order of the internal elements of $X_t^{m,d}$ is π_i.

Consequently, based on the calculation theorem of Shannon entropy, PE can be defined as

$$\mathrm{PE}(X, m, d) = -\sum_{i=1}^{m!} p(\pi_i) \ln p(\pi_i). \tag{4}$$

Nevertheless, PE enjoys some nonnegligible deficiencies, which led to its inability in describing the irregularity of the series. Firstly, from the theoretical point of view, the original PE algorithm only considers the effect of the ordinal structure of the time series on the entropy value, but the amplitude information of each mapped element in the series is ignored. Secondly, when there are elements with equal amplitude, their influence on the entropy value cannot be accurately estimated. In view of the aforementioned defects of PE, Azami proposed AAPE to significantly enhance the performance of PE [18]. The basic principle of the AAPE algorithm is as follows:

Supposing that the starting value of $p(\pi_i)$ is 0, for the reconstruction vector $X_t^{m,d}$, when the time t adds from 1 to $N-m+1$, the value of $p(\pi_i)$ is updated whenever the permutation is π_i.

$$p^{\mathrm{update}}(\pi_i) = p(\pi_i) + \left(\frac{\alpha}{m}\sum_{k=1}^{m}\left|x_{t+(k-1)d}\right| + \frac{1-\alpha}{m-1}\sum_{k=2}^{d}\left|x_{t+(k-1)d} - x_{t+(k-2)d}\right|\right), \tag{5}$$

where $\alpha \in [0, 1]$ denotes the adjustment coefficient which is utilized to adjust the weight of the time series amplitude average and the deviation between the amplitudes. Thus, the probability of $p(\pi_i)$ is

$$p(\pi_i) = \frac{p^{\text{update}}(\pi_i)}{\sum_{t=1}^{N-m+1}\left((\alpha/m)\sum_{k=1}^{m}\left|x_{t+(k-1)d}\right| + (1-\alpha/m-1)\sum_{k=2}^{m}\left|x_{t+(k-1)d} - x_{t+(k-2)d}\right|\right)}. \tag{6}$$

The AAPE of time series x can be defined as

$$\text{AAPE}(X, m, d, \alpha) = -\sum_{i=1}^{m!} p(\pi_i)\ln p(\pi_i). \tag{7}$$

2.1.2. mvAAPE. To describe the complexity of multichannel time series, it is necessary to extend the AAPE to multivariate analysis so as to put forward multivariate amplitude-aware permutation entropy (mvAAPE). The definition of mvAAPE is described as follows:

(1) Given a p-channel series $X = \{x_{c,1}, x_{c,2}, \ldots, x_{c,i}, \ldots, x_{c,N}\}$, $c = 1, 2, \ldots, p$, phase space reconstruction is performed as follows:

$$Z_t^{m,d} = \left[x_{c,t}, x_{c,t+d}, \ldots, x_{c,t+(m-2)d}, x_{c,t+(m-1)d}\right],\ t = 1, 2, \ldots, N-(m-1)d. \tag{8}$$

(2) Arrange the reconstruction time series $Z_i^{m,d}$ in ascending order as $[x_{c,i+(j_1-1)d} \le x_{c,i+(j_2-1)d} \le \cdots \le x_{c,i+(j\ \ m-1}-1)d \le x_{c,i+(j_m-1)d}]$. At the same time, there are m! potential permutations $\pi_i, 1 \le i \le m!$.

(3) For c-th channel, supposing that the starting value of $p(\pi_{c,i})$ is 0, for the reconstruction series $Z_i^{m,d}$, when t gradually increases from 1 to $N-m+1$, the value of $p(\pi_{c,i})$ will be renewed as $\pi_{c,i}$ appears.

$$p^{\text{update}}\left(\pi_{c,i}\right) = p\left(\pi_{c,i}\right) + \left(\frac{\alpha}{m}\sum_{k=1}^{m}\left|x_{c,t+(k-1)d}\right| + \frac{1-\alpha}{m-1}\sum_{k=2}^{d}\left|x_{c,t+(k-1)d} - x_{c,t+(k-2)d}\right|\right). \tag{9}$$

(4) Calculate the relative frequency of i-th permutation in c-th channel $\pi_{c,i}$ as follows:

$$p\left(\pi_{c,i}\right) = \frac{p^{\text{update}}\left(\pi_{c,i}\right)}{\sum_{c=1}^{p}\sum_{t=1}^{N-m+1}\left((\alpha/m)\sum_{k=1}^{m}\left|x_{c,t+(k-1)d}\right| + (1-\alpha/m-1)\sum_{k=2}^{m}\left|x_{c,t+(k-1)d} - x_{c,t+(k-2)d}\right|\right)}. \tag{10}$$

For p-channel time series, $p(\pi_{c,i})$ satisfies $\sum_{c=1}^{p}\sum_{j=1}^{m!} p(\pi_{c,i}) = 1$.

(5) The probability of the i-th pattern π_i in p-channel time series X can be calculated as follows:

$$p(\pi_i) = \sum_{c=1}^{p} p\left(\pi_{c,i}\right). \tag{11}$$

(6) Based on the definition of Shannon entropy, mvAAPE is expressed as

$$mv\text{AAPE}(\text{X}, m, \alpha, d) = -\sum_{i=1}^{m!} p(\pi_i)\ln p(\pi_i), \tag{12}$$

where mvAAPE actually extends the application of AAPE from univariate analysis to multivariate analysis. However, mvAAPE only analyzes the multichannel time series on one temporal scale, while the measured time series often contains information on multiple scales. Therefore, the key information will lose if only a single scale analysis is conducted. In response to this problem, mvMAAPE that is able to analyze time series on multiple scales is proposed.

2.2. mvMAAPE. The principle of mvMAAPE is as follows:

(1) For p-channel series $U = \{u_{k,1}, u_{k,2}, \ldots, u_{k,i}, \ldots, u_{k,L}\}$, $k = 1, 2, \ldots, p$, the multivariate coarse-grained time series at scale factor τ is defined as follows:

$$y_{k,j}^{\tau} = \frac{1}{\tau}\sum_{b=(j-1)\tau+1}^{j\tau} u_{k,i},\ 1 \le j \le \frac{L}{\tau},\ 1 \le k \le p. \tag{13}$$

When $\tau > 1$, the multivariate series is divided into coarse-grained time series of length $[L/\tau]$.

(2) Calculate the mvAAPE of τ multivariate coarse-grained time series and the result is as follows:

$$mv\text{MAAPE}(\text{U}, m, \alpha, d, \tau) = mv\text{AAPE}\left(y_{k,j}^{\tau}, m, \alpha, d\right), \tag{14}$$

where mvMAAPE overcomes the shortcomings that PE does not consider the amplitude information; meanwhile, the combination with multivariate analysis improves the utilization of multichannel information, which is essentially an assessment of the irregularity of multichannel data. The

evaluation principle can be summarized into two aspects: (1) if the entropy value of the multivariate series X is greater than that of series Y on most scale factors, it can be shown that X is more random than Y and more prone to dynamic mutations. (2) If the entropy value of X decreases significantly with the increase of the scale factor, it indicates that the information included in X mainly appears on a smaller scale factor, such as a random white noise signal. mvMAAPE considers the interrelationship of each time series in multichannel data and comprehensively evaluates each dimension of multichannel series. Therefore, mvMAAPE can effectively detect the mutation change of multichannel series.

2.3. Refined Composite Multivariate Multiscale Amplitude-Aware Permutation Entropy

2.3.1. Basic Principle. The mvMAAPE realizes multivariate and multiscale analysis by extending the mvAAPE method to multiple scales, so as to obtain more useful information. However, the coarse-graining method adopted by mvMAAPE has serious defects, which leads to incomplete information analysis. For instance, the calculation of mvMAAPE only considers the coarse-graining series starting from $u_{k,1}$ and ignores the coarse-graining series such as $u_{k,2}$ at scale factor τ. However, the remaining $\tau - 1$ time series also contain the key information, and the direct neglect will lead to insufficient analysis and affect the analysis effect. Therefore, the refined composite multiscale coarse-graining approach is employed to achieve accurate and sufficient analysis. The implementation principle of the coarse-graining method is presented in Figure 1.

The Detailed Procedures of RCmvMAAPE are Described as follows:

(1) For p-channel series $U = \{u_{k,1}, u_{k,2}, \ldots, u_{k,i}, \ldots, u_{k,L}\}$, $k = 1, 2, \ldots, p$, the coarse-grained multivariate time series are computed on a given scale factor τ and the elements of the a-th coarse-grained time series $Y_a^\tau = \{y_{k,i,1}^\tau, y_{k,i,2}^\tau, \ldots\}$ are computed by

$$y_{k,i,a}^\tau = \frac{1}{\tau} \sum_{b=a+(i-1)\tau}^{a+i\tau-1} u_{k,b}, \tag{15}$$

where $1 \le i \le L/\tau, 1 \le k \le p, 1 \le a \le \tau$. For the scale factor τ, there will be τ diverse coarse-grained multivariate time series.

(2) For each coarse-grained multivariate series, the marginal relative frequencies $p(\pi_j)$ are computed. Then, the average relative frequencies $\overline{p(\pi_j)}$ can be acquired by

$$\overline{p(\pi_j)} = \frac{1}{\tau} \sum_{a=1}^{\tau} p_a(\pi_j). \tag{16}$$

(3) The RCmvMAAPE of original multivariate time series is computed as follows:

$$RCmvMAAPE(U, m, \alpha, d, \tau) = -\sum_{j=1}^{m!} \overline{p(\pi_j)} \ln \overline{p(\pi_j)}. \tag{17}$$

In the RCmvMAAPE approach, there are three key parameters, namely, the m, α, and d. For the embedding dimension m, if the value is too small, the reconstructed vector includes too few states and the algorithm will lose its validity and significance, whereas if m is too large, the phase space reconstruction will homogenize the time series, which not only increases the amount of calculation but also cannot reflect the slight change of the time series. According to references [18, 29], the AAPE for univariate analysis usually sets the embedding dimension to 3–7, and the optimal parameters of the univariate analysis method and multivariate analysis are generally consistent, so this article sets the embedding dimension to $m = 5$. The adjustment coefficient α is usually set to 0.5 according to reference [18], so this article sets $\alpha = 0.5$. Time delay has little effect on the performance of the algorithm, so in this article, $d = 1$.

2.3.2. Performance Analysis. To validate the performance of RCmvMAAPE, other multivariate analysis approaches are compared with it to reflect its advantages in extracting the complexity of multichannel signals. White Gaussian noise (WGN) and $1/f$ noise are two signals that are widely adopted to evaluate the univariate and multivariate analysis method. Compared with WGN signals, the power spectrum of $1/f$ noise is more complicated and includes more mode information. The generation of WGN is randomly distributed, so the probability of its state transition matrix appearing is approximately equal. On the contrary, $1/f$ noise is a long-range correlation signal, and the irregularity of $1/f$ noise is lower than that of WGN. Consequently, the complexity of $1/f$ noise is higher than that of WGN. Considering the universality, WGN and $1/f$ noise are employed to create a multichannel signal with three different channels to analyze RCmvMAAPE, mvMAAPE, RCmvMSE, and RCmvMPE. They are (a) three channel WGN; (b) three channel $1/f$ noise; (c) two channel WGN and one channel $1/f$ noise; and (d) two channel $1/f$ noise and one channel WGN. There are 25 groups (length 2048) of the synthesized signals in each case.

For sake of verifying the advantages of the proposed approach in measuring the complexity of multivariate signals, RCmvMAAPE, mvMAAPE, RCmvMPE, and RCmvMSE of four kinds of multivariate synthetic signals are calculated. The mean standard deviation diagrams of the four methods are shown in Figure 2. Compared with mvMAAPE, RCmvMPE, and RCmvMSE, the standard deviation of RCmvMAAPE is significantly smaller than mvMAAPE and RCmvMSE, which indicates that the stability and robustness of RCmvMAAPE are stronger than mvMAAPE and RCmvMSE. It can be clearly seen from the figure that RCmvMAAPE can effectively separate four multivariate synthetic signals, proving that RCmvMAAPE

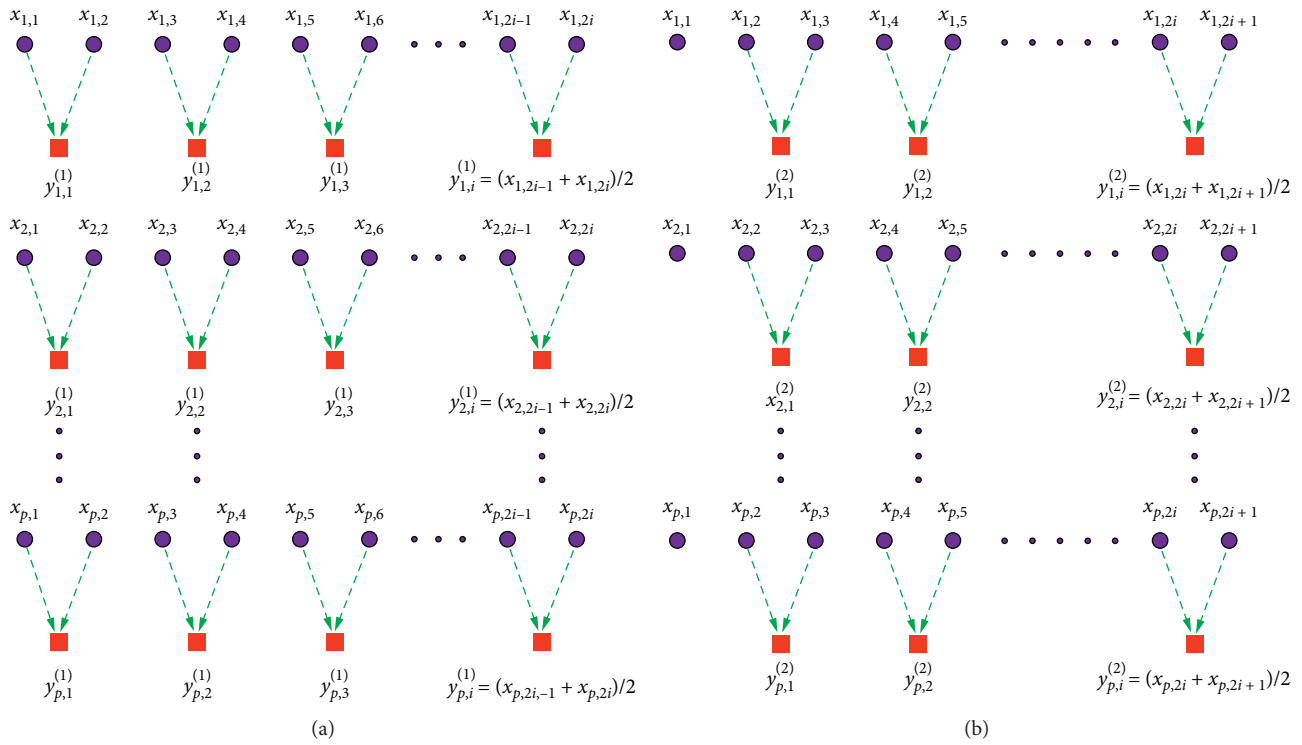


FIGURE 1: Illustration of refined composite coarse-grained approach for multivariate data with scale factor 2. (a) First coarse-grained time series; (b) second coarse-grained time series.

has better separation performance. What's more, the fluctuation of the RCmvMPE curve is greater than that of RCmvMAAPE, especially the fluctuation of (d) is obvious. This phenomenon shows that RCmvMAAPE is more stable when analyzing multivariate data and is not prone to large errors. In addition, when the scale factor is 14–20, RCmvMSE cannot effectively distinguish between (b) and (d). Similarly, mvMAAPE cannot effectively distinguish (a) and (c); meanwhile, the entropy value of four multivariate signals has extremely large fluctuation, which also verifies that the traditional coarse-graining method is prone to large errors. In a word, compared with the other three multivariate analysis methods, RCmvMAAPE enjoys better separation performance and robustness, thereby can better characterize the complexity of multivariate signals.

3. The Principle of the WOA-KELM

3.1. Kernel Extreme Learning Machine. Kernel extreme learning machine is a training algorithm based on single-hidden layer feedforward neural network. It does not require to repeatedly adjusting the hidden layer parameters [28]. In addition, the conventional single-hidden layer feedforward neural network parameter training problem is transformed into solving linear equations, and the smallest norm least-squares solution obtained is used as the network output weight. The whole training process is completed once. Therefore, the training speed is greatly improved and the generalization performance is better.

For input and output data, the goal of ELM is to simultaneously minimize training error and output weight norm, which can be expressed as follows:

$$\begin{cases} \min \sum \left\| \beta \cdot h(x_i) - t_i \right\|^2, \\ \min \|\beta\|, \end{cases} \tag{18}$$

where β is the connection weight vector between the hidden layer and the output layer and $h(x_i)$ is the kernel mapping of the hidden layer.

The optimization problem of equation (18) is simplified to the following constraint problem:

$$\begin{cases} \min L = \dfrac{1}{2}\|\beta\|^2 + C\dfrac{1}{2}\displaystyle\sum_{i=1}^{N} \xi_i^2, \\ \\ h(x_i)\beta = t_i - \xi_i,\ i = 1, 2, \ldots, N, \end{cases} \tag{19}$$

where ξ_i stands for training error and C denotes the penalty factor. Using the theory of orthogonal projection, the training process of ELM is equivalent to solving the following dual optimization problems:

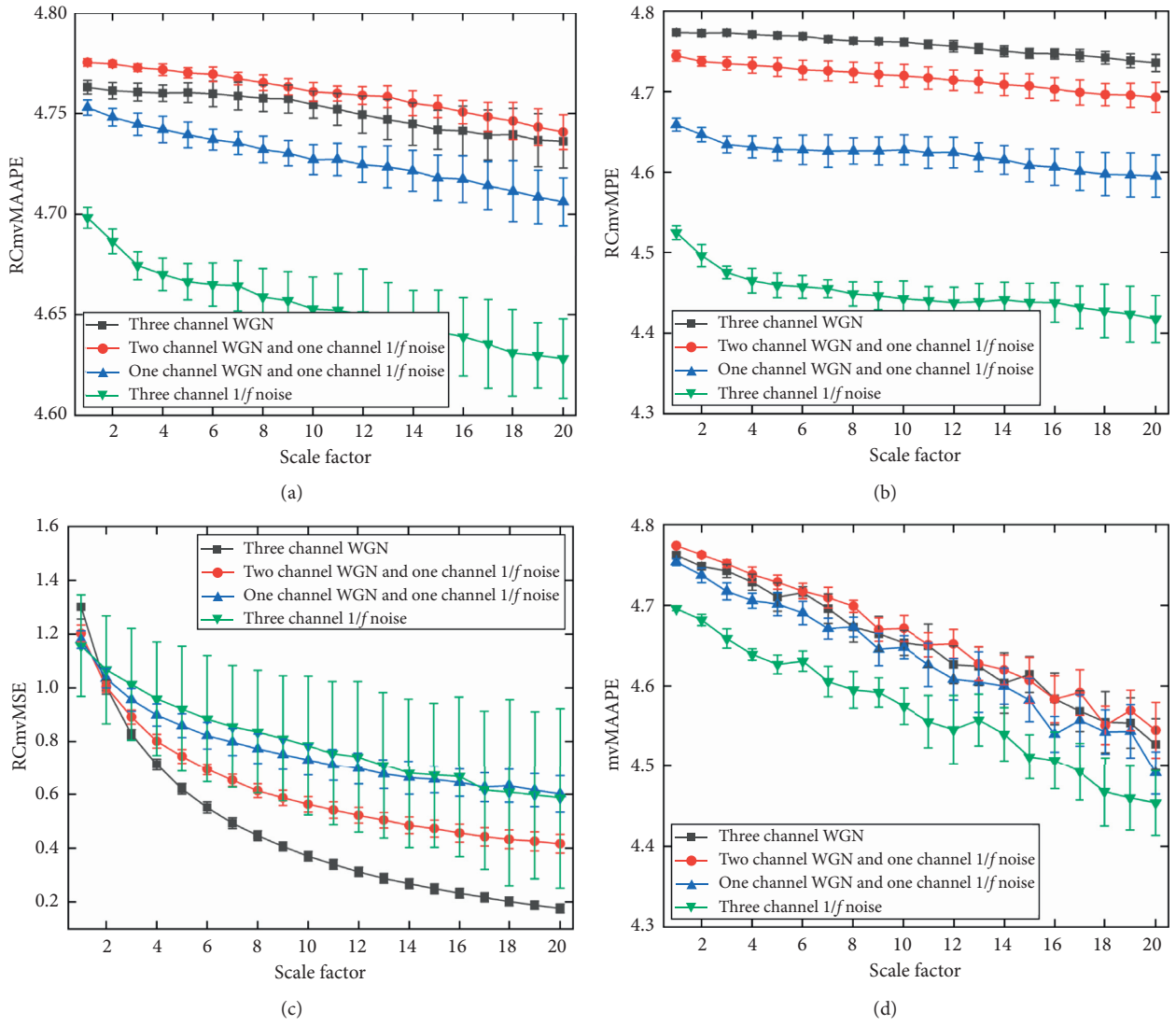


Figure 2: RCmvMAAPE, RCmvMPE, RCmvMSE, and mvMAAPE of multivariate synthetic signals. (a) RCmvMAAPE; (b) RCmvMPE; (c) RCmvMSE; and (d) mvMAAPE.

$$L_{\mathrm{ELM}} = \frac{1}{2}\|\beta\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i\left(h\left(x_i\right)\beta - t_i + \xi_i\right), \quad (20)$$

where α_i is the Lagrangian operator, and the derivative of it is

$$\frac{\partial L_{\mathrm{ELM}}}{\partial \beta} = 0 \Longrightarrow \beta = \sum \alpha i\left(h\left(x_i\right)\right)^T = H^T\alpha, \quad (21)$$

$$\frac{\partial L_{\mathrm{ELM}}}{\partial \xi_i} = 0 \Longrightarrow \alpha_i\xi_i = 0, \quad (22)$$

$$\frac{\partial L_{\mathrm{ELM}}}{\partial \alpha_i} = 0 \Longrightarrow h\left(x_i\right)\beta - t_i + \xi_i = 0, \quad (23)$$

where $\alpha = [\alpha_i, \ldots, \alpha_N]^T$.

Substituting formulas (20) and (21) into formula (22), the formula (23) can be equivalently written as follows:

$$\left(\frac{I}{C} + \mathrm{HH}^T\right)\alpha = T. \quad (24)$$

The corresponding output function of ELM is described as follows:

$$f\left(x\right) = h\left(x\right)\beta = h\left(x\right)H^T\left(\frac{I}{C} + \mathrm{HH}^T\right)^{-1}T. \quad (25)$$

It can be seen from the formula (25) that the parameter I/C is added to the main diagonal in the unit diagonal HH^T, thereby its eigenvalue cannot be 0. Then, the weight vector is

computed. ELM is more stable and has strong generalization ability in this way.

The kernel function is introduced into ELM and the KELM algorithm is proposed. Mercer condition is applied to define the kernel function matrix of KELM as follows:

$$\begin{aligned} \Omega &= HH^T, \\ \Omega_{i,j} &= h(x_i) \times h(x_j) = K(x_i, x_j), \end{aligned} \tag{26}$$

where $K(x_i, x_j)$ denotes the kernel function and the elements of the kernel matrix $\Omega_{i,j}$ in row i and column j, $i, j \in (1, 2, \ldots, N)$.

Therefore, it can be concluded that the actual output of the KELM model is

$$f(x) = h(x)H^T \pi\left(\frac{I}{C} + HH^T\right)^{-1} T = \begin{bmatrix} K(x, x_1) \\ \cdots \\ K(x, x_N) \end{bmatrix} \left(\frac{I}{C} + \Omega\right)^{-1} T. \tag{27}$$

3.2. Whale Optimization Algorithm. Whale optimization algorithm (WOA) is a novel heuristic search optimization algorithm [31]. Its advantages lie in its uncomplicated operation, less adjustment parameters, and strong capability to jump out of local optimum. The algorithm mainly imitates three behaviors of humpback whale, including encircling prey, hunting prey, and searching prey.

WOA supposes that the current best candidate solution is the target quarry or close to the best. After defining the best search agent, other search agents will therefore try to renew their best-located search agents. The update formula of WOA position is as follows:

$$\begin{aligned} D &= \left|CX^*(t) - X(t)\right|, \\ X(t+1) &= X^*(t) - A\,D, \end{aligned} \tag{28}$$

where A and C are the coefficients; t is the number of iterations; $X(t)$ represents the current position vector of the whale; and $X^*(t)$ denotes the best whale position vector so far. The mathematical expressions of A and C are as follows:

$$\begin{aligned} A &= 2ar_1 - a, \\ C &= 2 \cdot r_2, \\ a &= 2\left(1 - \frac{t}{T_{\max}}\right), \end{aligned} \tag{29}$$

where $T_{\max}$ represents the maximum number of iterations and r_1 and r_2 are random numbers in the interval $[0, 1]$. The value of a decreases linearly from 2 to 0, and t is the number of iterations.

When hunting, humpback whales not only swim to the prey in spiral shape but also contract the encircling circle. The position of whales is updated with 50% probability between the contraction mechanism and the spiral model.

$$X(t+1) = \begin{cases} X^*(t) - A \cdot D & \text{if } p < 0.5, \\ D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t) & \text{if } p \geq 0.5, \end{cases} \tag{30}$$

where $D' = |X^*(t) - X(t)|$ denotes the distance between the whale and its prey; the constant b is used to define the spiral shape; and l is a random number in $[-1, 1]$.

When the humpback whale attacks the prey, by linearly reducing the value of parameter a, the fluctuation range of A is continuously decreased and the value of A in the interval $[-a, a]$ decreases continuously as a decreases. When the value of A is in the interval [–1, 1], the solution position of the whale's next search agent will be any position between the current position and the prey position. By simulating the behavior of the humpback whale attacking the prey, the development capability of local search is shown. When the random value of A is greater than 1 or less than −1, the humpback whale search agent moves away from the prey to search, thereby finding a more suitable prey, which shows the exploration function of the whale optimization algorithm in the global search.

3.3. Whale Optimization Algorithm-Based Kernel Extreme Learning Machine (WOA-KELM). Considering that the performance of the KELM is easily affected by penalty factors and kernel parameters, a new method for optimizing the kernel extreme learning machine by whale optimization algorithm is raised. The optimization procedure is presented in Figure 3, and the detailed step is as follows:

(1) Input training set and testing set samples and normalize the two sample sets, respectively.

(2) Initialize the position of whale population and set the population number to N. The maximum iteration number is $T_{\max}$.

(3) Initialize the parameters of KELM and select the corresponding fitness function.

(4) The fitness of each whale is computed and sorted according to the fitness value, so as to continuously update the whale population.

(5) When the fitness value meets the conditions or reaches the maximum number of iterations, the optimization process is terminated.

(6) According to the optimal penalty factor and kernel function parameter, the KELM fault diagnosis model is established.

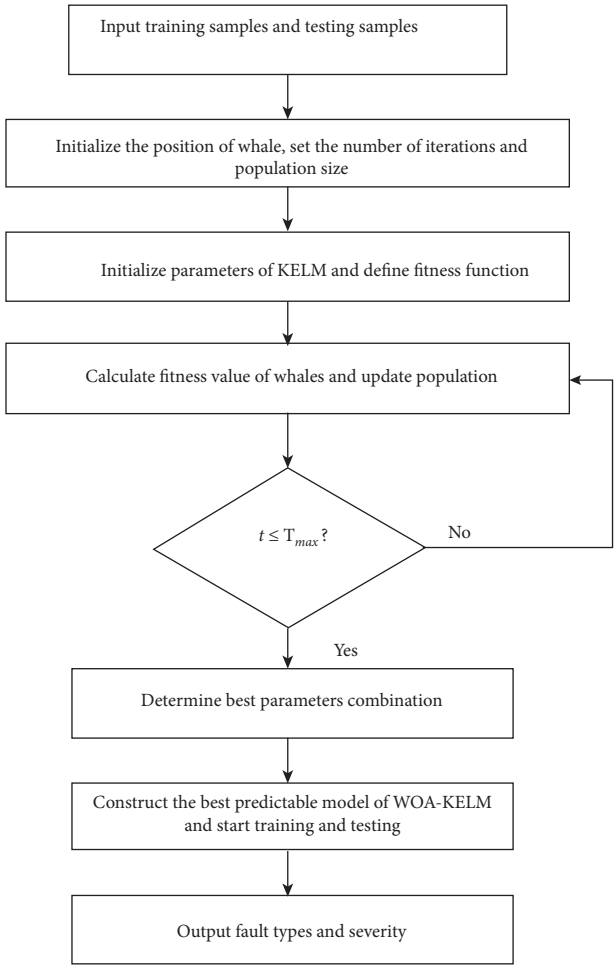


Figure 3: The flow chart of WOA-KELM algorithm.

(7) The trained KELM health condition detection model is employed to output the fault type and severity of the testing data.

4. The Proposed Approach

In this study, considering that RCmvMAAPE possesses excellent performance of processing multivariate time series, it is used to extract the fault features of rotating machinery. Combining mRmR and WOA-KELM, an integrated health condition detection method for rotating machinery is proposed. The method includes fault detection and health condition recognition.

4.1. Fault Detection. The ability of mvAAPE to measure the complexity of multivariate nonlinear data and the probability of dynamic mutation is the basis for fault diagnosis. Since mvAAPE is proposed based on mvPE, it inherits the ability of mvPE to detect failures. The inconsistent entropy values of mvAAPE corresponding to different states are a prerequisite for fault screening.

The mvAAPE values of the rotating machinery vibration signals in all fault states are greater than that in the normal state, and the difference is obvious. Therefore, mvAAPE can be applied for fault screening. In order to determine the screening criteria intuitively, a threshold based on mvMAAPE is set. When the mvMAAPE value of the vibration signal of rotating machinery in an unknown state is less than the threshold, the state is determined to be healthy. Conversely, if it is greater than the threshold, it is determined that there is a fault.

4.2. Health Condition Recognition. After fault detection, if it is detected that there is a fault in rotating machinery, further analysis is required to judge the type and severity of the fault. Firstly, RCmvMAAPE is employed to acquire the nonlinear complex information of fault multichannel vibration signals to form the initial fault feature vectors. However, the RCmvMAAPE values at all scales may include redundant information, so it is necessary to compress the feature dimensions to obtain sensitive feature vectors. The mRmR is a dimensionality reduction algorithm for nonlinear data, which uses mutual information to measure the correlation and redundancy of features, so as to realize the importance ranking of features. Therefore, the mRmR is utilized to screen the initial fault features to obtain sensitive feature vectors. Finally, the whale optimization algorithm is utilized to optimize the kernel function parameter and penalty factor of KELM to construct the optimal classification model and accomplish the health condition recognition of rotating machinery.

The flowchart of the raised approach is shown in Figure 4 and the implementation procedures of the integrated health condition detection method are listed as follows:

(1) Multichannel vibration signals of rotating machinery under diverse working conditions are collected.

(2) Divide the collected vibration data into multiple nonoverlapping samples of length N.

(3) Compute the mvAAPE value of the vibration signal and establish a threshold based on mvAAPE to determine the health condition of the rotating machinery. If the mvAAPE value of the vibration signal to be detected is less than the threshold value, it indicates that the rotating machinery is healthy. The output is normal and the diagnosis terminates. Otherwise, the next step is conducted to judge the fault type and severity of the rotating machinery.

(4) RCmvMAAPE is utilized to extract fault information from fault vibration signals of rotating machinery to generate the initial fault features.

(5) The mRmR method is employed to screen the sensitive feature from the initial fault feature to form the sensitive feature vectors.

(6) The training set samples are utilized to train the WOA-KELM-based multiclassifier.

(7) The testing set samples are fed to the trained multiclassifier for prediction. The fault type and severity are recognized in line with the output of WOA-KELM multifault classifier.

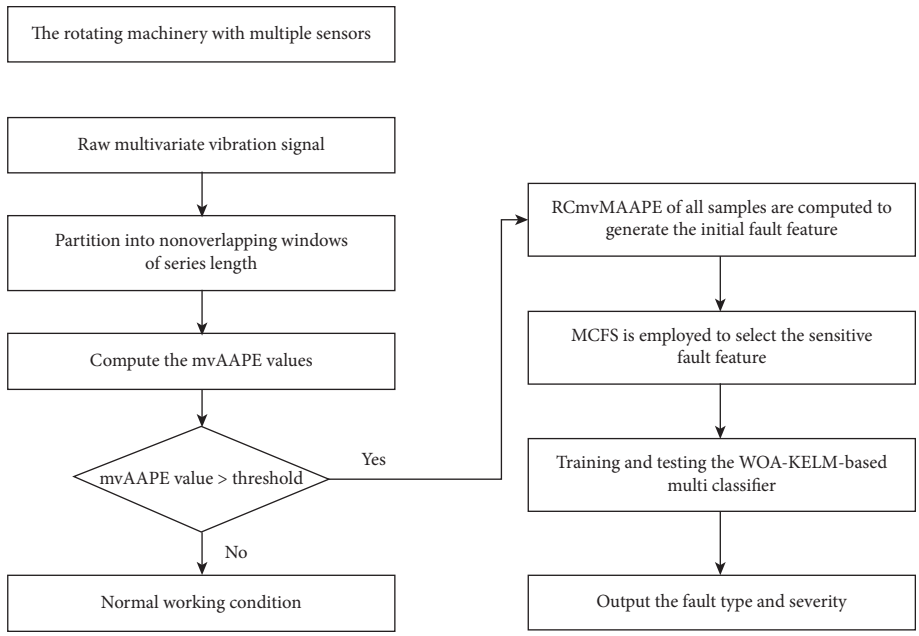


Figure 4: The flowchart of the proposed integrated health condition detection method.

5. Experimental Analysis and Results

In order to study the health condition detection method for rotating machinery raised in this paper to verify its universality and effectiveness for fault identification of general rotating machinery, experiments and analysis are conducted using two typical examples, namely, rolling bearings and gearboxes. The rolling bearing dataset was provided by CWRU [32]. The gearbox experiment data were collected on the QPZZ-II vibration analysis platform produced by Jiangsu Qianpeng Diagnostic Engineering Co., Ltd.

5.1. Health Condition Detection Experiment of Rolling Bearing

5.1.1. Experimental Rig and Data Introduction. The data were collected by the high-precision multichannel sensor installed on the bearing experimental rig. The specific structure of the bearing experimental rig is presented in Figure 5. The experimental rig includes a motor, a torque transducer/encoder, control electronics, and a dynamometer. The installation position of the acceleration sensors is at the 12 o'clock position at both the drive end and fan end of the motor housing, which are connected with the magnetic casing. The collected experimental data are the vibration waveforms of the motor, which are collected by the 16-channel data recorder. Single-point faults are set on SKF rolling bearings by electrical discharge machining. The fault diameter is 0.1778 mm, 0.3556 mm, and 0.5334 mm, respectively, and the fault depth is 0.2794 mm. The three fault diameters represent the different severity of the bearing fault. The experimental environment is set as follows: the motor load is 0 hp, the motor speed is 1797 r/min, and the sampling frequency is 12 kHZ. In this article, the data used include 10 categories, normal bearings, inner race faults, outer race faults, and ball faults. The fault diameter of each fault state is

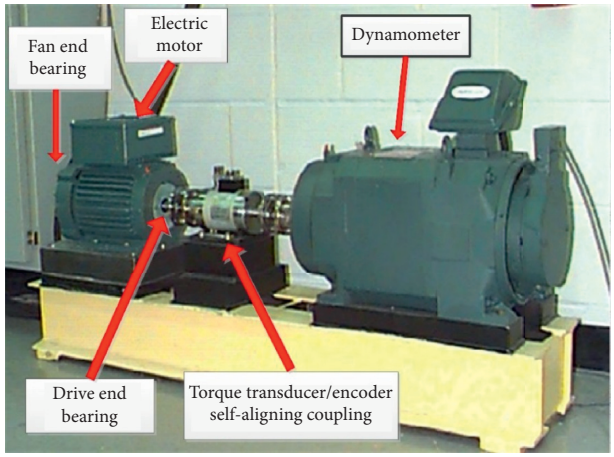


Figure 5: The rolling bearing test platform.

0.1778 mm, 0.3556 mm, and 0.5334 mm (label as NM, IRF1, IRF2, IRF3, ORF1, ORF2, ORF3, BF1, BF2, and BF3, respectively). For each fault state, the synchronous vibration signal at the drive end and fan end is used as dual-channel data. Generally, in the field of bearing fault diagnosis, the vibration signals are basically collected at the drive end. Since the data quality of the driver end is higher, which contains less noise and can directly reflect the vibration of the output part, however, for the fault diagnosis of mechanical equipment, high accuracy of fault identification is our goal. Therefore, it is necessary for us to use all available information to improve the utilization rate of information. The data of the fan end contains part of the fault information and the use of the data can significantly improve the characteristic quality, thus improving the fault recognition rate.

In this study, the vibration data of each working condition were divided into 58 samples without overlap, and the

number of sampling points of each sample was set to 2048. In order to be consistent with the engineering application under the actual condition, 28 samples for various working conditions are randomly selected for training, and the remaining 30 samples are the testing set. The effectiveness of the raised approach is validated by randomly selecting training and testing samples. The specific introduction of the dataset is presented in Table 1.

5.1.2. Fault Detection. The time domain waveforms of rolling bearing under ten working conditions are shown in Figure 6. Due to the lack of regularity, it is hard to directly recognize diverse working conditions based on their original vibration signals. According to previous analysis, PE has the ability to detect faults, the mvAAPE is obtained based on the theory of multidimensional embedding, and reconstruction also enjoys the same function. Therefore, mvAAPE can be used to detect whether the equipment is faulty. Figure 7 shows the mvAAPE values for all samples. As presented in Figure 7, the mvAAPE values in the fault states are generally large and the mvAAPE of the normal state is small, which is significantly different from the mvAAPE values of the fault states. Consequently, this method can be used to screen the normal state of the bearing. The value at the blue dotted line is defined as the mvAAPE threshold (2.9973). By comparing the mvAAPE value of the vibration signals with the threshold, the normal and fault states can be clearly distinguished. However, the samples of different fault types have poor separability, so mvAAPE cannot be used as the standard to judge the fault type and severity. A further analysis is needed to obtain more reliable characteristics.

The fault samples have the maximum mvAAPE value, which demonstrates that they are more complicated than normal samples. When the bearing is in normal operation, the vibration mainly comes from the interaction and coupling between the mechanical parts and the ambient noise, thereby the vibration signal shows certain regularity. Therefore, the mvAAPE value of normal condition is lower than that of the fault condition. When a fault occurs in the running process of the bearing, the vibration of the bearing will produce periodic pulse components. The high frequency vibration is mixed with the bearing vibration, which makes the frequency component and bandwidth of vibration signal more complex.

The first procedure in fault diagnosis is health detection. For a complicated mechanical system, it is necessary to judge whether there is a fault in the component firstly and then identify the type and severity of the fault. If the system does not detect the fault, it indicates that the system is running normally, and there is no need to disassemble and repair it.

5.1.3. Fault Recognition. Once a bearing fault is detected, the raised approach is used to distinguish the diverse fault types and severity. To validate the advantages of multivariate analysis, univariate analysis methods such as RCMAAPE are employed to test the bearing vibration signals at the drive end. By comparing with the univariate feature extraction method, the advantages of multichannel analysis in terms of information utilization are intuitively verified. Each method uses data from 9 fault conditions for experiments. The entropy results of univariate analysis method RCMAAPE and multivariate analysis methods RCmvMAAPE, RCmvMPE, RCmvMSE, and mvMAAPE are shown in Figures 8(a)–8(e).

Compared with other multivariate analysis methods shown in Figures 5(b)–5(d), the entropy deviation of RCmvMAAPE is smaller and the stability is higher. First of all, when the scale factor is 5–16, RCmvMPE has poor discrimination of NM, IRF3, and ORF3. In addition, mvMAAPE is generally poorly distinguished, and the entropy deviation of each fault state is very large, which indicates its performance is unstable and easily causes large errors. Except for NM and ORF2, the RCmvMSE curves of the other states are similar on most scales, and the degree of overlap is high, making it difficult to distinguish them. For the other two univariate analysis methods, entropy deviation is significantly greater than that of the multivariate analysis method, and the degree of entropy curve overlap is also greater than that of the multivariate analysis method. This is mainly because the univariate analysis method only uses the vibration information of one channel, so the utilization rate of information is relatively low, while the multivariate analysis method realizes the effective use of information by comprehensively considering the vibration information of multiple channels, thus improving the stability and robustness of the analysis. Therefore, based on the abovementioned analysis, RCmvMAAPE is more effective in feature extraction than RCmvMPE, RCmvMSE, mvMAAPE, and RCMAAPE, while the quality of the extracted features is also higher.

According to the abovementioned analysis, although the features extracted by the RCmvMAAPE method have high quality and can represent the fault state well, the fault features on the partial scale enjoy low separability and cannot achieve satisfactory distinguishing effect. For the sake of reducing the redundancy between features and enhancing the separability of fault features, the mRmR approach is utilized to reduce the dimension of original features. The distribution of multiscale features after the rearrangement is visually described in Figure 9. The dimensionality of the new multiscale fault features is selected as 9 according to the correlation with the main fault information and the importance of the features. Finally, the obtained new fault features are input into the WOA-KELM classifier to determine the fault type and severity. Figure 10 shows the failure classification results for one trial. It can be clearly observed from the figure that all the faults have been accurately identified and the classification accuracy has reached 100%, which indicates that the proposed approach can availably distinguish the types and severity of faults.

In addition, for the sake of avoiding the influence of random factors such as contingency on the experimental results, 20 trials are repeated to obtain more accurate and reliable classification results. Moreover, four other entropy-based methods are also used to diagnose rolling bearing faults. The detailed classification results of the five approaches for 20 trials are presented in Figure 10 and Table 2.

Table 1: The detailed introduction of experiment sample.

Fault location	Fault diameter (mm)	Abbreviation	Training sample number	Testing sample number	Class label
Normal	0	NM	28	30	0
Inner race	0.1778	IRF1	28	30	1
	0.3556	IRF2	28	30	2
	0.5334	IRF3	28	30	3
Outer race	0.1778	ORF1	28	30	4
	0.3556	ORF2	28	30	5
	0.5334	ORF3	28	30	6
Ball	0.1778	BF1	28	30	7
	0.3556	BF2	28	30	8
	0.5334	BF3	28	30	9

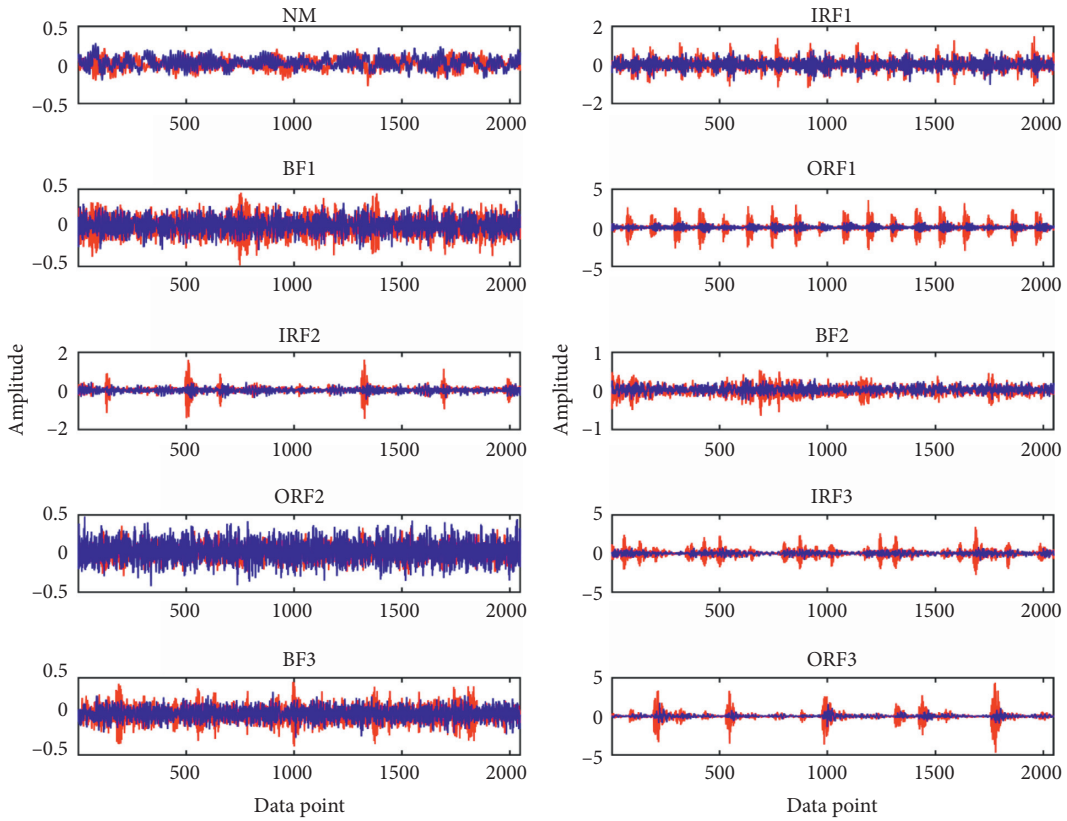


Figure 6: The waveforms of diverse classes of rolling bearing, where red denotes data of drive end and blue denotes fan end.

From Figure 11 and Table 2, it is obvious that the average classification accuracy of the raised approach is higher than that of other approaches, and the average accuracy rate is 99.96%. Moreover, the accuracy of the multivariate analysis methods (RCmvMAAPE, RCmvMPE, RCmvMSE, and mvMAAPE) is generally higher than that of the univariate analysis method (RCMAAPE), which is consistent with the previous analysis. Therefore, the comparison results indicate that the raised approach can effectively extract fault features and obtain high fault recognition rate.

To verify the necessity of mRmR feature selection, two-dimensional projections of two random features selected without adopting the mRmR method are presented in Figure 12(a), while the first two sensitive features obtained applying the mRmR method are visualized as Figure 12(b). By comparing Figures 12(a) and 12(b), it can be clearly found that RCmvMAAPE combined with mRmR has a better recognition effect than using RCmvMAAPE alone. Moreover, nine random features ($\tau = 8, 19, 1, 17, 9, 3, 20, 14, 6$) are directly inputted into WOA-KELM to identify the fault type and the identification results are presented in Table 3. According to the results in Table 3, it can be clearly found that the fault recognition accuracy rate gained without using the mRmR method is lower than that gained with adopting the mRmR method. In addition, it can be noticed that the recognition accuracy of RCmvMAAPE is still higher than that of other methods without using mRmR. Thus, the experimental results again verify that RCmvMAAPE can extract fault features from multichannel signals effectively and improve the quality of fault information. The mRmR method can select sensitive low-dimensional features from high-dimensional fault

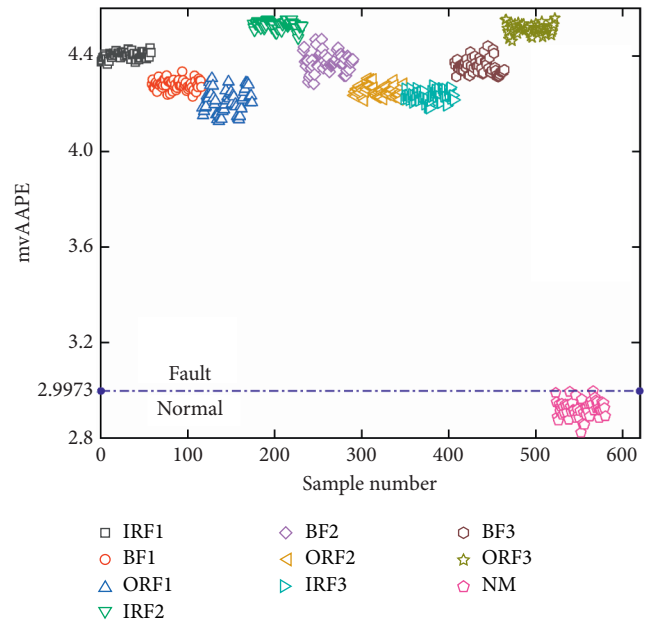


Figure 7: The multivariate amplitude-aware permutation entropy (mvAAPE) distribution of all samples.

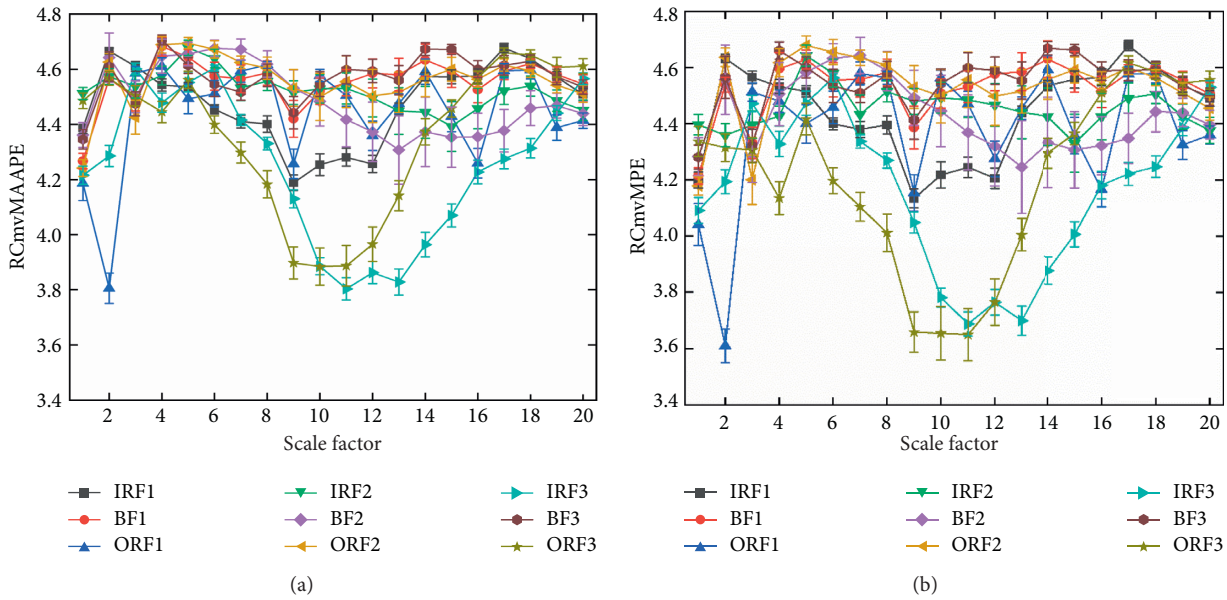


Figure 8: Continued.

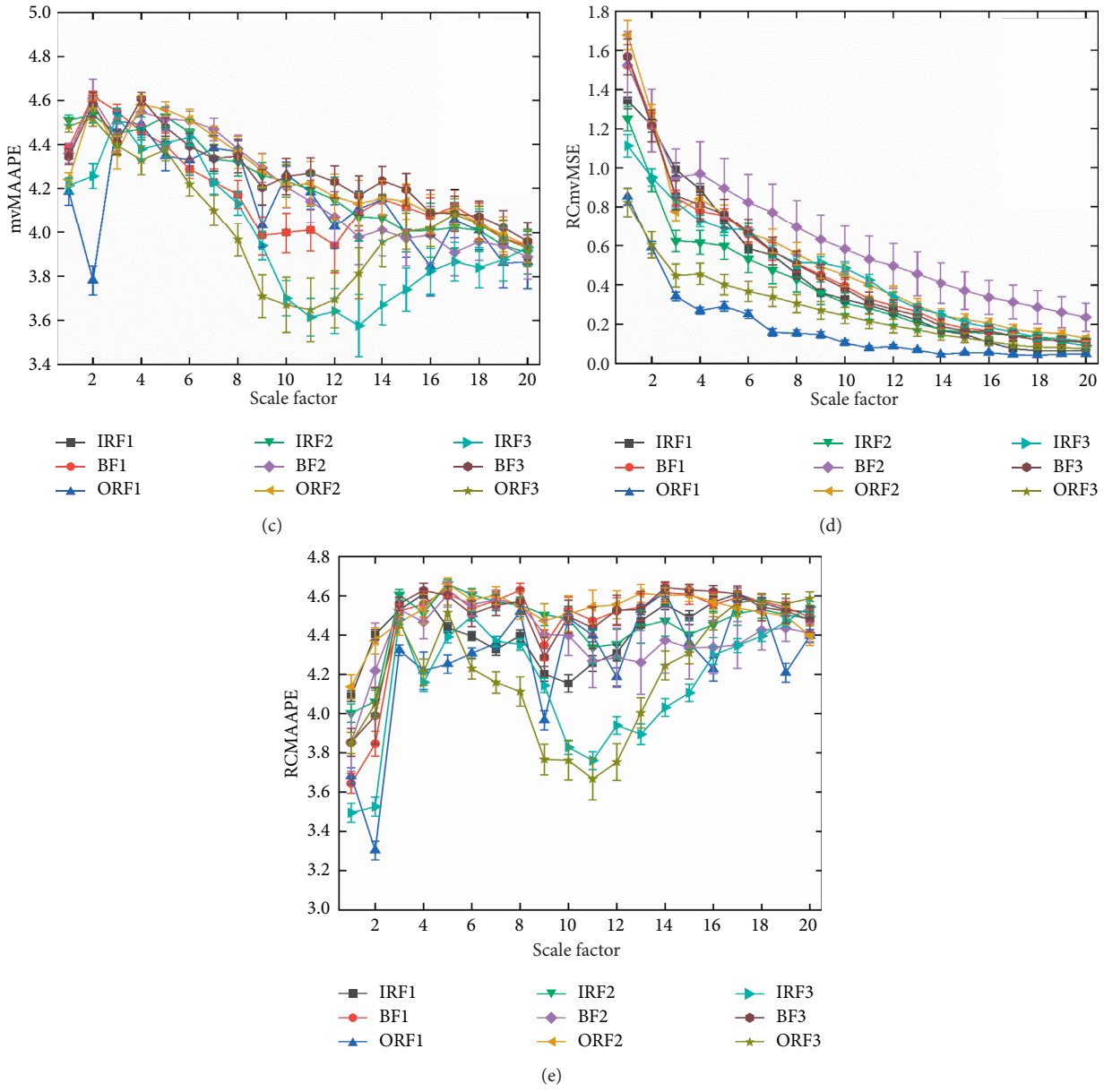


Figure 8: The entropy results of rolling bearing data analyzed by adopting five approaches. (a) RCmvMAAPE; (b) RCmvMPE; (c) mvMAAPE; (d) RCmvMSE; (e) RCMAAPE.

features, which not only improves the recognition accuracy but also improves the classification efficiency.

This section discusses the superiority of using WOA algorithm to optimize KELM in fault identification. For comparison, three commonly used classifiers are used for comparison, namely, support vector machine (SVM), extreme learning machine (ELM), and kernel extreme learning machine (KELM). The ratio of training samples to testing samples remains the same. The diagnostic results of the five approaches using diverse classifiers are listed in Table 4. It can be seen that when the four classifiers are combined with the five feature extraction methods, the classification accuracy of WOA-KELM is the highest, which shows that WOA-KELM is an effective classifier. In addition, it can be clearly found that when the features obtained by different feature extraction methods are input to the four classifiers, the classification accuracy of RCmvMAAPE is the highest, which further verifies that the raised RCmvMAAPE approach has excellent performance in feature extraction.

5.2. Health Condition Detection Experiment of Gearbox

5.2.1. Experimental Rig and Data Introduction. The gearbox experiment data were collected from the experiment platform QPZZ-II that is built by Jiangsu Qianpeng Diagnosis Engineering Co., Ltd. The overall structure of the experimental platform is shown in Figure 13. The experimental platform is composed of gearbox, motor, iron base, capacitance, and sensors. The sensors are installed above the

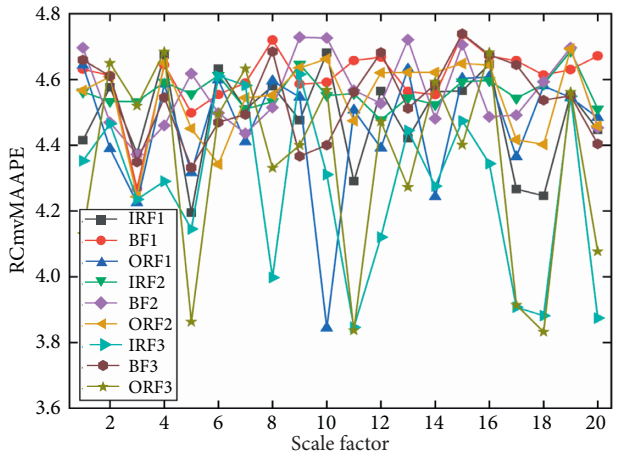


Figure 9: Distribution of multiscale feature after applying the mRmR approach.

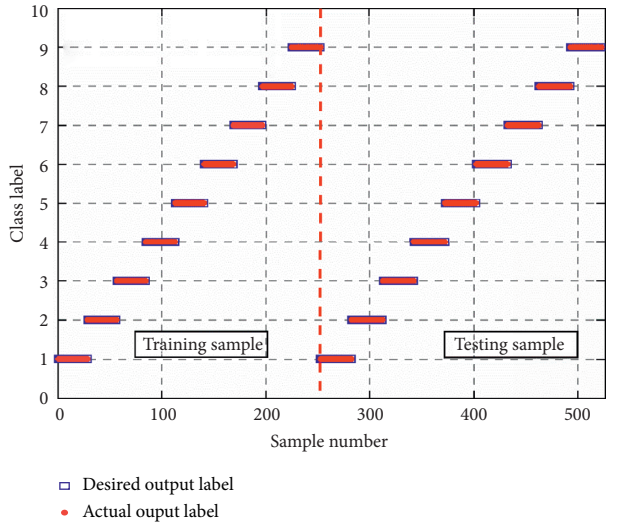


Figure 10: The recognition results of raised approach for rolling bearing.

Table 2: Identification result of five approaches for rolling bearings with mRmR feature selection.

Diverse approaches	Accuracy (%)			
	Max	Min	Mean	SD
The proposed method	100	99.26	99.96	0.1655
RCmvMPE and mRmR	100	97.41	98.54	0.8358
mvMAAPE and mRmR	91.11	86.30	88.69	1.2900
RCmvMSE and mRmR	95.56	92.22	93.92	1.0024
RCMAAPE and mRmR	90.37	85.56	87.52	1.4565

gearbox. The experimental data consist of eight channels of vibration signals and one channel of tachometer signals, in which the motor speed is 880 r/min. In the experiment, a total of four operating conditions were set up, including normal condition, gear pitting fault (pitting), gear tooth breaking (tooth breaking), pinion wear fault (wearing), and gear pitting fault coupling with pinion wear fault (pitting and wearing). The detailed introduction of gearbox

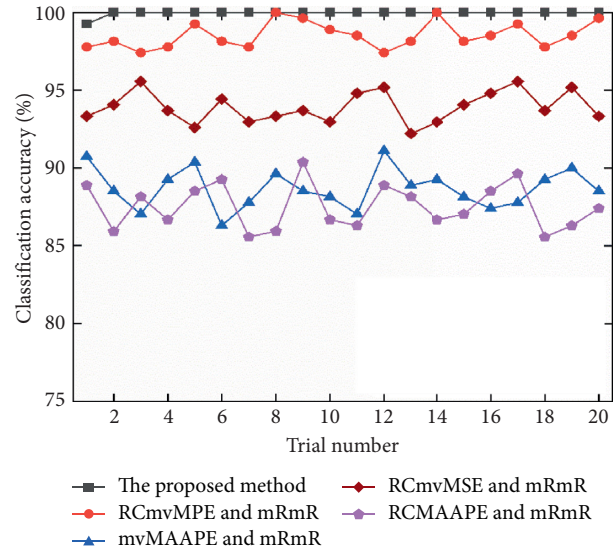


Figure 11: The diagnostic result of the five methods for 20 trials.

experimental data is shown in Table 5. The data acquisition equipment is QPZZ-II produced by Jiangsu Qianpeng Diagnostic Engineering Co., Ltd., with a sampling frequency of 5.12 kHZ and sampling time of 6 s. Therefore, each health state contains 53248 data points. The selected channels are the acceleration signal collected by the bearing X on the motor side of the input shaft and the bearing Y on the load side of the output shaft. The collected vibration signals are divided into 26 nonoverlapping samples with length 2048. Among them, 10 samples were used for training, and the remaining 16 groups were used for testing.

5.2.2. Fault Detection. The time domain waveforms of the gearbox under four working conditions are shown in Figure 14. It is difficult to directly judge the type of gear failure based on the amplitude and frequency changes of the waveforms. According to the previous analysis, mvAAPE can be used to detect whether mechanical equipment is faulty and is successfully used to detect the health condition of rolling bearings. Due to the complicated structure of the gearbox, it is difficult to disassemble and inspect the gearbox. Therefore, it is necessary to detect the health condition of the gearbox. Figure 15 shows the mvAAPE values of all samples of the gearbox. It can be observed from the figure that all faulty samples have larger mvAAPE values, while all normal samples have smaller mvAAPE values. The value shown by the blue dashed line is defined as the mvAAPE threshold (4.2342). By comparing the mvAAPE value of the sample to be tested with the threshold, it can be judged whether the gearbox is faulty. However, the entropy values between different fault samples are relatively close, and the fault type cannot be judged intuitively. Therefore, the mvAAPE value cannot be used as a criterion for judging the fault type and further analysis is needed to obtain more obvious characteristics.

The fault samples have larger mvAAPE values, which indicates that the vibration signals of the fault samples are

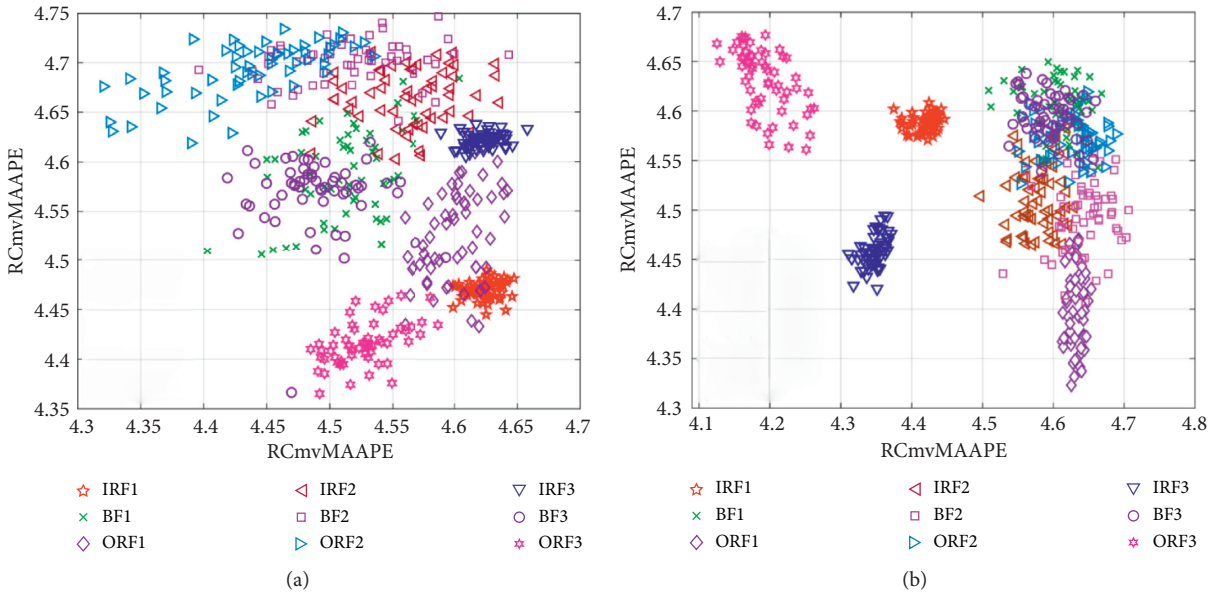


Figure 12: (a) Two-dimensional visualization of two random selected features without adopting mRmR. (b) Two-dimensional visualization of two new features selected utilizing mRmR.

Table 3: Identification result of five approaches without mRmR feature selection.

Methods	Accuracy (%)		
	Max	Min	Mean
RCmvMAAPE	94.44	89.26	92.32
RCmvMPE	91.85	88.15	90.69
mvMAAPE	87.04	84.44	85.86
RCmvMSE	90.37	87.78	89.63
RCMAAPE	86.30	81.85	84.27

more complicated than that of the normal samples. After the gearbox fails, the vibration signals enjoy obvious modulation characteristics, which are composed of multiple AM and FM signals. Compared with the vibration signals of the normal samples, the fault signals contain more impact components; meanwhile, due to the influence of random factors such as noise in the signal, the signal component is more complex, so it has a larger entropy value.

5.2.3. Fault Recognition. After detecting the gearbox failure, for the sake of identifying different fault types, the raised approach is utilized to process the fault vibration signals to obtain stronger features. Similarly, to verify the advantages of multivariate analysis, the univariate analysis method (RCMAAPE) is used for the motor side vibration signals. In addition, for the sake of studying the effectiveness of the RCmvMAAPE approach for extracting fault features, the RCmvMPE, mvMAAPE, and RCmvMSE approaches are used to analyze multichannel vibration signals. The analysis result is shown in Figures 16(a)–16(e).

It can be observed from Figure 16 that the overall trend of the RCmvMAAPE curve is consistent with that of RCmvMPE and mvMAAPE, but RCmvMAAPE has smaller entropy deviation, which indicates that the RCmvMAAPE method has better stability. Compared with the RCmvMSE method, the RCmvMAAPE curve has more obvious fluctuation, so it can effectively highlight the earth oscillation component of gearbox fault vibration signal, so as to extract fault features more effectively. In addition, compared with the univariate analysis method RCMAAPE, the entropy deviation of RCmvMAAPE is significantly smaller, that is, its performance is better. The main reason is that the univariate analysis method only makes rough use of the fault information in the single channel vibration signal, while the rich information in other channels is not used reasonably. However, after gearbox fails, the transmission path of internal vibration is complex and has multiple directions. The vibration signals collected from each channel contain the fault information, so it is impossible to fully characterize the fault state only by performing univariate analysis. Based on the abovementioned analysis, RCmvMAAPE can effectively analyze multichannel vibration signals and has stable performance.

It can be observed from Figure 16 that the fault features extracted by RCmvMAAPE are redundant at some scales, which indicates that not all features can be used for fault classification. It is necessary to screen them to select sensitive features. In order to improve the separability of fault features, the mRmR approach is used to process the features. The distribution of multiscale features after the rearrangement is visually described in Figure 17. The dimensionality of the new multiscale fault features is selected as 9 according to the correlation with the main fault information and the importance of the features. Finally, the obtained new fault features $\tau = (19, 8, 7, 16, 5, 13, 10, 3, 2)$ are fed into the WOA-KELM classifier to determine the fault type. Figure 18 shows the fault classification results for one trial. It can be clearly observed from the figure that except two samples of pitting and wear fault are misclassified as tooth breaking fault, the other faults are accurately identified, and the

Table 4: The diagnostic results gained by combining diverse methods with four classifiers.

Diverse classifiers	The testing accuracy of classifiers with diverse approaches (%)					Average accuracy (%)
	RCmvMAAPE (%)	RCmvMPE (%)	mvMAAPE (%)	RCmvMSE (%)	RCMAAPE (%)	
ELM	97.04	95.93	88.52	94.44	85.19	92.22
SVM	95.56	94.07	87.04	92.96	83.33	90.59
KELM	98.52	96.30	89.26	94.81	85.92	92.96
WOA-KELM	100	99.26	91.48	95.93	88.89	95.11
Average accuracy (%)	97.78	96.39	89.07	94.54	85.83	—

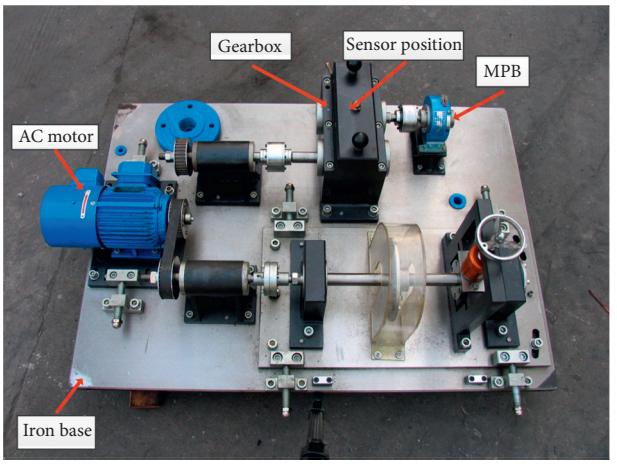


Figure 13: The experimental rig of the gearbox from QPZZ-II.

Table 5: The brief introduction of the experimental sample.

Fault type	Training sample number	Testing sample number	Class label
Normal	10	16	0
Wearing	10	16	1
Tooth breaking	10	16	2
Pitting and wearing	10	16	3

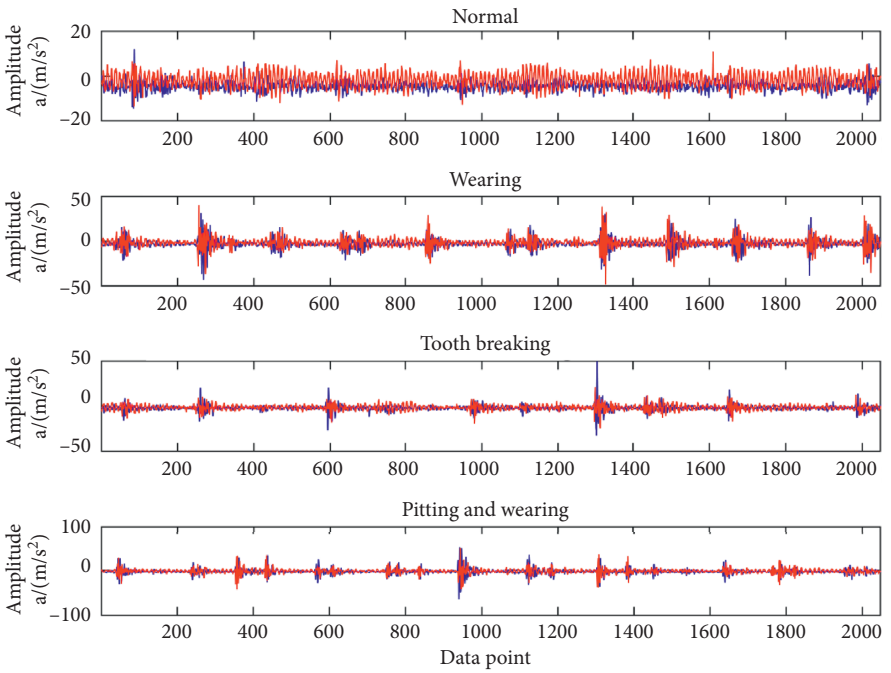


Figure 14: The vibration signal waveforms of the gearbox in different health conditions, where red denotes data of the motor side and blue denotes the load side.

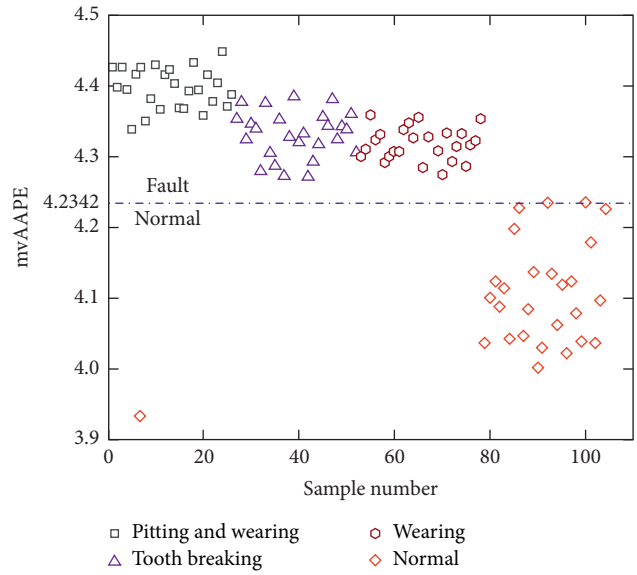


Figure 15: The multivariate amplitude-aware permutation entropy (mvAAPE) distribution of all samples.

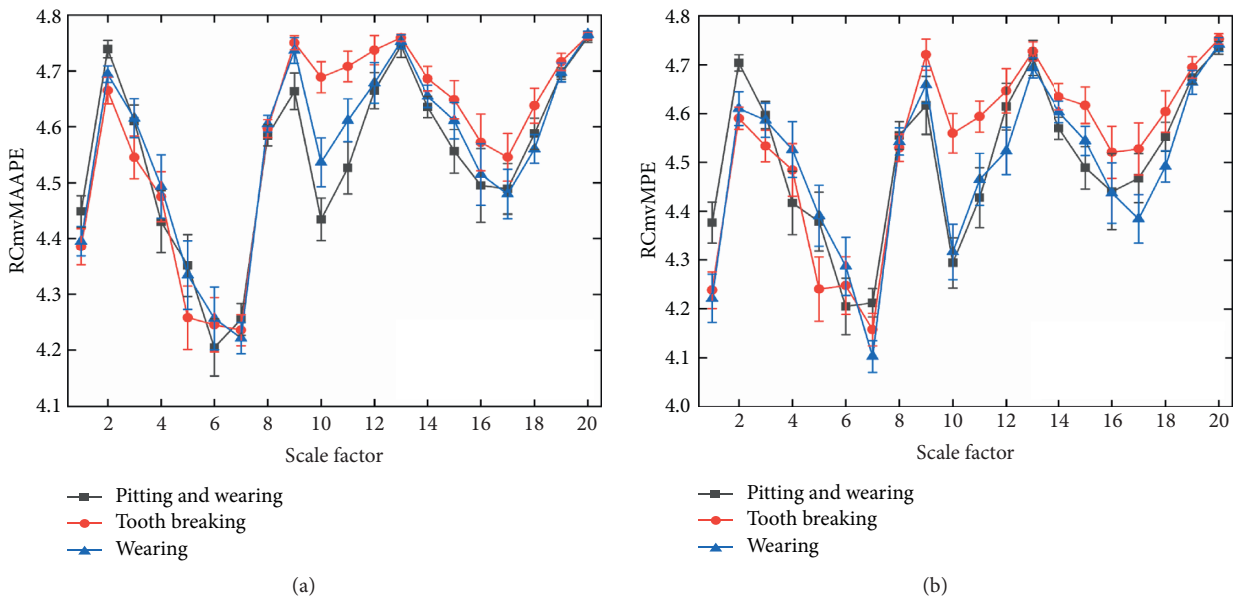


Figure 16: Continued.

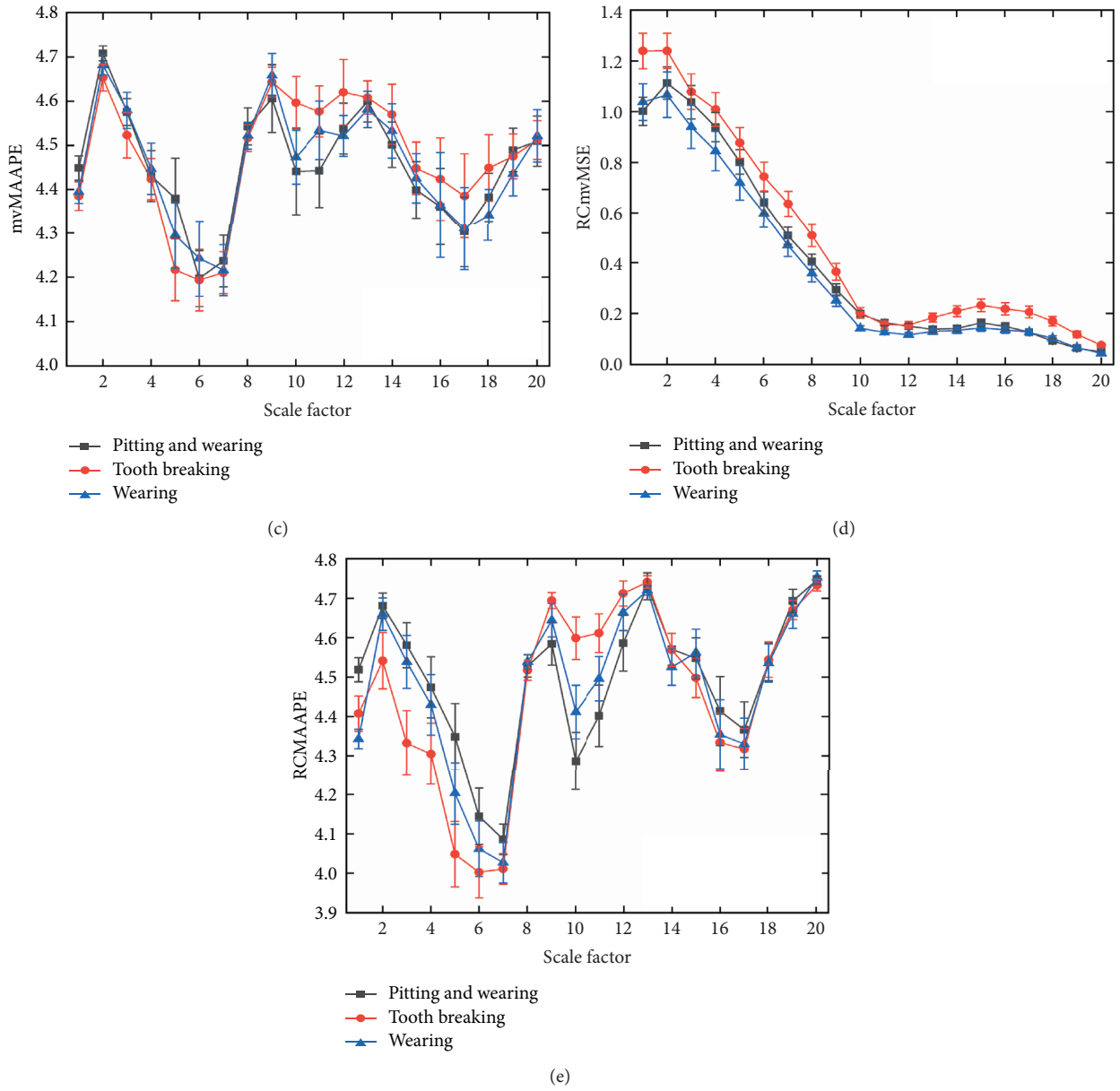


Figure 16: The entropy results of gearbox data analyzed by adopting five approaches. (a) RCmvMAAPE; (b) RCmvMPE; (c) mvMAAPE; (d) RCmvMSE; (e) RCMAAPE.

overall classification accuracy rate reaches 95.83%, which shows that the raised approach can availably distinguish different fault types of gearbox.

Similarly, in order to reduce the large randomness of experimental results due to only performing one trial, 20 trials are repeated to obtain more reliable and accurate classification results. In addition, in order to intuitively verify the advantages of RCmvMAAPE method, four other entropy-based methods are used to diagnose gearbox faults. The detailed classification results of five approaches for 20 trials are shown in Figure 19 and Table 6. It is obvious from Table 7 that the average recognition accuracy of the presented approach is the highest and the standard deviation is the smallest, which indicates that the raised approach has stable and excellent performance. The accuracy of RCmvMPE approach is slightly lower than that of the proposed approach, which indicates that RCmvMPE can also effectively diagnose gearbox faults. But the standard difference is large, indicating that the recognition rate is not stable. In addition, the accuracy of the multivariate analysis method is higher than that of the univariate analysis method, which verifies the necessity of multivariable analysis in gearbox fault diagnosis.

As before, for the sake of investigating the importance of mRmR feature selection, two-dimensional projections of two random features selected without adopting the mRmR method are presented in Figure 20(b), while the first two sensitive features obtained applying the mRmR approach are

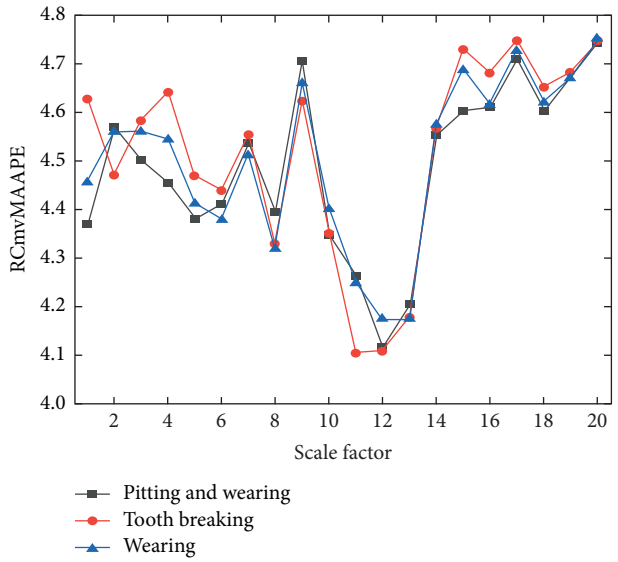


Figure 17: Distribution of multiscale feature after applying the mRmR approach.

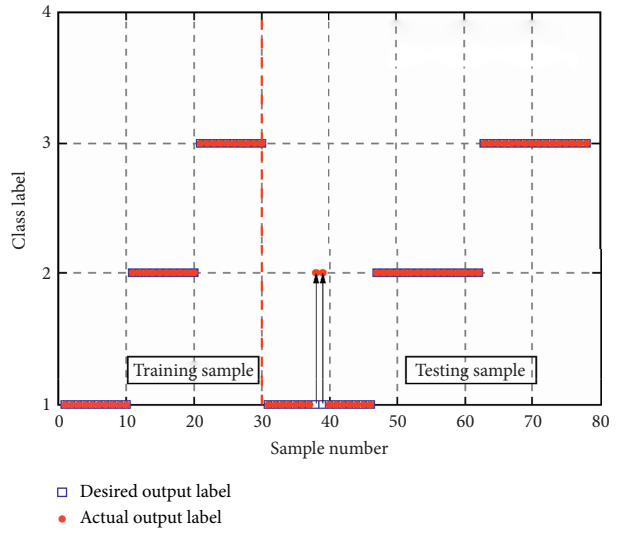


Figure 18: The recognition results of raised approach for gearbox.

visualized as Figure 20(a). It can be seen from the figure that the features without mRmR feature selection are disorderly and have no obvious clustering center, which indicates that the quality of features is not high and further processing is needed to obtain separable features. After mRmR feature selection, although no obvious clustering center is obtained, the separability of the three fault states becomes stronger. It can be concluded that mRmR feature selection can improve the recognition of features and has better recognition effect. Then, nine features are randomly selected and input into the WOA-KELM classifier to determine the fault type of gearbox. Similarly, each method was repeated 20 times. Table 7 shows the gearbox identification results of five methods without using mRmR feature selection for 20 trials. As can be seen from Table 7, although the highest recognition rate of the RCmvMAAPE approach is lower than that of the RCmvMPE method, the average recognition rate is still the highest, which indicates that the performance of RCmvMAAPE is more stable. Consistent with the previous analysis, the recognition accuracy of the multivariate analysis approach is higher than that of the univariate analysis approach, which directly verifies the necessity of multivariate analysis. In a word, mRmR dimension reduction can significantly improve the fault recognition rate, that is, improve the reliability of fault identification.

To validate the necessity of utilizing WOA-KELM, three commonly used classifiers are used for comparison: SVM, ELM, and KELM. The same proportion of training and test samples is employed to train and test the classifier. Table 8 shows the classification results of five approaches using diverse classifiers. It can be seen that the RCmvMAAPE approach still has the highest fault recognition rate when using different classifiers, which is higher than that of the RCmvMPE method. Obviously, amplitude-aware

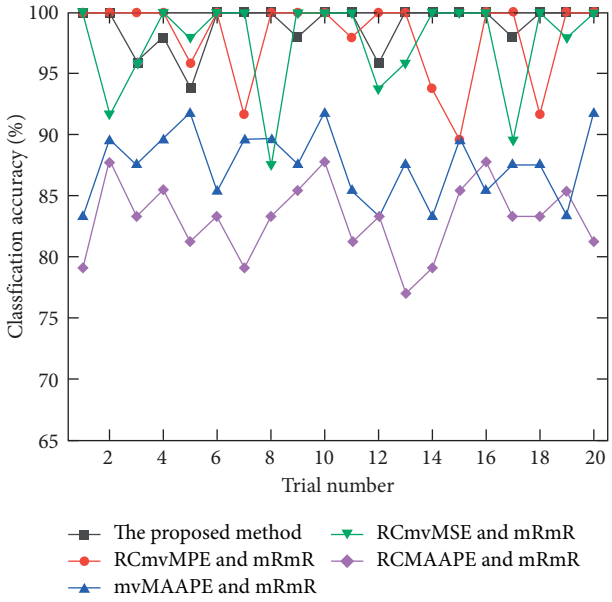


Figure 19: The diagnostic result of the five methods for 20 trials.

Table 6: Identification result of five approaches for gearbox with mRmR feature selection.

Diverse methods	Accuracy (%)			
	Max	Min	Mean	SD
The proposed method	100	93.75	98.96	1.8514
RCmvMPE and mRmR	100	89.58	98.02	3.4778
mvMAAPE and mRmR	91.67	83.33	87.50	2.8685
RCmvMSE and mRmR	100	87.50	97.5	3.9183
RCMAAPE and mRmR	87.80	77.08	83.17	3.0853

Table 7: Identification result of five approaches without mRmR feature selection.

Approaches	Accuracy (%)		
	Max	Min	Mean
RCmvMAAPE	89.58	85.42	87.22
RCmvMPE	91.67	83.33	86.94
mvMAAPE	81.25	72.92	78.46
RCmvMSE	87.5	81.25	84.32
RCMAAPE	75	66.67	72.53

Table 8: The diagnostic results gained by combining diverse methods with four classifiers.

Diverse classifiers	The testing accuracy of classifiers with diverse approaches (%)					Average accuracy (%)
	RCmvMAAPE (%)	RCmvMPE (%)	mvMAAPE (%)	RCmvMSE (%)	RCMAAPE (%)	
ELM	93.75	91.67	83.33	87.50	85.42	88.33
SVM	91.67	89.58	79.17	87.50	83.33	86.25
KELM	97.92	93.75	87.50	89.58	85.42	90.83
WOA-KELM	100	95.83	89.58	93.75	87.50	93.33
Average accuracy (%)	95.84	92.71	84.90	89.58	85.42	—

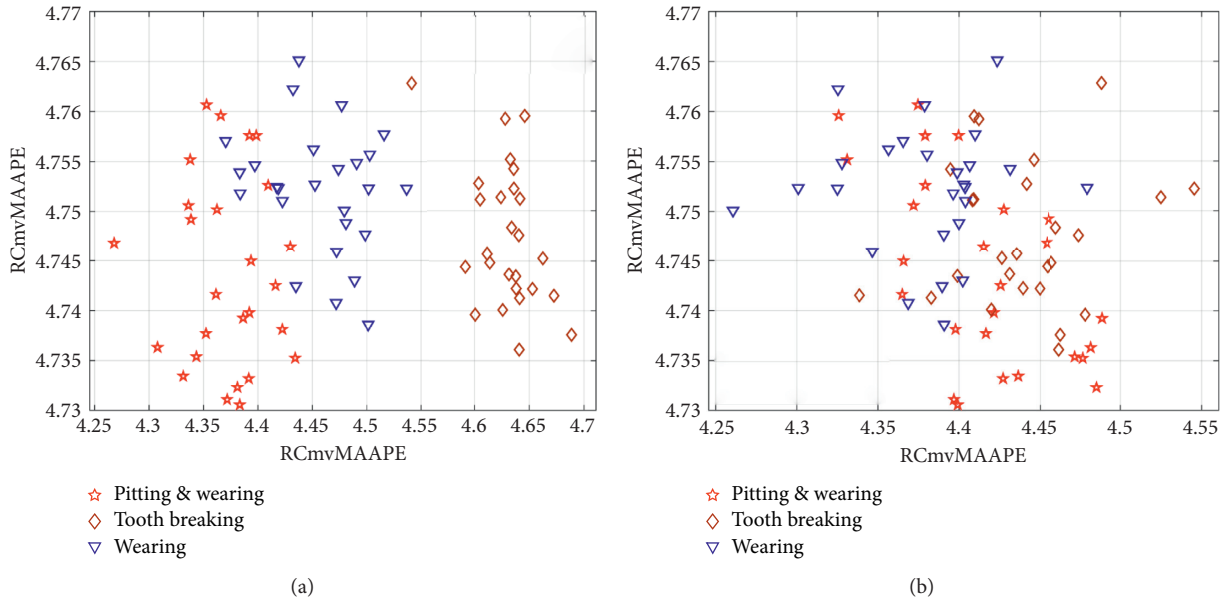


Figure 20: (a) Two dimensional visualization of two random selected features using mRMR. (b) two-dimensional visualization of two new features selected without using mRMR.

permutation entropy has better performance than permutation entropy by considering the amplitude and frequency information of time series. In addition, when the five methods are combined with different classifiers, the WOA-KELM classifier has the highest average recognition rate of 93.33%, which is higher than that of the KELM classifier alone. Since the performance of KELM is affected by the kernel parameters and penalty factor. The artificial setting cannot achieve the best classification effect. In conclusion, the WOA-KELM classifier has excellent performance, and the generalization performance is better than the commonly used classifiers.

6. Conclusions

In this study, a novel nonlinear analysis approach called RCmvMAAPE is raised. Various synthetic signals are analyzed and compared with RCmvMPE, mvMAAPE, and RCmvMSE. The results verify that RCmvMAAPE could effectively measure the complexity of multivariate time series and enjoys more stable performance. In the fault detection part, the mvAAPE is used to define a threshold. If the mvAAPE value of the measured sample is less than the threshold value, the equipment is normal, so as to realize the fault detection of the equipment. When a fault is detected, RCmvMAAPE is employed to extract fault features to construct initial feature vectors, and then mRmR is used to select sensitive features to form sensitive features to be classified. Finally, the sensitive feature vectors are input into the WOA-KELM classifier to determine the type and severity of the fault. The validity of the raised approach is verified by two typical examples, namely, rolling bearing and gearbox. The results demonstrate that the raised approach can not only accurately detect the fault of rotating machinery but also effectively identify the fault type. In addition, compared with other methods, RCmvMAAPE can extract higher quality fault features from multichannel vibration signals and is superior to that of common entropy-based methods, which verifies its effectiveness in feature extraction. From the perspective of practical application, the proposed method avoids the mode classification that is full of uncertainty and improves the effectiveness and timeliness of fault diagnosis by detecting the state of rotating machinery, thereby is more in line with the actual engineering needs.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 108–126, 2013.

[2] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: a review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33–47, 2018.

[3] A. Rai and S. H. Upadhyay, "A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings," *Tribology International*, vol. 96, pp. 289–306, 2016.

[4] P. Henriquez, J. B. Alonso, M. A. Ferrer, and C. M. Travieso, "Review of automatic fault diagnosis systems using audio and

vibration signals," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 642–652, 2013.

[5] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7319–7327, 2020.

[6] Q. Gao, C. Duan, H. Fan, and Q. Meng, "Rotating machine fault diagnosis using empirical mode decomposition," *Mechanical Systems and Signal Processing*, vol. 22, no. 5, pp. 1072–1081, 2008.

[7] J. Cheng, Y. Yang, and Y. Yang, "A rotating machinery fault diagnosis method based on local mean decomposition," *Digital Signal Processing*, vol. 22, no. 2, pp. 356–366, 2012.

[8] Y. Chen, T. Zhang, Z. Luo, and K. Sun, "A novel rolling bearing fault diagnosis and severity analysis method," *Applied Sciences*, vol. 9, no. 11, p. 2356, 2019.

[9] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2013.

[10] F. Zhou, X. Yang, J. Shen, and W. Liu, "Fault diagnosis of hydraulic pumps using PSO-VMD and refined composite multiscale fluctuation dispersion entropy," *Shock and Vibration*, vol. 2020, Article ID 8840676, , 2020.

[11] Y. Li, X. Wang, Z. Liu, X. Liang, and S. Si, "The entropy algorithm and its variants in the fault diagnosis of rotating machinery: a review," *IEEE Access*, vol. 6, pp. 66723–66741, 2018.

[12] R. Yan and R. X. Gao, "Approximate Entropy as a diagnostic tool for machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 824–839, 2007.

[13] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[14] Z. Jinde, C. Junsheng, and Y. Yang, "A rolling bearing fault diagnosis approach based on LCD and fuzzy entropy," *Mechanism and Machine Theory*, vol. 70, pp. 441–453, 2013.

[15] R. Yan, Y. Liu, and R. X. Gao, "Permutation entropy: a nonlinear statistical measure for status characterization of rotary machines," *Mechanical Systems and Signal Processing*, vol. 29, pp. 474–484, 2012.

[16] X. Zhang, Y. Liang, J. Zhou, and Y. Zang, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164–179, 2015.

[17] K. Moshen, C. Gang, P. Yusong, and L. Yong, "Research of planetary gear fault diagnosis based on permutation entropy of CEEMDAN and ANFIS," *Sensors*, vol. 18, no. 3, p. 782, 2018.

[18] H. Azami and J. Escudero, "Amplitude-aware permutation entropy: illustration in spike detection and signal segmentation," *Computer Methods and Programs in Biomedicine*, vol. 128, pp. 40–51, 2016.

[19] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, Article ID 068102, 2002.

[20] Y. Chen, T. Zhang, W. Zhao et al., "fault diagnosis of rolling bearing using multiscale Amplitude-aware permutation entropy and random forest," *Algorithms*, vol. 12, no. 9, 2019.

[21] X. Wang, S. Si, Y. Li, and X. Du, "An integrated method based on refined composite multivariate hierarchical permutation entropy and random forest and its application in rotating machinery," *Journal of Vibration and Control*, vol. 26, no. 3-4, pp. 146–160, 2020.

[22] H. Azami and J. Escudero, "Refined composite multivariate generalized multiscale fuzzy entropy: a tool for complexity analysis of multichannel signals," *Physica A: Statistical Mechanics and its Applications*, vol. 465, pp. 261–276, 2017.

[23] M. U. Ahmed and D. P. Mandic, "Multivariate multiscale entropy: a tool for complexity analysis of multichannel data," *Physical Review E*, vol. 84, no. 6, Article ID 61918, 2011.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *ieee Transactions on Pattern Analysis and Machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[25] G. B. Huang, "An insight into extreme learning machines: random neurons, random features and kernels," *Cognitive Computation*, vol. 6, no. 3, pp. 1–15, 2014.

[26] Á. Arnaiz-González, A. Fernández-Valdivielso, A. Bustillo, and L. N. López de Lacalle, "Using artificial neural networks for the prediction of dimensional error on inclined surfaces manufactured by ball-end milling," *The International Journal of Advanced Manufacturing Technology*, vol. 83, no. 5–8, pp. 847–859, 2016.

[27] X. Liu and J. Tang, "Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method," *IEEE Systems Journal*, vol. 8, no. 3, pp. 910–920, 2014.

[28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[29] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks*, Perth, Australia, August 1995.

[30] W. Deng, J. Xu, and H. Zhao, "An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem," *IEEE Access*, vol. 7, pp. 20281–20292, 2019.

[31] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, no. 95, pp. 51–67, 2016.

[32] Case Western Reserve University, "Bearing data center," 2019, http://csegroups.case.edu/%20bearingdatacenter/pages/download-data-fifile.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6661975, 8 pages
https://doi.org/10.1155/2020/6661975

Research Article

The Extraction Method of Gearbox Compound Fault Features Based on EEMD and Cloud Model

Ling Zhao[1], **Jiaxing Gong**[1] **and Hu Chong**[2]

[1]*Chongqing Jiaotong University, School of Information Science and Engineering, Chongqing 400074, China*
[2]*Chongqing Weibiao Technology Co., Ltd., Chongqing 401121, China*

Correspondence should be addressed to Ling Zhao; zhao.ling@163.com

Received 6 November 2020; Revised 26 November 2020; Accepted 4 December 2020; Published 18 December 2020

Academic Editor: Yong Chen

When a compound fault occurs, the randomness and ambiguity of the gearbox will cause uncertainty in the collected signal and reduce the accuracy of signal feature extraction. To improve accuracy, this research proposes a gearbox compound fault feature extraction method, which uses the inverse cloud model to obtain the signal feature value. First, EEMD is used to decompose the collected vibration signals of gearbox faults in normal and fault states. Then, the mutual information method is used to select the sensitive eigenmode function that can reflect the characteristics of the signal. Subsequently, the inverse cloud generator is used to extract cloud digital features and construct sample feature sets. On this basis, the concept of synthetic cloud is introduced, and the cloud-based distance measurement principle is used to synthesize new clouds, reduce the feature dimension, and extract relevant features. Finally, a simulation experiment on a rotating machinery unit with a certain type of equipment verifies that the proposed method can effectively extract the feature of gearbox multiple faults with less feature dimension. And comparing with the feature set extracted by the single cloud model, the results show that the method can better represent the fault characteristic information of the signal.

1. Introduction

Gear transmission is one of the commonly used transmission methods in mechanical equipment and is often used in high-speed trains, wind power generation, aviation, shipping, petrochemical, mining, lifting, and transportation industries. According to domestic and foreign statistics, about 10.3% of mechanical failures are caused by gearbox failure, so it is particularly important to predict and diagnose gearbox failures [1].

Due to the complex and harsh working environment of mechanical equipment, the vibration signals collected on-site are often doped with noise. To eliminate the influence of noise in the signal, a large number of researchers have carried out relevant research work in recent years. To reduce the noise in the signal, some researchers applied the wavelet denoising method to feature extraction and achieved good results [2–4]. However, this method has difficulties in selecting wavelet bases and determining thresholds in practical applications. Empirical mode decomposition (EMD) has no fixed basis, so compared with wavelet analysis methods, it solves the problem of difficult selection of wavelet basis, and it has a better processing effect on nonstationary signals than wavelet, but there is a problem with model confusion. To solve the above problems, Wu et al. [5] proposed the ensemble empirical mode decomposition (EEMD) to denoise the original signal, which overcomes the inherent mode confusion problem compared with the original EMD method. Also, there are some other methods used for fault feature extraction [5–8]. For example, Deng et al. [9] proposed an improved quantum heuristic differential evolution method to construct the best deep confidence network and propose a new fault classification method. The advantage of this method is to integrate the fault feature extraction process in the fault diagnosis algorithm.

The cloud model theory proposed by Professor Wang et al. in 1995 has been widely used in data mining [10, 11],

intelligent control [12–14], decision analysis [15, 16], intelligent transportation [17], image processing, and other fields in the past 20 years. Han et al. [18] proposed that EEMD can be combined with the cloud model to perform feature extraction of bearing faults and achieved good results, but there is a problem of more fault feature dimensions. Therefore, this article has improved based on the literature [18] and proposed a fault feature extraction method based on EEMD and synthetic cloud model, which can effectively extract fault features while avoiding difficult parameter selection problems. First, EEMD is used to decompose multiple IMF components of the vibration signal, and the mutual information method is used to select the sensitive eigenmode function that can reflect the characteristics of the signal. Subsequently, the cloud model is used to extract cloud digital features and use them as sample features. Then, the concept of synthetic cloud is introduced, the cloud similarity criterion is used to determine the choice of the base cloud, and then the number of features is reduced by synthetic cloud. Finally, by comparing with the feature sets extracted by the single cloud model, the result shows that this method can better represent the feature information of the fault signal.

2. Related Theories

2.1. EEMD Decomposition Principle. Ensemble empirical mode decomposition (EEMD) uses the statistical characteristics of Gaussian white noise with uniform time-frequency distribution to solve the problem of mode confusion, to achieve the purpose of improving EMD. It adds Gaussian white noise to the signal for multiple EMD decompositions and finally defines the overall average of the IMF decomposed multiple times as the final IMF. Based on the above, the principal steps of the EEMD algorithm are rough as follows:

(1) Initialize the overall average number M and the added noise amplitude, and set $m = 1$.

(2) Perform the mth EMD decomposition.

(1) Add white noise $n_m(t)$ of constant amplitude to the signal $x(t)$ to be analyzed;

$$x_m(t) = x(t) + n_m(t). \tag{1}$$

In the above formula, $n_m(t)$ is the white noise added for the mth time, and $x_m(t)$ is the signal after the mth noise is added.

(2) Use EMD to decompose the noised signal $x_m(t)$ to obtain a set of IMF $c_{n,m}$ $(n = 1, 2, \ldots, N)$, where $c_{n,m}$ is the nth IMF obtained from the mth decomposition

(3) If $m < M$, then return to step (1) and make $m = m + 1$. Repeat steps (1) and (2) until $m = M$.

(3) Calculate the overall average y_n of the M IMFs

$$y_n = \frac{1}{M}\sum_{m=1}^{M} c_{n,m}, \quad n = 1, 2, \ldots, M. \tag{2}$$

(4) Save the average y_n $(n = 1, 2, \ldots, N)$ of the previous N IMF decompositions as the final IMF.

2.2. Cloud Model Related Theories

2.2.1. Cloud Model. The cloud model [19] is a qualitative and quantitative conversion model proposed by the academicians Li and Du. The cloud generator can realize the mutual conversion between qualitative concepts and quantitative data. The cloud model uses expectations *Ex*, entropy *En*, and hyper-entropy *He* as digital features to represent qualitative concepts. The expected value *Ex* is the value that best represents the current qualitative concept, reflecting the information center value of the corresponding qualitative knowledge, and entropy *En* is a measure of the randomness of a qualitative concept, reflecting the degree of dispersion of cloud drops that can represent this qualitative concept. The hyper-entropy *He* is the entropy of the entropy *En*, reflecting the random degree of the numerical value belonging to the qualitative concept, and it also indirectly reflects the thickness of the cloud. As is shown in Figure 1, it is a simple cloud model ($Ex = 18$, $En = 2$, $He = 0.2$), and its ordinate μ is the degree of certainty of the cloud drop on the qualitative concept, which represents the certainty of the current cloud drop on its concept.

The above-mentioned cloud digital feature expectations *Ex*, entropy *En*, and hyper-entropy *He* are calculated using the algorithm of backward cloud algorithm [20]. The specific calculation method is as follows:

Input: N cloud drops x_i;

Output: the qualitative concept expectations *Ex*, entropy *En*, and hyper-entropy *He* represented by these N cloud drops.

(1) The estimated value of *Ex* is

$$\widehat{E}_x = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3}$$

(2) The estimated value of *En* is

$$\widehat{E}_n = \sqrt{\frac{\pi}{2}} \times \frac{1}{N}\sum_{i=1}^{N} \left|x_i - E_x\right|. \tag{4}$$

(3) The estimated value of *He* is

$$\widehat{H}_e = \sqrt{S^2 - \widehat{E}_n^2}. \tag{5}$$

The one-dimensional forward cloud algorithm is

Input: Three numerical characteristic values *Ex*, *En*, *He*, cloud drop N representing the qualitative concept $\widetilde{A}$;

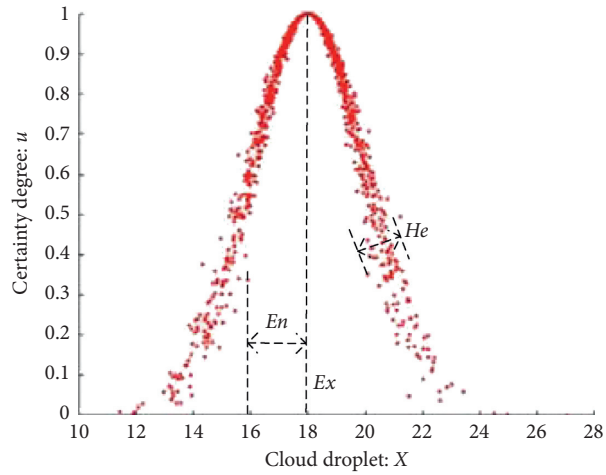


Figure 1: An example of cloud models and digital features.

Output: the quantitative value of N cloud drops, and the certainty μ that each cloud drop represents a concept $\widetilde{A}$.

(1) Generate a normal random number En' with En as the expected value and He as the standard deviation;

(2) Generate a normal random number x with Ex as the expected value and En' is the standard deviation, x is the cloud drop;

(3) Calculate $y = e^{-(x-Ex)^2/2(En')^2}$, which is the certainty of x;

(4) Repeat the above steps until N cloud drops are generated.

2.2.2. Cloud Synthesis. Cloud synthesis [21] is the process of superimposing two cloud models to obtain a comprehensive cloud model. $\widehat{C}_1 = (Ex_1, En_1, He_1)$, $\widehat{C}_2 = (Ex_2, En_2, He_2)$ are two cloud models, and a and b are two constants. According to the independent normal distribution algorithm, the synthesis method of the integrated cloud can be expressed as follows:

$$\begin{aligned} a\widehat{C}_1 + b\widehat{C}_2 &= a(Ex_1, En_1, He_1) + b(Ex_2, En_2, He_2) \\ &= \left(aEx_1 + bEx_2, \sqrt{(aEn_1)^2 + (bEn_2)^2}, \sqrt{(aHe_1)^2 + (bHe_2)^2}\right). \end{aligned} \tag{6}$$

The method of selecting the base cloud to be synthesized is based on the similarity criterion of the cloud. To consider the basic structure of the original base cloud as far as possible, the cloud similarity [19] is used as the judgment of the base cloud to be synthesized. According to the guidelines, the definition of cloud similarity is mainly described as follows:

Input: two cloud models $\widehat{C}_1(Ex_1, En_1, He_1)$ and $\widehat{C}_2 = (Ex_2, En_2, He_2)$, and the number of cloud drops n_1 and n_2;

Output: the distance between two cloud models $d(\widehat{C}_1, \widehat{C}_2)$.

(1) The two cloud models generate n_1 and n_2 cloud drops respectively through the cloud generator.

(2) Sort the cloud drops according to the abscissa from largest to smallest.

(3) Filter the cloud drops and keep the cloud drops in the range of $[Ex - 3\,En, Ex + 3\,En]$.

(4) Assuming $n_1 \le n_2$, randomly select n_2 cloud drops from n_1 cloud drops in cloud 1, and sort them in sequence, and keep them in set Drop1 and set Drop2 respectively. If $n_1 > n_2$, the same is true.

(5) Calculate the distance between each cloud drop in the two sets Drop1 and Drop2 in the corresponding order:

$$d\left(\widehat{C}_1, \widehat{C}_2\right) \approx d(\text{Drop1}, \text{Drop2}) = \frac{1}{n_2} \sum_{k=1}^{n_2} \sqrt{(x_{1k} - x_{2k})^2 + (\mu(x)_{1k} - u(x)_{2k})^2}. \tag{7}$$

In the above steps, in step 3, since the cloud satisfies the normal distribution, most of the cloud drops remain in the interval $[Ex - 3\,En, Ex + 3\,En]$, so the number of cloud drops outside the interval can be ignored. In the cloud similarity measurement, it is difficult to distinguish the similarity by setting a threshold. In this article, the distance is directly used as the similarity selection, and the two clouds with the smaller distance are selected as the base cloud to be synthesized.

2.2.3. Mutual Information Method. Mutual information (MI) can be used to describe the relationship between two random variables. It is regarded as the amount of information contained in one random variable about another random variable. The mutual information between two variables can be described as

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{8}$$

In the formula, $H(X)$ and $H(Y)$ are the entropy of variables X and Y, respectively, $H(X,Y)$ is the joint entropy of variables X and Y, and the distribution can be expressed as

$$\begin{aligned} H(X) &= -\sum_i p(x_i)\log p(x_i), \\ H(Y) &= -\sum_i p(y_i)\log p(y_i), \\ H(X,Y) &= -\sum_i \sum_j p\left(x_i, y_j\right)\log p(x_i, y_i), \end{aligned} \tag{9}$$

where $p(x)$ and $p(y)$ are the probability density functions of X and Y; $p(x,y)$ is the joint probability density function.

2.3. Feature Extraction Method Based on the EEMD Cloud Model. The cloud model is used as a composite fault signal feature characterization method. The feasibility of its cloud digital feature entropy as a fault signal feature characterization has been demonstrated by related experiments [18]. Also, cloud digital features have related applications in fault diagnosis applications [22–24]. Therefore, it is theoretically feasible to use the digital feature of the cloud model as a feature representation of the fault signal.

The cloud model can be used as a feature extraction method to obtain cloud digital features, but for gearbox multifault vibration signals, the cloud digital features obtained with a single cloud model have a high dimensionality in numbers, and some features are difficult to distinguish effectively. Therefore, this paper uses the synthetic cloud model as the feature extraction method to extract the features of the gearbox multifault vibration signal. According to the previous analysis, the feature extraction method based on EEMD and cloud model can be completed by the following steps:

(1) IMF_j $(j = 1, 2, \ldots, n)$ is obtained by decomposing the vibration signal collected by the EEMD experiment.

(2) Calculate all mutual information values between all IMF_j $(j = 1, 2, \ldots, n)$ components and the original signal. Select the sensitive IMF based on the mutual information threshold.

The threshold is determined according to reference [25].

$$u_h = \frac{\max(u_i)}{10 \times \max(u_i) - 3}, \quad i = 1, 2, \ldots, n. \tag{10}$$

In the above formula, it is the mutual information between the u_i IMF and the original signal n is the number of IMFs and $\max(u_i)$ is the maximum value of the mutual information.

(3) Keep the IMF components whose mutual information value with the original signal is greater than the threshold u_h, and delete the IMF components whose mutual information value with the original signal is less than the threshold.

(4) Perform cloud model feature extraction and transformation on the retained IMF components, synthesize the cloud into a new cloud, and calculate the cloud digital features of the new cloud as a new sample feature set.

The algorithm flow diagram of the method for extracting the fault feature of the gearbox compound fault based on the EEMD and cloud model is shown in Figure 2.

3. Experimental Verification and Result Analysis

To verify the effectiveness of the feature extraction method proposed in this paper, it is applied to the actual diagnosis of multiple faults in a certain type of equipment bearing. The experimental data [26] is collected from the rubber expansion dryer and extrusion dehydrator simulation platform of the Guangdong Petrochemical Equipment Fault Diagnosis Key Laboratory. By replacing various faulty gears, bearings, transmission shafts, and other components, the simulation cantilever centrifugal compression realized common single failures and compound failures of the engine or expander unit.

Aiming at common bearing and gear faults of complex equipment, combined with the typical industrial unit structure and load, based on the above simulation experiment platform, a set of fault accessories matching the system is designed, including bearing external cracks, bearing internal cracks, bearing ball wear, bearing lack of balls, cracked teeth, and gear wear. Some parts of the experimental failure parts are shown in Figures 3–5. Based on the above fault accessories, the test selects the NSK NN3021 bearing model for multiple fault simulation, and each fault sample is set to 40.

Based on the above fault accessories, the experiment selects NSK NN3021 bearing model for multiple fault simulation and designs 5 types of multiple fault types, namely, type 1-normal, type 2-gearbox large and small gear missing teeth + Left bearing inner ring missing the ball, type

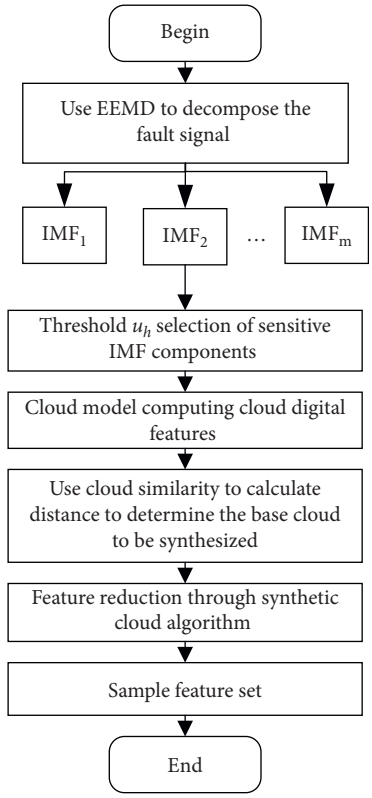


Figure 2: The algorithm flow diagram of the method for extracting the fault feature of the gearbox compound fault based on the EEMD and cloud model.

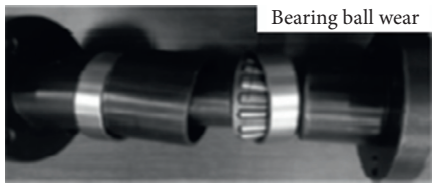


Figure 3: Bearing ball wear failure parts.

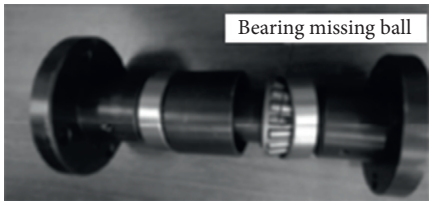


Figure 4: Bearing missing ball accessories.

Figure 5: Cracked tooth fault accessories.

3-gearbox large and small gears missing teeth + Outer ring wear on the right bearing, type 4-gearbox large and small gears missing teeth + Left bearing inner ring wear, and type 5-gearbox large and small gears missing teeth + Left bearing outer ring wear. The original signal of the five sample data is shown in Figure 6.

The EEMD parameter sets the total average time $M = 100$, and the added noise amplitude is 0.01 times the standard deviation of the original signal. After the above signal is decomposed by EEMD, 9 groups of IMF components are obtained. Usually, the most important information of the original signal is concentrated in the decomposed EEMD among the first few IMF components, as shown in Figure 7, and the MI values of IMF_1~IMF_9 and the original signal are calculated by the MI method in the five states. The abscissas in the figure represent the IMF components, the threshold is calculated by formula (10), and the thresholds are 0.1861, 0.1550, 0.1565, 0.1421, and 0.1359, respectively. It can be seen from Figure 8 that both IMF_1 and IMF_2 are higher than the corresponding threshold, and IMF_3 in type 3 is higher than the threshold, so IMF_1, IMF_2, and IMF_3 are selected as the sensitive IMF components after EEMD decomposition. To facilitate subsequent experimental simulations, IMF_4 is selected as the sensitive IMF component at the same time, so IMF_1~IMF_4 components were selected as the sensitive IMF components.

IMF_1~IMF_4, respectively, represent the first 4 sensitive IMF components selected, and the cloud digital features are calculated by formulas (3)–(5). The cloud digital feature average values of each category signal and IMF component are shown in Table 1.

For the convenience of calculation, in the paper, the clouds of IMF_1~IMF_4 components are defined as base clouds $\widehat{C}_1$~$\widehat{C}_4$. In this paper, the synthetic cloud is used to extract cloud digital features, the number of cloud drops is set to 1000, the cloud digital information obtained by IMF components is calculated by similarity to calculate the distance, and the two IMF components with the smaller distance are selected as the base cloud as the synthetic cloud algorithm. Calculate the distance between the cloud and the cloud by formula (7), and use this as the basis to determine the base cloud to be synthesized, and get $d(\widehat{C}_1, \widehat{C}_2) = 0.2642$, $d(\widehat{C}_3, \widehat{C}_4) = 0.1642$. Therefore, IMF_1 and IMF_2, IMF_3, and IMF_4 are selected as the base cloud to be synthesized. In the synthetic cloud algorithm, the value of a is set to 1, and the calculation method of the value of b is calculated as follows:

$$b_i = \frac{Ex_1}{Ex_2}. \tag{11}$$

In the above formula, b_i is the value of the coefficients of different synthetic base clouds, and Ex_1 and Ex_2 are the average expected values of the base cloud to be synthesized, where $Ex_1 > Ex_2$. Therefore IMF_1 and IMF_2, IMF_3, and IMF_4 are, respectively, used as base clouds to perform synthetic cloud, and calculation of the digital characteristics of the synthetic cloud is shown in Table 2.

As the final pattern recognition algorithm, there are many classifier algorithms, such as the literature [9, 27, 28], and the proposed method boosts the classification

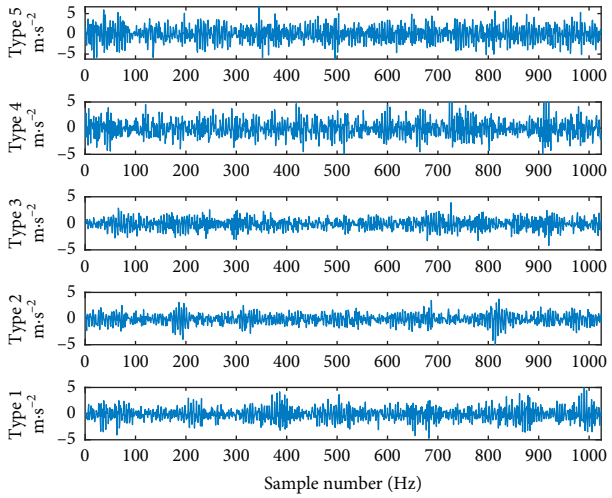


Figure 6: 5-state source signal diagram.

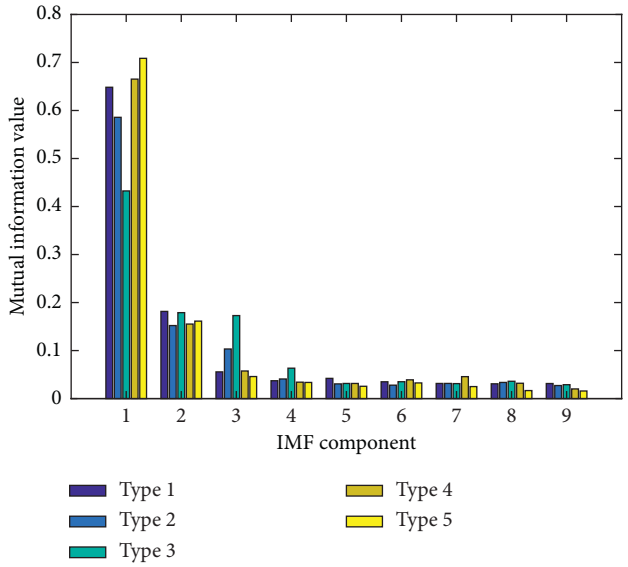


Figure 7: MI value between IMF_1~IMF_9 and original signal in five states.

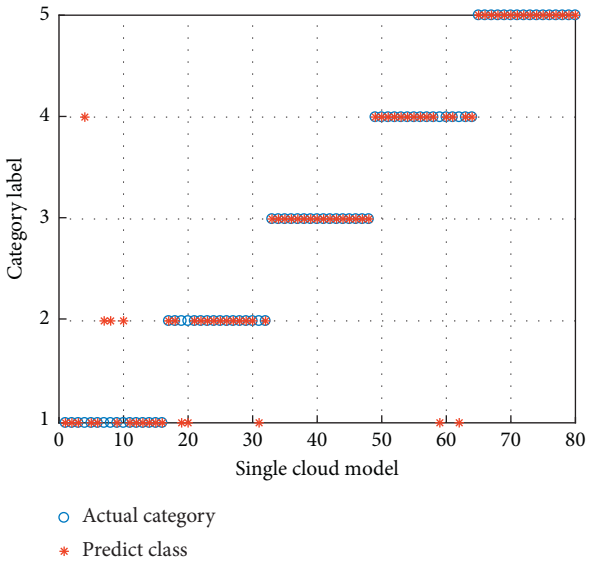


Figure 8: Single cloud model.

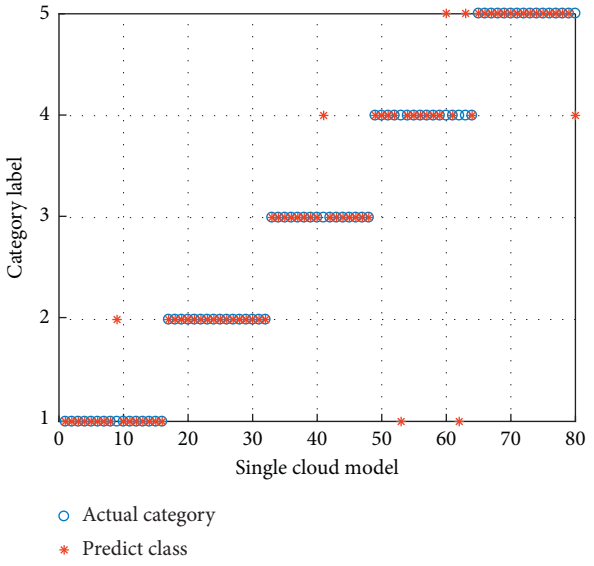


Figure 9: Synthetic cloud model.

performance across the classes of the data. Since the fault sample data is small, considering the time efficiency issue, this paper directly uses the support vector machine as the classifier for experimental verification. For the calculated synthetic cloud digital features, 200 samples were selected from the samples at a ratio of 6 : 4, as 120 samples were used for training and 80 samples were used for testing. In the support vector machine (SVM) algorithm, the penalty factor $C = 150$, $\sigma = 1$, the experimental results are shown in Table 3, and the test classification effects of the two methods are shown in Figures 7 and 9. The results show that, in the feature extraction method of the EEMD synthetic cloud model, compared with the single cloud model, the feature dimension is reduced, and the degree of discrimination is also improved.

From Figure 7 and Table 3, it can be seen that the cloud digital features extracted by the single cloud model are used as the fault feature extraction method to verify that the classification accuracy is up to 88.25%, which verifies that the cloud model as a method for extracting composite fault features is reliable and effective. It can be seen from Figure 9 and Table 3 that the synthetic cloud model feature extraction method proposed in this paper has a verification classification accuracy of 91.25%. At the same time, analyzing Figure 7 shows that fault category 1 and fault category 2 in the single cloud model are prone to misdiagnosis. Analysis of Figure 9 shows that fault categories 2 and 4 in the synthetic cloud model have fault identification phenomena. In the synthetic cloud algorithm, the choice of parameters will also directly affect the category of features, so it depends on the situation. But overall, in terms of feature dimension and

Table 1: Single cloud model characteristics.

Single cloud model	Type 1 (*Ex*, *En*, *He*)	Type 2 (*Ex*, *En*, *He*)	Type 3 (*Ex*, *En*, *He*)	Type 4 (*Ex*, *En*, *He*)	Type 5 (*Ex*, *En*, *He*)
IMF_1	−0.0079	0.0031	−0.0029	0.0099	−0.0088
	0.9095	0.9678	0.7992	0.8819	1.8971
	0.0707	0.0053	0.0036	0.0031	—
IMF_2	−0.0016	0.000145	−0.00063	−0.0013	0.0017
	0.4378	0.4501	0.3933	0.4161	0.8557
	0.1680	0.1294	0.0812	0.1069	0.2029
IMF_3	0.0017	0.00037	−0.0040	−0.0007732	0.0005829
	0.2791	0.3269	0.4423	0.2785	0.2540
	0.0575	0.0397	0.1109	0.0523	0.0359
IMF_4	−0.0014	−0.0016	0.0002721	0.0001742	0.0014
	0.1458	0.2288	0.1364	0.1758	0.2702
	0.0297	0.0270	0.0177	0.0157	0.0426

Table 2: Synthetic cloud model characteristics.

Synthetic cloud model	Type 1 (*Ex*, *En*, *He*)	Type 2 (*Ex*, *En*, *He*)	Type 3 (*Ex*, *En*, *He*)	Type 4 (*Ex*, *En*, *He*)	Type 5 (*Ex*, *En*, *He*)
$IMF_1 + IMF_2$	−0.0083	0.0029	0.0031	−0.00017	−0.0028
	0.9308	0.2452	0.9891	0.2909	0.8187
	0.0869	0.0485	0.0060	0.0347	0.0042
$IMF_3 + IMF_4$	−0.0028	0.0096	−0.000415	−0.0085	−0.00098
	0.3890	0.9018	0.1957	1.9362	0.4094
	0.1087	0.0046	0.0366	—	0.0678

Table 3: Comparison of the two feature extraction methods.

Methods	Feature dimension	Accuracy (SVM)
Synthetic cloud model	6	90% ± 1.25%
Single cloud model	12	87% ± 1.25%

classification accuracy, the synthetic cloud model method and the single cloud model fault feature extraction method have certain advantages.

4. Conclusions

This paper proposes a feature extraction method for gearbox composite fault signals based on EEMD and synthetic cloud model. The EEMD algorithm is used for signal decomposition and then uses the mutual information method to select the sensitive IMF to obtain the feature information. Then, the concept of synthetic cloud is introduced, and the cloud-based distance measurement principle is used to select the cloud to be synthesized, synthesize the new cloud, and reduce the number of features at the same time, and relevant features are extracted. Finally, use the actual composite fault data set for verification and compare it with the feature set extracted by the single cloud model. Also, the time complexity of the method proposed in this article mainly depends on the choice of parameters. There are mainly several parameters to determine including the number of decomposition k in the EEMD algorithm, and the similarity distance of the cloud model. In terms of judgment, the number of cloud drops needs to be calculated, and the number of generated cloud drops determines the timeliness of the entire algorithm. In practical applications, the method proposed in this paper is mainly determined by the number of cloud drops, depending on the scale of the data. The experimental results prove that this method is effective and superior to the single cloud model fault extraction method, which has certain engineering practical application significance.

Data Availability

The fault-related database used for this research can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Darong Huang for his comments and suggestions. This work was supported in part by the National Natural Science Foundation of P.R. China under Grants (62073051, 61304104); the Science and Technology Research Project of the Chongqing Municipal Education Commission of P.R. China under Grants (KJZD-K 201900704); Chongqing Technology Innovation and Application Special Key Project under Grants cstc2019jscx-

mbdxX0015; and the Innovation Foundation of Chongqing Postgraduate Education under Grants CYS20282.

References

[1] Z. Wang, J. Wang, and W. Du, "Research on fault diagnosis of gearbox with improved variational mode decomposition," *Sensors*, vol. 18, no. 10, p. 3510, 2018.

[2] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," *Journal of Sound and Vibration*, vol. 234, no. 1, pp. 135–148, 2000.

[3] P. Liang, C. Deng, J. Wu, Z. Yang, J. Zhu, and Z. Zhang, "Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform," *Computers in Industry*, vol. 113, Article ID 103132, 2019.

[4] W. Deng, J. Xu, Y. Song, and H. Zhao, "Differential evolution algorithm with wavelet basis function and optimal mutation strategy for complex optimization problem," *Applied Soft Computing*, Article ID 106724, 2020.

[5] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[6] H. Zhao, S. Zuo, M. Hou et al., "A novel adaptive signal processing method based on enhanced empirical wavelet transform technology," *Sensors*, vol. 18, no. 10, p. 3323, 2018.

[7] X. Lyu, Z. Hu, H. Zhou, and Q. Wang, "Application of improved MCKD method based on QGA in planetary gear compound fault diagnosis," *Measurement*, vol. 139, pp. 236–248, 2019.

[8] W. Cai, Z. Yang, Z. Wang, and Y. Wang, "A new compound fault feature extraction method based on multipoint kurtosis and variational mode decomposition," *Entropy*, vol. 20, no. 7, p. 521, 2018.

[9] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, 2020.

[10] S. Wang, D. Li, W. Shi et al., "Cloud model-based spatial data mining," *Geographic Information Sciences*, vol. 9, no. 1-2, pp. 60–70, 2003.

[11] S. N. Rao and P. R. Kumar, "Time series data mining in cloud model," in *Proceedings of the International Conference on E-Business and Telecommunications*, pp. 237–244, Springer, Prague, Czech Republic, July 2019.

[12] G. Hongbo, X. Guotao, L. Hongzhe, Z. Xinyu, and L. Deyi, "Lateral control of autonomous vehicles based on learning driver behavior via cloud model," *The Journal of China Universities of Posts and Telecommunications*, vol. 24, no. 2, pp. 10–17, 2017.

[13] X. Xiang, A. Luo, and Y. Li, "Intelligent control method of power supply for tundish electromagnetic induction heating system," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, 2020.

[14] M. Tao, R. Qu, and Z. Ke, "The cloud model theory of intelligent control method for non-minimum-phase and non-self-balancing system in nuclear power," in *Proceedings of the International Conference on Nuclear Engineering*, p. 51432, American Society of Mechanical Engineers (ASME), London, UK, July 2018.

[15] H. Tang, M. Lei, Q. Gong, and J. Wang, "A BP neural network recommendation algorithm based on cloud model," *IEEE Access*, vol. 7, pp. 35898–35907, 2019.

[16] H. Ren, Y. Yan, T. Zhou, and X. Xiang, "Evaluation on cooperative partners in organization coalition for mega projects based on cloud model and gray correlation analysis," *China Civil Engineering Journal*, vol. 44, no. 8, pp. 147–152, 2011.

[17] Z. Deng, D. Huang, J. Liu, B. Mi, and Y. Liu, "An assessment method for traffic state vulnerability based on a cloud model for urban road network traffic systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2020.

[18] L. Han, C. Li, and H. Liu, "Feature extraction method of rolling bearing fault signal based on EEMD and cloud model characteristic entropy," *Entropy*, vol. 17, no. 10, pp. 6683–6697, 2015.

[19] D. Li and Y. Du, *Artificial Intelligence with Uncertainty*, CRC Press, Boca Raton, FL, USA, 2007.

[20] C. Liu, M. Feng, X. Dai, and D.-Y. Li, "A new algorithm of backward cloud," *Acta Simulata Systematica Sinica*, vol. 11, 2004.

[21] S. D. Xu, X. L. Geng, and X. Q. Dong, "Improved FMEA approach for risk evaluation based on cloud model," *Computer Engineering and Applications*, vol. 54, no. 2, pp. 228–233, 2018.

[22] Y. Zhang, D. N. Zhao, and D. Y. Li, "The similar cloud and the measurement method," *Information and Control*, vol. 33, no. 2, pp. 129–132, 2004.

[23] Y. Jiang, C. Tang, X. Zhang, W. Jiao, G. Li, and T. Huang, "A novel rolling bearing defect detection method based on bispectrum analysis and cloud model-improved EEMD," *IEEE Access*, vol. 8, pp. 24323–24333, 2020.

[24] H. J. Xu, Z. Y. Wang, and H. Y. Su, "Dissolved gas analysis based feedback cloud entropy model for power transformer fault diagnosis," *Power System Protection and Control*, vol. 41, pp. 115–119, 2013.

[25] A. Ayenu-Prah and N. Attoh-Okine, "A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition," *Advances in Adaptive Data Analysis*, vol. 2, no. 1, pp. 1–24, 2010.

[26] Z. Xu, H. Darong, G. Sun, and W. Yongchao, "A fault diagnosis method based on improved adaptive filtering and joint distribution adaptation," *IEEE Access*, vol. 8, pp. 159683–159695, 2020.

[27] R. Chen, S.-K. Guo, X.-Z. Wang, and T.-L. Zhang, "Fusion of multi-RSMOTE with fuzzy integral to classify bug reports with an imbalanced distribution," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 12, pp. 2406–2420, 2019.

[28] W. Deng, J. Xu, X.-Z. Gao, and H. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, p. 1, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6682216, 9 pages
https://doi.org/10.1155/2020/6682216

Research Article

A Short-Term Traffic Flow Reliability Prediction Method considering Traffic Safety

Shaoqian Li,[1] Zhenyuan Zhang ,[1] Yang Liu,[1] and Zixia Qin[2]

[1]*Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[2]*School of Design and Arts, Beijing Institute of Technology, Beijing 100081, China*

Correspondence should be addressed to Zhenyuan Zhang; zzhenyuan@cqjtu.edu.cn

Received 4 November 2020; Revised 30 November 2020; Accepted 30 November 2020; Published 10 December 2020

Academic Editor: Yong Chen

With the rapid development and application of intelligent traffic systems, traffic flow prediction has attracted an increasing amount of attention. Accurate and timely traffic flow information is of great significance to improve the safety of transportation. To improve the prediction accuracy of the backward-propagation neural network (BPNN) prediction model, which easily falls into local optimal solutions, this paper proposes an adaptive differential evolution (DE) algorithm-optimized BPNN (DE-BPNN) model for a short-term traffic flow prediction. First, by the mutation, crossover, and selection operations of the DE algorithm, the initial weights and biases of the BPNN are optimized. Then, the initial weights and biases obtained by the aforementioned preoptimization are used to train the BPNN, thereby obtaining the optimal weights and biases. Finally, the trained BPNN is utilized to predict the real-time traffic flow. The experimental results show that the accuracy of the DE-BPNN model is improved about 7.36% as compared with that of the BPNN model. The DE-BPNN is superior to the performance of three classical models for short-term traffic flow prediction.

1. Introduction

In recent years, with the development of traffic detection technology, big data technology, and data mining technology, accurate and real-time traffic flow operation data and traffic accident data are easy to collect [1]. By studying the changing characteristics of traffic flow before and after traffic accidents, the traffic safety status is analyzed, evaluated, and forewarned according to the collected real-time traffic flow data. Real-time and dynamic release of early warning information can adjust and control traffic flow parameters in time, greatly reduce the occurrence of traffic accidents and the degree of accident damage, and thus improve the operation efficiency of the expressway network. Among them, the intelligent prediction of traffic flow plays a key role in various technologies. Short-term traffic flow forecasting is the most valuable practice in traffic application, and it is the foundation and basis for the realization of advanced traffic management system and traffic information service system. The accuracy of short-term traffic flow prediction directly affects the effects of traffic flow guidance and traffic control, which is of great significance for maintaining traffic safety. Real-time and accurate traffic flow prediction is the premise and key to the realization of both traffic flow guidance systems and traffic control systems [2].

As a matter of fact, short-term traffic flow prediction largely relies on the historical and real-time traffic data collected through various sensors (e.g., induction coils, radar, cameras, mobile global positioning systems, and social media) to build corresponding models and algorithms. The fundamental principle of short-term traffic flow prediction is as follows: first, a rational prediction model is built by a dedicated structure and parameters based on a certain amount of sensor data such as the historical traffic flow, vehicle density, and vehicle speed; then, the prediction model is trained by the corresponding learning algorithm based on the collected data to obtain a set of optimal solutions; finally, the traffic sensor data to be determined are fed back into the trained model to predict the future traffic flow. By principle, existing short-term traffic flow prediction

methods are broadly classified into parametric methods, nonparametric methods, and simulation approaches [3]. Parametric methods principally include time series models, Kalman filtering, autoregression, and exponential smoothing. Nonparametric methods include *K*-nearest neighbor methods, support vector machines (SVMs), and artificial neural networks. Simulation approaches predict traffic flow using existing traffic simulation tools. These classical short-time traffic flow prediction methods offer favourable results for theoretical analysis and simulation. Unfortunately, their application in practical engineering scenarios is greatly limited due to the explosive growth of traffic big data.

Accordingly, to solve the data explosion problem arising from the explosive growth of traffic data, many short-term traffic flow prediction methods based on improving the parameters of the aforementioned classical models have been developed. For instance, based on a classical parametric prediction method, the seasonal autoregressive integrated moving average (ARIMA) model, Williams and Hoel [4] built a short-term traffic flow prediction model by considering the impact of seasonal factors on road traffic flow. Furthermore, considering the influencing factors of affect spatiotemporal correlations such as the road network topology and time-varying speed, Duan et al. [5] proposed a spatiotemporal model based on the Space-Time Autoregressive Integrated Moving Average (STARIMA) model, which further enhances the accuracy of short-term traffic flow prediction. However, it is difficult to achieve high traffic flow prediction accuracy with the limited small samples. Therefore, Kumar [6] assumed a linear traffic flow and attached great importance to the temporal correlation of traffic flow at a particular location with a relatively stable traffic flow, thereby proposing a Kalman filter-based prediction scheme, which improves the prediction accuracy using small samples. However, such methods neglect the impacts of complex and changeable actual traffic environment challenges, such as spatiotemporal interaction and coupling.

To solve this problem, building new short-term traffic flow prediction models by combining models based on nonparametric prediction methods is considered to be a solution for short-term traffic flow prediction [7, 8]. Duo et al. [9] optimized the parameters of an SVM and built a short-term traffic flow prediction model by decomposing the traffic flow sequence into different frequency components and then introducing the crossover and mutation factors of the genetic algorithm into particle swarm optimization (PSO). Dai et al. [10] proposed a gated recurrent unit-(GRU-) based short-term traffic flow prediction model based on an analysis of the spatiotemporal characteristics of traffic flow data. Chen et al. [11] attempted to build a number of prediction models with different time delays to propose the least squares support vector regression (LSSVR) based short-term traffic flow prediction model and to achieve a favourable prediction performance. Zhao et al. [12] proposed a hybrid model by combing *K*-nearest neighbor (KNN) with support vector regression (SVR), imitating the KNN search mechanism to rebuild a historical traffic flow sequence Unfortunately, since the road network traffic system is affected by uncertainties such as the road traffic environment, weather conditions, and pedestrians, the actual traffic flow data are evidently nonlinear, time-varying, and susceptible to random noise. Therefore, the above traffic flow prediction methods are not suitable for short-term traffic flow prediction in complex conditions because they are still limited by dedicated model parameters, low prediction accuracy, and poor generalization. So, exploring more effective methods to achieve higher short-term traffic flow prediction accuracy has become a great concern for traffic researchers.

Recently, deep learning-based methods, such as backward-propagation neural networks (BPNNs), have been successfully applied to various tasks in traffic flow prediction. Some scholars have introduced artificial neural networks with many hidden layers to build short-term traffic flow prediction models [13–15], which achieve better prediction performance. However, BPNNs have two obvious shortcomings, including a high involvement in local optimal solutions and a low convergence rate. Meanwhile, these models lack the interpretability of the learning process. Hence, how to optimize the structural parameters of a BPNN and building a practicable short-term traffic flow prediction model is the main focus of this article.

For the above problems, based on the influence of traffic volume and traffic safety, this paper proposes an adaptive differential evolution (DE) algorithm-optimized BPNN (DE-BPNN) model for short-term traffic flow prediction. First, the DE algorithm is used for heuristic random optimization of the group difference of the BP parameters based on a brief description of the BPNN to make up for the random defects of the BPNN in terms of the initial weight and bias selection. Second, to accelerate the convergence rate in short-term traffic flow prediction, the BPNN is combined with the DE algorithm to build a novel short-term traffic prediction model for global optimization and generalization of short-time traffic flow. Finally, simulation verification is performed for the proposed algorithms and models using the standard data set collected by the Caltrans Performance Measurement System (PeMS), USA. The simulation results show that the proposed algorithm has a better learning ability and global optimization performance compared to conventional algorithms such as ARIMA, wavelet neural networks (WNNs), and BPNNs.

2. Traffic Flow Prediction Modelling Based on DE-BPNN

In general, traffic flow prediction can be classified by the prediction period into long-term prediction, medium-term prediction, and short-term prediction. In fact, once travellers learn the evolution trend of short-term traffic flow in real-time, they can change their routes for fast, convenient, and comfortable travel. Therefore, travellers extremely concern about short-term traffic flow prediction. In actual traffic environments, the 5- to 30-minute traffic flow evolution trend is chosen as the time range for short-term traffic flow prediction. To ensure the prediction accuracy, the traffic flow data sequence observed within n identical time intervals at an observation point in the traffic network is assumed to be

$\{x_i\}, i = 1, 2, \ldots, n$, and the predicted traffic flow of a certain period in the future is y. With a rational traffic flow prediction model, it is possible to observe the traffic flow data sequence for accurate prediction within a short time, thereby providing an effective decision-making basis for travellers to choose their travel routes. This paper adopts the DE-BPNN-based model for the optimization and improvement of accuracy. For the simplicity of analysis and the integrity of the overall frame structure, the basic structure of the BPNN-based short-term traffic flow prediction model is introduced first.

2.1. Overview of the BPNN-Based Short-Term Traffic Flow Prediction Model. It is well known that the traffic system is a large, complex, nonlinear, time-varying, and stochastic system. BPNN can identify complex nonlinear systems and constantly adjust the parameters based on a large number of collected data sequences. Moreover, it can approximate any nonlinear continuous function with an arbitrary precision through the deep data fusion of parallel structures and the data processing capability of self-learning. Such properties help to remarkably reduce the computing workload of online prediction. Therefore, BPNN-based methods are widely applied in the field of short-term traffic flow prediction [16, 17]. Generally, the BPNN structure includes an input layer, a hidden layer, and an output layer. Each layer is connected by weights and bias. The weights and bias value range is typically $[-1, 1]$. As shown in Figure 1, a neuron model contains an input layer with n nodes, an intermediate layer with m nodes, and an output layer with 1 node. In short-term traffic flow prediction, the processing procedure basic execution process consists of the forward propagation of traffic information and the backward propagation of error, as shown in Figure 2.

It is assumed that x_i represents the traffic flow of an observation point in the traffic network in the i^{th} time interval; the input is $(x_1, x_2, \ldots, x_n)^T$, and the output is y.

$$\begin{cases} \text{net}_j = \sum_{i=1}^{n} \omega_{ij} * x_i + \theta_j, \\ y_j = f_j(\text{net}_j), \\ \text{net} = \sum_{j=1}^{m} v_j * y_j + \theta, \\ y = f(\text{net}), \end{cases} \quad j = 1, 2, \ldots m, \tag{1}$$

where w_{ij} and v_j are the connection weight between the input-hidden layers and hidden-output layers, respectively; θ_j and θ are the biases of the hidden layer and output layer, respectively; $f_j(\cdot)$ and $f(\cdot)$ are the activation functions of the hidden layer and output layer, respectively.

The BPNN weights and biases can generate three matrices and one bias, including, the weight matrix W from the input layer to the hidden layer, the weight matrix V from the hidden layer to the output layer, and the bias matrix T of the hidden layer and the output layer bias θ. Each matrix is represented as follows:

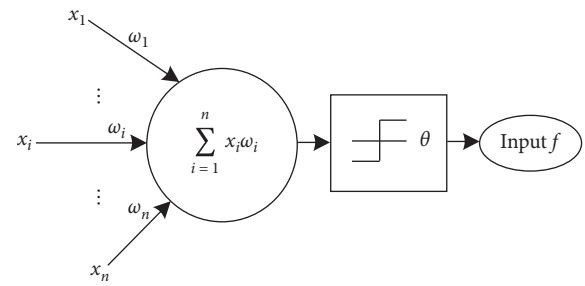


Figure 1: Neuron model structure.

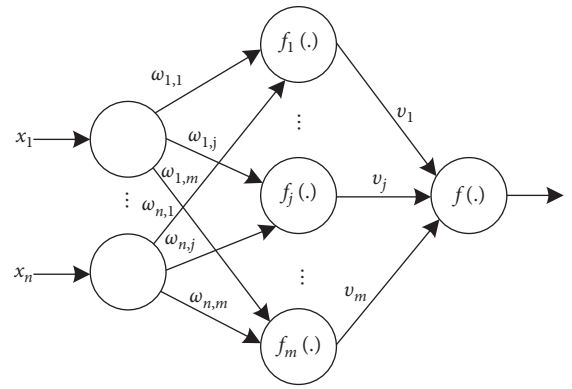


Figure 2: Single hidden layer BPNN structure diagram.

$$W = \begin{bmatrix} \omega_{1,1} & \omega_{1,2} & \cdots & \omega_{1,n} \\ \omega_{2,1} & \omega_{2,2} & \cdots & \omega_{2,n} \\ \vdots & \vdots & & \vdots \\ \omega_{m,1} & \omega_{m,2} & \cdots & \omega_{m,n} \end{bmatrix}, \tag{2}$$
$$V = [v_1, v_2, \ldots, v_m],$$
$$T = [\theta_1, \theta_2, \ldots, \theta_m]^T.$$

Compared with classical traffic flow prediction algorithms, the BPNN has obvious superiorities, such as higher prediction accuracy and noise robustness. However, such a black box-like learning mode still faces several challenges [7]. Firstly, since the training process utilizes the current real-time data, the trained network may be no longer applicable when the traffic flow changes. Secondly, the connection weights and bais of each layer in BPNN are set randomly, which may make the training process fall into local minimization. To alleviate these challenges, this paper relies on the DE algorithm to optimize BPNN parameters, resulting in faster convergence, simpler implementation, and higher prediction accuracy.

2.2. Improvement of the Model by DE Algorithm-Based BPNN Parameter Optimization. As a group difference-based heuristic global search algorithm, the DE algorithm optimizes the distribution of the weights and biases for each layer in BPNN through real number encoding. During the iteration process, the optimal weights and biases are chromosomes obtained by the assistance of the DE algorithm. In

the end, the local optimization of the BPNN-based traffic flow prediction model ultimately leads to the global optimal solution. Conventionally, the DE algorithm employs a different strategy for mutation operations, that is, the difference vector between individuals in the population to perturb individuals to achieve individual variation. The mutation method of the DE algorithm improves the search capacity by using the characteristics of the population distribution effectively.

In fact, the DE optimization process adopts population initialization, mutation, crossover, and selection strategies to map matrices, including W, V, T, and θ, into the chromosome. The mapping relationship is expressed as follows:

$$\left\{\omega_{1\ ,\ 1}, \omega_{1\ ,\ 2}, \ldots, \omega_{n,m}, v_1, v_2, \ldots, v_m, \theta_1, \theta_2, \ldots, \theta_m, \theta\right\}, \tag{3}$$

where the set of mappings is D, which represents the number of dimensions of a variable. The initial variables are calculated and assigned based on equation (4), and the initial population $\left\{z_i(0) | z_{i.j}^L \le z_{i.j}(0) \le z_{i.j}^U, i = 1, 2, \ldots, \mathrm{Np}, j = 1, 2, \ldots, D\right\}$ is generated randomly as given in the following equation.

$$z_{i,j}(0) = z_{i.j}^L + \mathrm{rand}(0,1)\left(z_{i.j}^U - z_{i.j}^L\right), \tag{4}$$

where $z_i(0)$ represents the i^{th} individual of the 0^{th} generation in the population; $z_{i.j}(0)$ represents the j^{th} "gene" (number of dimensions) of the i^{th} individual in the 0^{th} generation in the population; $z_{i.j}^L$ and $z_{i.j}^U$, respectively, represent the minimum and maximum numbers of dimensions of the individual; Np represents the population size; and $\mathrm{rand}(0,1)$ is a random number uniformly distributed in the interval $(0,1)$.

A mutation operation is performed to achieve individual mutation through a differential strategy. Three different individuals are randomly selected from Np numbers of individuals; two of them are scaled by the vector difference, and another vector is added thereto, that is,

$$v_i(g+1) = z_{r1}(g) + F\left(z_{r2}(g) - z_{r3}(g)\right), \tag{5}$$

where $z_i(g)$ represents the i^{th} individual in the g^{th} generation population; $r1, r2, r3$, and target vector i are different from each other; and F is a scale factor, which has been assigned a value between $[0, 2]$ and used to control the scaling of differential variables [18, 19].

The crossover operation is performed between individuals for the g^{th} generation population $\{z_i(g)\}$ and the intermediate $\{v_i(g+1)\}$

$$u_{i.j}(g+1) = \begin{cases} v_{i.j}(g+1), if\, \mathrm{rand}(0,1) \le \mathrm{CR}\ \mathrm{or}\ j = j_{\mathrm{rand}}, \\ z_{i.j}(g), \mathrm{otherwise}, \end{cases} \tag{6}$$

where CR is the crossover factor and j_{rand} is a random integer in $[1, 2, \ldots, D]$. To ensure at least one "gene" in the intermediate individual is passed to the next generation, the $j_{\mathrm{rand}}^{\mathrm{th}}$ gene of each individual is passed in the first crossover "gene" operation. $z_i(g)$ or $v_i(g+1)$ is chosen as the allele of $u_i(g+1)$, which depends on the CR probability for subsequent crossover operations.

The selection strategy focuses on the population selection after the crossover operation in the differential algorithm. Based on the complexity of the actual traffic environment, the DE algorithm employs the greedy algorithm to choose the individuals inputting to the next generation of the traffic flow population; that is,

$$z_i(g+1) = \begin{cases} u_i(g+1), & f\left(u_i(g+1)\right) \le f\left(x_i(g)\right), \\ x_i(g), & \mathrm{otherwise}. \end{cases} \tag{7}$$

For the DE algorithm, the mutation, crossover, and selection operations are continuously performed through equations (5)–(7) until meeting the conditions. Thus, the DE-BPNN parameter optimization is achieved. Note that, the scale and crossover factors of the DE algorithm are fixed values based on experience. In actual traffic flow prediction, this algorithm requires a wider search range to avoid becoming trapped in local optimal solutions in the early stage, while it requires a narrower search range to prevent the algorithm from missing the extreme points in the later stage. Therefore, to conform to the dynamics of short-term traffic flow, it is necessary to improve the method of determining the scale and crossover factors.

2.3. Method of Determining the Dynamics of Adaptive Scale and Crossover Factors for DE Algorithm. To ensure the accuracy and effectiveness of short-term traffic flow prediction, the established adaptive scale factors need to guarantee the following characteristics: as the number of iterations increases, the mutation rate should gradually decrease; at the beginning of an iteration, a larger scale factor should be selected to increase the diversity of the traffic flow population; a smaller scale factor should be selected in the later stage to preserve the superior individuals of the traffic flow population. Based on these considerations, adaptive scale factors are generated with the following equation:

$$F = F_0 * \widehat{2}\, e^{1-\left(G_m/G_m + (1-G)\right)}, \tag{8}$$

where F_0 represents the initial scale factor, G_m is the maximum number of iterations, and G stands for the current number of iterations.

Similarly to the scale factor, as the number of iterations increases, the crossover rate changes dynamically. The larger crossover factor at the initial stage ensures the global traffic flow state mutation. The smaller crossover rate in the later period is more focused on local traffic state convergence. Hence, the design adaptive crossover factor is shown in the following equation:

$$\mathrm{CR} = \mathrm{CR}_{\max} - \left(\frac{G\left(\mathrm{CR}_{\max} - \mathrm{CR}_{\min}\right)}{G_m}\right), \tag{9}$$

where $\mathrm{CR}_{\max}$ is the maximum value of the crossover parameter and $\mathrm{CR}_{\min}$ represents the minimum value of the crossover parameter.

In regard to traffic flow prediction, because the DE algorithm optimizes the BP parameters, it can prevent the

BPNN from being trapped in local optimal solutions and improve the accuracy of the BPNN-based traffic flow algorithm.

3. Design of the Short-Term Traffic Flow Prediction Algorithm Based on Adaptive DE-BPNN

The preceding section introduces the BPNN-based short-term traffic prediction model and the DE algorithm, as well as the method for determining the dynamic parameter factors in the DE algorithm. The basic procedure of the hybrid short-term traffic flow prediction model algorithm is given below, as shown in Figure 3.

Step 1: initialize the population. N_p populations are randomly initialized based on equation (4). Each individual has D dimensions, each of which represents a parameter in a neural network.

Step 2: the error between the neural network output and actual values is defined as the population-dependent fitness function. The fitness of each individual is calculated, and the minimum fitness value is the global minimum; the globally optimal individual is updated. The fitness function can be represented by the mean square error (MSE) or the root mean square error (RMSE) as given in the following equation:

$$\begin{aligned} \text{MSE} &= \frac{1}{N_{ts}} \sum_{i=1}^{N_{ts}} \left(\widetilde{y}_i - \widehat{y}_i\right)^2, \\ \text{RMSE} &= \sqrt{\frac{1}{N_{ts}} \sum_{i=1}^{N_{ts}} \left(\widetilde{y}_i - \widehat{y}_i\right)^2}, \end{aligned} \tag{10}$$

where N_{ts} represents the number of trained samples; $\widetilde{y}_i$ is the actual value of the predicted traffic flow; and $\widehat{y}_i$ stands for the value of the actual traffic flow.

Step 3: the next-generation individual $x_i(g+1)$ is obtained based on the DE mutation, crossover, and selection operations.

Step 4: step 3 is repeated until the next-generation population is obtained.

Step 5: determine whether or not the termination condition (the global minimum meets the predefined accuracy requirements, or the maximum number of iterations is reached) is met; if yes, the iteration will be stopped with the optimal individuals as parameters of the neural network; otherwise, go to the next step.

Step 6: if $g = g + 1$; go back to Step 2.

Step 7: enter the test set and perform prediction with the trained network.

4. Simulation Experiment

4.1. Experiment Conditions. To verify the performance of the proposed short-term traffic flow prediction model, the PeMS data set, one of the most commonly used data set in short-term traffic flow prediction [20], was selected. The data acquired from one road segment were chosen for detection; the data collected by these detectors were summarized once in every 5 minutes, and the traffic flow was summarized for three one-way lanes. Then, the data were subjected to preprocessing, including the removal of redundant data, the correction of erroneous data, and normalization. Figure 4 shows the traffic flow at a detection point along a freeway overtime during the week. It can be seen that the traffic flow on weekdays tends to be consistent. To more closely assess the similarity of weekdays, the daily traffic flow was compared in the same plane. As shown in Figure 5, the morning and evening rush hours are almost at the same time on the weekdays, which reflects the consistency of travel patterns. Hence, to ensure the accuracy of the prediction, weekends and weekdays are distinguished, and the experimental study was conducted on weekday traffic flow data collected from May 2 to May 6, 2018. The data collected on May 2 and 3, 2018, were used as the training set, while the data acquired on May 4 were used as the prediction set.

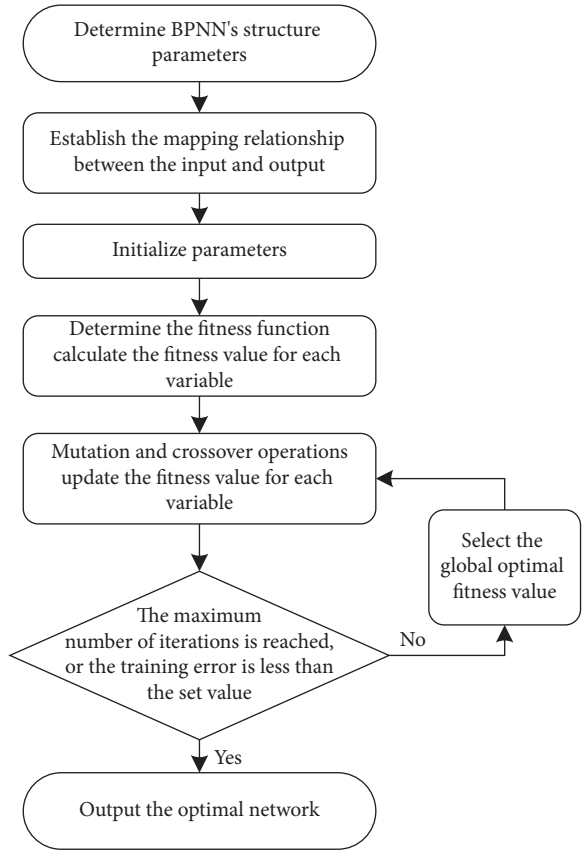


Figure 3: DE-BPNN algorithm flowchart.

4.2. Assessment Indicators. To evaluate the effectiveness of the DE-BPNN model and some conventional models, the four most commonly used performance indicators were selected for regression problems: the mean absolute error (MAE), the MSE, the mean absolute percentage error

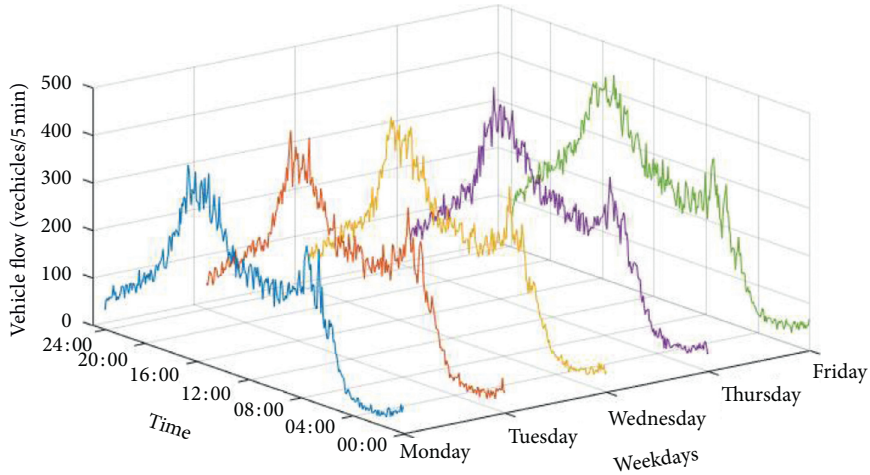


Figure 4: Correlation of time series on weekdays.

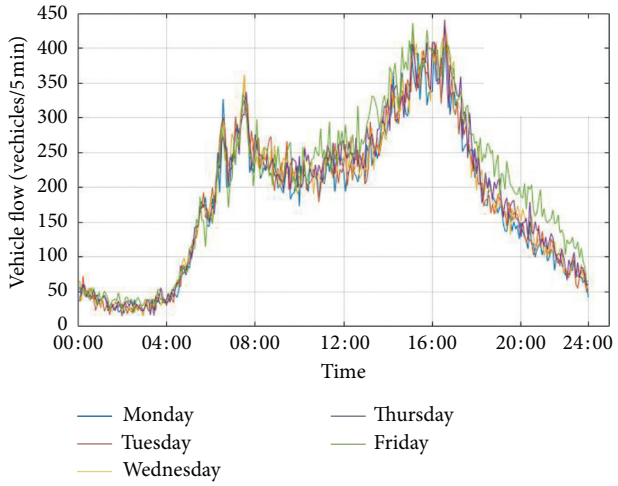


Figure 5: Typical traffic flow pattern on weekdays.

(MAPE), and the mean square percent error (MSPE). All indicators are defined as follows: y_t and $\hat{y}_t$ represent the detection value and the model prediction value of traffic flow, respectively [21, 22].

(1) Mean absolute error (MAE): it can avoid the problem of mutual cancellation of errors, so it can well reflect the actual situation of predicted value errors.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right|. \tag{11}$$

(2) Mean square error (MSE): it is a measure reflecting the difference between the estimated quantity and the estimated quantity, which can evaluate the change degree of the data. The smaller the MSE value, the better the accuracy of the prediction model in describing the experimental data:

$$\text{MSE} = \frac{1}{n}\sqrt{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}. \tag{12}$$

(3) Mean absolute percentage error (MAPE): it is used as a statistical indicator to measure the accuracy of prediction. A smaller MAPE indicates a better model effect.

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|. \tag{13}$$

(4) Mean square percent error (MSPE): it is used to test the degree of model fitting. The smaller the MSPE is, the better the fitting degree is, and the model can be accepted.

$$\text{MSPE} = \frac{1}{n}\sqrt{\sum_{t=1}^{n}\left(\frac{y_t - \widehat{y}_t}{y_t}\right)^2}. \quad (14)$$

4.3. Model Parameters. The experiment used the BPNN-based method with a single hidden layer ($m - 2m + 1 - 1$). As shown in Figure 2, the input data were set as $(x_1, x_2, \ldots, x_n)^T$, and the output result was y[23, 24]. The parameters of the BPNN model are as follows: the maximum number of trainings is 1,000, the training error target is 0.001, the learning rate is 0.01, the activation function of the hidden layer is tansig: $f(x) = (2/(1 + e^{-2x})) - 1$, the activation function of the output layer is purelin: $f(x) = x$, and the training function is trainlm. The parameter settings for the DE algorithm are as follows: the population size is $N_p = 10$, the number of iterations is $g = 100$, the results of several trial calculations determine the mutation factor is $F_0 = 0.9$, and the crossover factor $\text{CR}_{\min} = 0.1$ and $\text{CR}_{\max} = 0.7$. The experiment was performed on a PC with an Intel i7 2.4 GHZ CPU and an 8 GB RAM; the algorithm was written in the MATLAB R2018a environment.

4.4. Interpretation of Results

4.4.1. Model Prediction Results. In this paper, the traffic flow data observed at an observation point in the road network on May 4 were analyzed. The prediction results are shown in Figure 6. It can be seen from Figure 6 that the DE-BPNN model yields excellent prediction results. The predicted values at each time point are basically consistent with the actual values, so the traffic flow trend throughout the day is excellently predicted, and the trend of the change in the traffic flow is accurately identified.

4.4.2. Comparative Analysis of Models. Theoretically, the adaptive DE-BPNN model proposed in this paper offers higher convergence rates and smaller prediction errors compared with the use of the BPNN alone. In the early stage, the DE algorithm can effectively avoid the local extremum problem and offer fast convergence and optimization. Furthermore, this algorithm can find the optimal initial parameters of the BPNN during training and continuously optimize the BPNN, and the parameter values are more accurate than the initial values randomly generated by the neural network, thereby enhancing the prediction accuracy. To compare the predictive performance of the DE-BPNN model, three classical prediction models were selected for comparison, including the ARIMA-based, WNN-based, and BPNN-based models. The results of the short-term traffic flow predictions performed with the ARIMA, BPNN, and WNN models are given below. The data in Figures 7–9 include the actual values, the predicted values, and the prediction errors, including the emergence of morning and evening peaks.

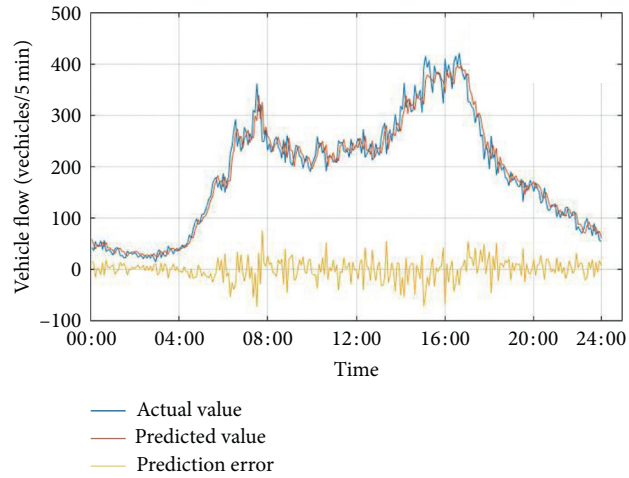


Figure 6: Traffic flow prediction results of the DE-BPNN model.

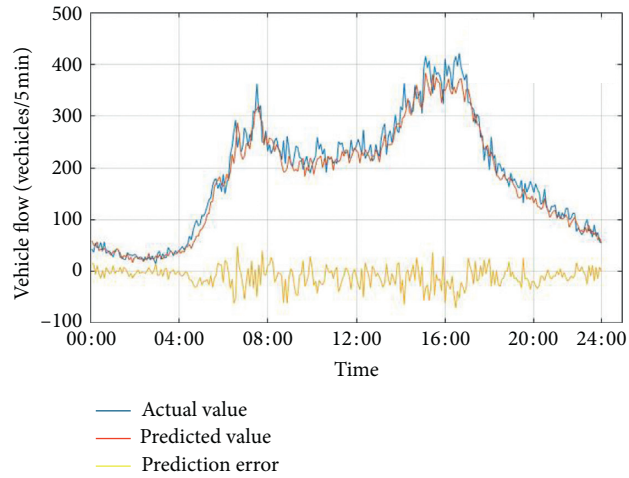


Figure 7: Traffic flow prediction results with the ARIMA model.

For a more intuitive comparison of the model prediction results, the four performance evaluation indicators are used to evaluate the four prediction models.

The performances of the ARIMA, WNN, BPNN, and DE-BPNN models were compared. We used the same data set in all cases. Table 1 shows the prediction result of the 5 min freeway traffic flow verification data set. It should be noted that we only used the traffic flow data as the input for prediction without considering engineering factors related to the traffic flow, such as weather conditions, accidents, and other traffic flow parameters (density and speed). As shown in Table 2, the MAE values of DE-BPNN model decreases 49.04%, 8.16%, and 7.36% as compared with those of the ARIMA, WNN, and BPNN models, respectively; the MSE values of DE-BPNN model decreases 44.97%, 5.88%, and 6.66% as compared with those of the ARIMA, WNN, and BPNN models, respectively; the MAPE values of the DE-BPNN decreases 33.43%, 19.68%, and 18.55% as compared with those of the ARIMA, WNN, and BPNN models, respectively; and the MSPE values decreases 23.58%, 29.85%,

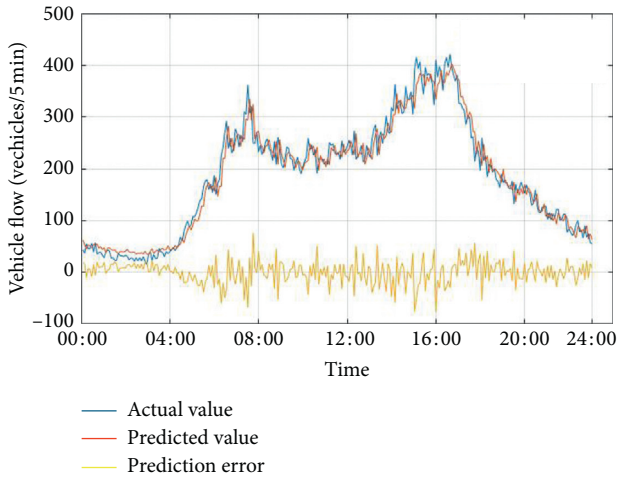


Figure 8: Traffic flow prediction results with the WNN model.

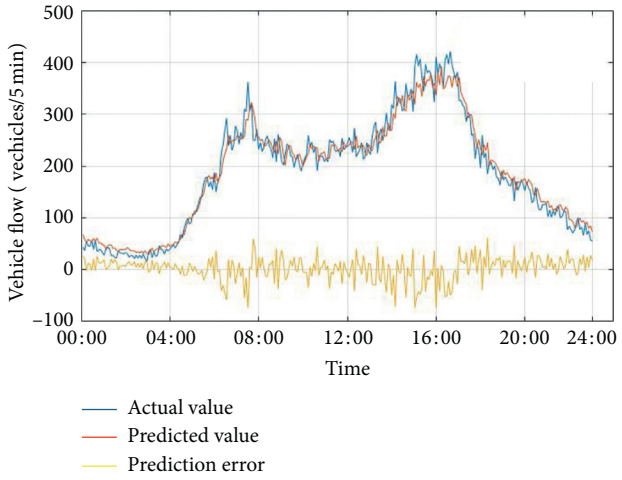


Figure 9: Traffic flow prediction results with the BPNN model.

Table 1: Prediction performances of the ARIMA, WNN, BPNN, and DE-BPNN models.

Model	MAE	MSE	MAPE	MSPE
ARIMA	31.7049	2.3084	0.1735	0.0123
WNN	17.5927	1.3496	0.1438	0.0134
BPNN	17.4396	1.3609	0.1418	0.0130
DE-BPNN	16.1563	1.2702	0.1155	0.0094

Table 2: Percentage improvement in the prediction results with the DE-BPNN model compared with the ARIMA, WNN, and BPNN models.

Model	MAE (%)	MSE (%)	MAPE (%)	MSPE (%)
ARIMA	49.04	44.97	33.43	23.58
WNN	8.16	5.88	19.68	29.85
BPNN	7.36	6.66	18.55	27.69

and 27.69% as compared with those of the ARIMA, WNN, and BPNN models, respectively.

5. Conclusions

Short-term traffic flow prediction is of great social and economic significance for reducing traffic safety hazards, reducing traffic accidents, providing safe and efficient experience for highway travellers, and improving highway traffic and transportation efficiency. By integrating intelligent optimization algorithm theory with machine learning methods, this paper proposes a DE-BPNN model for short-term traffic flow prediction. Restricted by a variety of internal and external factors, the actual traffic system exhibits strong nonlinearity and uncertainty. The BPNN is suitable for any nonlinear fitting; furthermore, to avoid being trapped in local extrema during the conventional BPNN training process, a difference-based heuristic random search DE algorithm is used for global preoptimization, and then, the weights and biases obtained from the DE algorithm are used to train the BPNN, thus improving the prediction accuracy. The results show that the DE-BPNN prediction model effectively improves the prediction accuracy of short-term traffic flow. Compared with the ARIMA, WNN, and BPNN models, the DE-BPNN model leads to lower values for the MAE, MSE, MAPE, and MSPE error evaluation indicators. The research scope can be expanded, and more complex road network data can be used for experiments in the future. Traffic flows may be affected by weather, traffic accidents, and other factors. How to use such auxiliary information to improve the prediction accuracy will also be a focus in future studies.

Data Availability

The data used to support the findings of this study are available at http://pems.dot.ca.gov.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61573076), Science Technology Research Program of Chongqing Municipal Education Commission (Grant no. KJZD-K201800701), Program of Chongqing Innovation and Entrepreneurship for Returned Overseas Scholars of P.R. China (Grant no. cx2018110), Chongqing Natural Science Foundation (Grant no. cstc2020jcyj-msxmX0797), the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant no. KJQN202000717), and the Innovation Foundation of Chongqing Postgraduate Education (Grant no. CYS20282).

References

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[2] D. Huang, Z. Deng, S. Wan, B. Mi, and Y. Liu, "Identification and prediction of urban traffic congestion via cyber-physical link optimization," *IEEE Access*, vol. 6, pp. 63268–63278, 2018.

[3] Y. Lv, Y. Duan, W. Kang, and Z. Li, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[4] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[5] P. Duan, G. Mao, W. Yue et al., "A unified STARIMA based model for short-term traffic flow prediction," in *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1652–1657, IEEE, Maui, Hawaii, USA, 2018.

[6] S. V. Kumar, "Traffic flow prediction using kalman filtering technique," *Procedia Engineering*, vol. 187, pp. 582–587, 2017.

[7] J. Zhang, S. Zhao, Y. Wang, and X. Zhu, "Improved social emotion optimization algorithm for short-term traffic flow forecasting based on back-propagation neural network," *Journal of Shanghai Jiaotong University (Science)*, vol. 24, no. 2, pp. 209–219, 2019.

[8] M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 871–882, 2013.

[9] M. Duo, Y. Qi, G. Lina et al., "A short-term traffic flow prediction model based on EMD and GPSO-SVM," in *Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 2554–2558, IEEE, Chongqing, China, 2017.

[10] G. Dai, C. Ma, and X. Xu, "Short-term traffic flow prediction method for urban road sections based on space-time analysis and GRU," *IEEE Access*, vol. 7, pp. 143025–143035, 2019.

[11] X. Chen, X. Cai, J. Liang, and Q. Liu, "Ensemble learning multiple LSSVR with improved harmony search algorithm for short-term traffic flow forecasting," *IEEE Access*, vol. 6, pp. 9347–9357, 2018.

[12] L. Zhao, D. Wei, Y. Dong-Mei et al., "Short-term traffic flow forecasting based on combination of k-nearest neighbor and support vector regression," *Journal of Highway & Transportation Research & Development*, vol. 12, no. 1, pp. 89–96, 2018.

[13] Y. Ma, Z. Zhang, and A. Ihler, "Multi-lane short-term traffic forecasting with convolutional LSTM network," *IEEE Access*, vol. 8, pp. 34629–34643, 2020.

[14] Y. Gu, W. Lu, X. Xu et al., "An improved bayesian combination model for short-term traffic prediction with deep learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1332–1342, 2019.

[15] W. Cheng and P. Feng, "Network traffic prediction algorithm research based on PSO-BP neural network," in *Proceedings of the 2015 International Conference on Intelligent Systems Research and Mechatronics Engineering*, Atlantis Press, Zhengzhou, China, 2015.

[16] S. A. Zargari, S. Z. Siabil, A. H. Alavi et al., "A computational intelligence-based approach for short-term traffic flow prediction," *Expert Systems*, vol. 29, no. 2, pp. 124–142, 2012.

[17] K. Kumar, M. Parida, and V. K. Katiyar, "Short term traffic flow prediction for a non urban highway using artificial neural network," *Procedia—Social and Behavioral Sciences*, vol. 104, no. 2, pp. 755–764, 2013.

[18] J. Liu and J. Lampinen, "A fuzzy adaptive differential evolution algorithm," *Soft Computing*, vol. 9, no. 6, pp. 448–462, 2005.

[19] J. Ye, J. Zhao, K. Ye et al., "How to build a graph-based deep learning architecture in traffic domain: a survey," 2020, https://arxiv.org/abs/2005.11691.

[20] Caltrans, "Performance measurement system (PeMS)," 2020, http://pems.dot.ca.gov.

[21] D. Huang, Z. Deng, L. Zhao et al., "A short-term traffic flow forecasting method based on markov chain and grey verhulst model," in *Proceedings of the 2017 6th Data Driven Control and Learning Systems (DDCLS)*, pp. 606–610, IEEE, Chongqing, China, 2017.

[22] D. Huang, Z. Deng, and B. Mi, "A new synergistic forecasting method for short-term traffic flow with event-triggered strong fluctuation," *Journal of Control Science and Engineering*, vol. 2018, Article ID 4570493, 8 pages, 2018.

[23] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: comparison of modeling approaches," *Journal of Transportation Engineering*, vol. 123, no. 4, pp. 261–266, 1997.

[24] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: overview of objectives and methods," *Transport Reviews*, vol. 24, no. 5, pp. 533–557, 2004.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6662662, 11 pages
https://doi.org/10.1155/2020/6662662

Research Article

Balancing Access Control and Privacy for Data Deduplication via Functional Encryption

Bo Mi,[1] Ping Long ,[1] Yang Liu,[1] and Fengtian Kuang[2]

[1]*College of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[2]*College of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400074, China*

Correspondence should be addressed to Ping Long; longpingcq@163.com

Received 2 November 2020; Revised 19 November 2020; Accepted 24 November 2020; Published 10 December 2020

Academic Editor: Yong Chen

Data deduplication serves as an effective way to optimize the storage occupation and the bandwidth consumption over clouds. As for the security of deduplication mechanism, users' privacy and accessibility are of utmost concern since data are outsourced. However, the functionality of redundancy removal and the indistinguishability of deduplication labels are naturally incompatible, which bring about a lot of threats on data security. Besides, the access control of sharing copies may lead to infringement on users' attributes and cumbersome query overheads. To balance the usability with the confidentiality of deduplication labels and securely realize an elaborate access structure, a novel data deduplication scheme is proposed in this paper. Briefly speaking, we drew support from learning with errors (LWE) to make sure that the deduplication labels are only differentiable during the duplication check process. Instead of authority matching, the proof of ownership (PoW) is then implemented under the paradigm of inner production. Since the deduplication label is light-weighted and the inner production is easy to carry out, our scheme is more efficient in terms of computation and storage. Security analysis also indicated that the deduplication labels are distinguishable only for duplication check, and the probability of falsifying a valid ownership is negligible.

1. Introduction

As a flourishing service mode, cloud computing adopts load balancing, distributed computing, and other technologies to conveniently provide computation and storage functions for remote follow-up users, thus saving local resources and promoting work efficiency. However, if the users immoderately outsource their data to the cloud, a serious problem may occur due to massive duplicated data. As reported in [1], almost half of the cloud storage is wasted because of data redundancy. Consequently, the budget for managing duplicate data raises up to eight times than that of source data maintenance [2, 3]. With the explosive growth of data nowadays, the tremendous storage requirements or the exorbitant administrative expenses have put enormous pressure on cloud service providers. Therefore, how to store and manage data economically and efficiently has become a serious challenge for these enterprises.

To cut down the costs caused by redundant data, deduplication technology has been widely used by cloud service providers [4]. In such a technology, duplication check and proof of ownership are two key problems. Till now, the problem of how to balance the conflict between comparability and confidentiality for secure duplication check remains unsolved [5]. Meanwhile, the problems of how to efficiently validate the access authority and how to achieve complex access structures are also urgent to address, considering that the mechanism of query matching is cumbersome and the downloading certificates may be abused to launch various attacks.

As a research hotspot, lots of attentions are put on the efficiency and security of data deduplication. In the published literature, Li et al. [6] suggested carrying out deduplication by comparing the fingerprint of the outsourced file with the uploaded ones in a direct way. However, this method is deficient since the communication and comparison of those fingerprints are inefficient and the contents

of data are exposed. To reduce the traffic of deduplication labels and conceal the data, Puzio et al. [7] used the hash function to code the same plaintexts into identical values, which serve as the labels for duplication check. Although this method achieved the goals of transmission efficiency and storage saving, it is vulnerable to dictionary attacks since the hash values are overt.

In order to ensure the confidentiality of deduplication labels, Chen et al. [8] utilized the message lock encryption (MLE) to encrypt those hash values of data. However, the traditional MLE scheme is not semantic secure and vulnerable against quantum attacks [9].

Fortunately, cryptographers have been devoted to design secure, efficient, and effective crypto systems to resist quantum attacks in recent years. In 2005, Regev et al. [10] proposed a novel paradigm as an underpinning of cryptography, namely, learning with errors. They proved that the difficulty of solving it is equivalent to the hardness of shortest vector problem (SVP) over lattice, and thus, it can resist the attacks based on quantum computing. Besides, it is provided with the capacity of homomorphic and linear computation. Therefore, we consider exploiting it in our scheme to ensure the functionality, efficiency, and security of deduplication labels.

As for the proof of ownership, the best solutions till now are all based on Merkle hash tree (MHT) [11, 12]. In detail, the cloud and the user independently hold an MHT computed from the outsourced data. Thus, the user can upload the same MHT to the cloud for comparison. The disadvantages of such scheme are not only high storage and communication overheads but also low computation efficiency. Therefore, Chen et al. [13] improved it by randomly asking the cloud to select some leaf nodes of the MHT to challenge the user. The user must trace the path from the root to these leaves as a reply to prove that he possesses the same tree. Although this method does not require the transmission of the whole MHT for comparison, it demanded that the user and the cloud should construct and store a complete MHT for each file. Moreover, the challenge-response mode implies a long delay.

In order to promote the performance of PoW, the advantages of inner product predicate gradually entered the researchers' sight [14–16]. Roughly speaking, only if the inner product results 0, the user can be granted a permission to access the corresponding file. The most significant merit of this method is using computation instead of comparison to efficiently perform ownership proof. Therefore, we adopted it in our scheme to balance the conflict between the variety of access structures and the security of users' privacy.

Aiming at checking replication over semantic secure deduplication labels and achieving fine-grained access control, this paper proposed a novel cloud data deduplication scheme by exploiting LWE (learning with errors) together with inner product predicate. Our contributions are abbreviated as follows:

(i) Though designed for the purpose of deduplication, the deduplication labels are indistinguishable to any process except for duplication check. This property is achieved in virtue of semantic secure and homomorphic LWE, which is also resistant to quantum attack.

(ii) The proof of ownership is carried out by inner product, which is computationally efficient. Besides, we impose the accessibility of users on their attributes, implying the functionality of the elaborate access structure and ownership transfer.

(iii) For each file, only one light-weighted downloading certificate should be stored by the cloud, while the clients should only carry out and upload its corresponding proof on demand. That is to say that both the storage and bandwidth are economic for cross-user access.

The rest of this paper is organized as follows. In Section 2, some formal definitions related to LWE and inner product predicate are given. Section 3 depicts our deduplication scheme, including the detailed way for duplication check and ownership proof. The correctness of our scheme is formally validated in Section 4, followed by security and performance analysis in Sections 5–7 that concludes the paper.

2. Preliminaries

For better understanding of our scheme, the concepts related to learning with errors and inner product predicate [2, 17] will be introduced in advance.

Definition 1 (Integer lattice). An integer lattice Λ is the integer linear combination of vectors $a_1, a_2, \ldots, a_k a_1, a_2, \ldots, a_k$ over $\mathbb{Z}^m$, expressed as

$$\Lambda(a_1, a_2, \ldots, a_k) = \left\{ \sum_{i=1}^{k} a_i z_i : z_i \in \mathbb{Z} \right\}. \tag{1}$$

Definition 2 (LWE hardness assumption). On parameters n, m, q, α and a discrete Gaussian distribution χ, where

$$\Pr[x \longleftarrow \chi : |x| > \alpha q] < \mathrm{negl}(n), \tag{2}$$

for $x \in \mathbb{Z}_q$, we select a noise e from χ^m and uniformly sample a vector $s \in \mathbb{Z}_q^n$ together with a matrix $P \in \mathbb{Z}_q^{n \times m}$. Based on the value of

$$b = [sP + e]_q, \tag{3}$$

two versions of LWE hardness can be defined as follows:

(a) LWE-Search hardness: Given multiple pairs of (P, b) on constant P and s, searching for the value of s is difficult.

(b) LWE-Determination hardness: For uniformly sampled $b' \in \mathbb{Z}_q^n$, the tuples of (P, b) and (P, b') are statistically indistinguishable. It means that it is difficult to tell if the second term of those tuples are randomly chosen or computed from formula (3).

In fact, the LWE-search hardness is equivalent to the problem of finding a short enough vector in lattice (GapSVP), and the LWE-determination hardness can be reduced to the problem of solving linearly independent shortest vectors (SIVP) of a lattice in the worst case. Therefore, the LWE assumption can be used to guarantee the one-way property for encryption with semantic security.

Definition 3 (Inner product predicate). The inner product predicate $P_{n,q}$ is defined on the Cartesian product $K \times I$ that

$$P_{n,q}(\overrightarrow{v}, t\overrightarrow{w}) = \begin{cases} 1, & \sum_{i=1,\ldots,n} v_i w_i = 0, \\ 0, & \text{other.} \end{cases} \tag{4}$$

From the perspective of functional encryption (FE), I can be deemed as the space of ciphertexts and K is composed of secret keys. Once a correct key $\overrightarrow{v}$ is known, we are able to learn the output of function $P_{n,q}(\overrightarrow{v}, t\overrightarrow{w})$.

To construct an attribute-based access control policy, the access structure is coded as a vector $\overrightarrow{w}$, thus the access authority can be verified with respect to the consistency of authorization certificate $\overrightarrow{v}$.To avoid obfuscation, the symbols used in this paper is listed in advance, as in Table 1.

3. Duplication Check Based on LWE

To prevent dictionary attacks caused by the exposure of deduplication labels, we intended to make them indistinguishable except for the process of duplication check. Therefore, LWE is adopted to randomize the hash value of file to ensure the indistinguishability of deduplication labels and resist the attacks of quantum computation. In addition, we exploit inner product predicate to control the accessibility of clients, which is flexible for functions such as cross-user sharing and ownership transfer. The logical idea of our scheme is illustrated below, which is shown in Figure 1.

4. File Upload

A user denoted as A, who possesses a file M_A and expects to upload it, is not aware of its existence over cloud at the very beginning. To avoid unnecessary storage and bandwidth, he is supposed to check if there is a copy already held by the server.

Drawing support from any strong-collision resistant hash function

$$H\colon \{0,1\}^* \longrightarrow \{0,1\}^{\ell}, \tag{5}$$

the user figures out the hash value of file M_A as

$$h_A = H(M_A), \tag{6}$$

and codes it as a vector of ℓ elements. On fixed public matrix $P \in \mathbb{Z}_q^{n\times m}$ and a pseudorandom sequence generator (PSRG), he produces a vector

$$\overrightarrow{s}_A = (\mathrm{PSRG}(h_A), -1) \in \mathbb{Z}_q^{n+1}, \tag{7}$$

and exploits LWE to obtain

$$\overrightarrow{v_A} = (P, b_A) \cdot r_A \in \mathbb{Z}_q^{n+1}. \tag{8}$$

Herein, $b_A = [\mathrm{PSRG}(h_A)P + e_A]_q$ stands for the last row of (P, b_A), where $\mathrm{PSRG}(h_A)$ is considered as a n dimensional vector and r_A is randomly chosen from $\{-1, 0, 1\}^m$.To this point, the user is ready to take the $n+1$ dimensional vector $\overrightarrow{v_A}$ as a deduplication label and upload it to the cloud. Since the subsequent actions he should take depend directly on the result of duplication check, we will discuss the situations for original uploader and repeated uploader, respectively, who are denoted as A and B for clarity.

4.1. The Process of Original Uploading. We defer the description of duplication check to the circumstance of repeated upload, if suppose that user A is informed with the inexistence of file M_A. For further deduplication, he should secretly upload the deduplication certificate $\overrightarrow{s_A}$ to the cloud. To ensure the confidentiality of his file, its hash value can be taken as a symmetric key $\mathrm{sk}_A = h_A$ to hide the plaintext as

$$\mathrm{Enc}_{\mathrm{sk}_A}(M_A) = C_A. \tag{9}$$

Then, the cloud preserves the uploaded ciphertext C_A for storage and the deduplication certificate $\overrightarrow{s_A}$ for duplication check. To further retrieve the file, user A ought to upload a downloading certificate as well, like the following.

Assuming that the attributes of user A correspond to a secret vector $\overrightarrow{\mu_A} = (\mu_{A,0}, \mu_{A,1}, \ldots, \mu_{A,n-1}) \in Z_q^n$, which can also be regarded as a polynomial

$$f(\overrightarrow{\mu_A}) = \mu_{A,0} + \mu_{A,1}x + \cdots + \mu_{A,n-1}x^{n-1} \bmod q. \tag{10}$$

It is worth mentioning that the user is aware of the elements of $\overrightarrow{\mu_A}$ only if he corresponds to those attributes. To actualize a functional encryption which reflects the access structure in covert manner, he uniformly samples two vectors $w_i\,(i = 0, 1, \ldots, n-2)$ and $\overrightarrow{u_A} = (u_{A,0}, u_{A,n-1}, u_{A,n-2}, \ldots, u_{A,1})$. Similarly, the vector $\overrightarrow{u_A} = (u_{A,0}, u_{A,n-1}, \ldots, u_{A,1}) \in Z_q^n$ can also be expressed as a polynomial $g(\overrightarrow{u_A}) = u_{A,0} + u_{A,n-1}x + \cdots + u_{A,1}x^{n-1} \bmod q$, which is equivalent to a cyclic matrix

$$U_A = \begin{bmatrix} u_{A,0} & u_{A,n-1} & \cdots & u_{A,1} \\ u_{A,1} & u_{A,0} & \cdots & u_{A,2} \\ \vdots & \vdots & \cdots & \vdots \\ u_{A,n-1} & u_{A,n-2} & \cdots & u_{A,0} \end{bmatrix} \in Z_q^{n\times n}, \tag{11}$$

with respect to the homorganic between polynomials and cyclic matrices.

In order to construct the correct downloading certificate, he computes $U_A \cdot \overrightarrow{\mu_A} = \overrightarrow{X_A} = (x_{A,0}, x_{A,1}, \ldots x_{A,n-1})^T$ and figures out w_{n-1} for $<\overrightarrow{X_A}, \overrightarrow{w}> = 0 \bmod q$ by

$$w_{n-1} = -\frac{\sum_{i=0}^{n-2} x_{A,i} w_i}{x_{n-1}} \bmod q. \tag{12}$$

Table 1: Symbols and notations.

Symbol	Notation
f	Length of a file
b	Length of each file block
g	Length of the hash value
N	Number of users participating in key aggregation
K	Total number of bloom filters
L	Length of bloom filter array
$Q(n)$	Number of common attributes in a user group
n	Number of attributes for an individual user
$Hash$	Computational cost of performing a hash function
CE_K	Computational cost of performing a key aggregation
Enc	Computational overhead of performing a symmetric encryption
PoW	Computational cost of performing a proof of ownership
Add	Computational cost of performing an addition

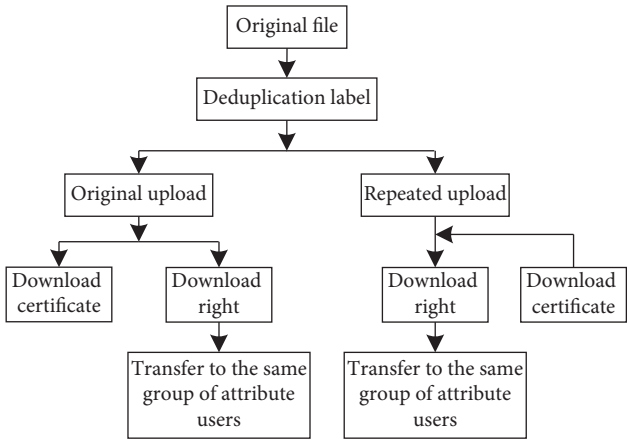


Figure 1: The overall framework of the deduplication program.

After that, the user uploads $\overrightarrow{w} = (w_0, w_1, \ldots, w_{n-1})$ as the downloading certificate and submits

$$y = w_{n-2}x_{A,n-2} + w_{n-1}x_{A,n-1} = -\left(\sum_{i=0}^{n-3} w_i x_{A,i}\right), \tag{13}$$

to the cloud for further expansions on access structure.

At the end, the user preserves the hash value sk_A, the essential elements $\overrightarrow{u_A} = (u_{A,n-1}, \ldots, u_{A,1}, u_{A,0})$ of matrix U_A, and the replied link of outsourced file. While the ciphertext C_A can be held by the cloud server, attached with $\overrightarrow{s_A}$, y, and $\overrightarrow{w}$ for duplication check, access expansion, and ownership proof.

4.2. The Process of Repeated Uploading. As mentioned before, once a deduplication label is figured out, any user should firstly hand it over to the cloud for duplication check. Assume that the deduplication certificate of an existing file M_A is $\overrightarrow{s_A}$, the cloud can inspect its consistency with another deduplication label $\overrightarrow{v_B}$ as the following:

When user B expects to upload his file M_B, he submits its deduplication label $\overrightarrow{v_B} = (P, b_B) \cdot r_B$ to the cloud and keeps the hash value h_B private.

Based on an outsourced deduplication certificate $\overrightarrow{s_A}$, the cloud computes within a lifted interval $[-(q-1)/2, (q-1)/2]$ which is as follows:

$$< \overrightarrow{s_A}, \overrightarrow{v_B} > \; = \; < \mathrm{PSRG}(h_A)P - \mathrm{PSRG}(h_B)P - e_B, r_B > . \tag{14}$$

It can be seen that, if the two files are identical, only $\langle -e_B, r_B \rangle$ will remain in formula (14). Therefore, when the result satisfies

$$\left\| \langle \overrightarrow{s_A}, t\overrightarrow{v_B} \rangle \right\|_\infty \leq \alpha q, \tag{15}$$

the cloud can ensure the duplication of file M_B with negligible false positive.

To validate his accessibility, user B should also figure out the downloading right of the corresponding file. However, it is more reasonable to use existing download rights $\overrightarrow{w}$ held by the cloud server for the purpose of storage saving. Based on this, user B can use the following subprotocol to obtain the download right of the duplicate file, and the cloud will simply send the link back to him for further retrieval.

4.3. The Subprotocol for Access Expansion. Denoting the secret corresponding to the attributes of repeated uploader B as $\overrightarrow{\mu_B} = (\mu_{B,0}, \mu_{B,1}, \ldots, \mu_{B,n-1})^T$. To bind the access structure with his own attributes, he should also figure out a cyclic matrix U_B which can be used to compute his proof of ownership which is as follows:

$$< U_B \cdot \overrightarrow{\mu_B}, \overrightarrow{w} > \; = 0. \tag{16}$$

Though the downloading certificate $\overrightarrow{w}$ cannot be exposed to prevent unauthorized access, the cloud can provide user B with the values of $(w_{n-3}, w_{n-2}, w_{n-1})$ and y to help him calculate the correct cyclic matrix U_B. Thus, the downloading right can be carried out by user B in Algorithm 1.

5. Proof of Ownership

Once any legal user obtained his downloading right, he should be authorized to retrieve the corresponding file from the cloud. To improve the efficiency of ownership proof, access authorization is executed in a computational way.

After uploading, the legal user A will be provided with the last row $\overrightarrow{u_A} = (u_{A,n-1}, \ldots, u_{A,1}, u_{A,0})$ of the cyclic matrix. Therefore, he only needs to form the cyclic matrix U_A and combines it with his attribute vector $\overrightarrow{u_A}$ to figure out the downloading right. Based on the resulted vector, the cloud can easily verify his accessibility by functional encryption. The process of PoW is completely given in Algorithm 2.

After obtaining the ciphertext C_A, user A can decrypt the file by computing $\mathrm{Dec}_{\mathrm{sk}_A}(C_A) = M_A$ because he is aware of the secret key $\mathrm{sk}_A = H(M_A)$.

In fact, the ownership proof process for user B is similar to that of user A. The reason why user B can also decrypt the file C_A is due to the equality of plaintexts M_A and M_B. Since

$sk_B = H(M_A)$ he is able to obtain the corresponding file via $\mathrm{Dec}_{sk_B}(C_A) = M_A$.

6. Downloading Right Transfer

On noting that, without the secret vectors corresponding to the attributes of legal users, other users are incapable of computing the downloading right even if the last row of cyclic matrix is known. Since the access controls subprotocol, any legal user can directly transfer the resultant downloading right to other users to avoid redundant operations such as peer to peer transmission. However, it may lead to the abuse of downloading right and violate the confidentiality of user's attributes. Practically, legal users are prone to transfer the downloading right of their file to others who share party of common attributes with him. Therefore, we designed a protocol that any legal user can update the downloading right and transfer it to a group of users with the same set of attributes. In this way, the owner does not have to download the file from the cloud and only needs to transfer the downloading right to other users to complete file sharing, which effectively reduces the consumption of communication bandwidth.

Definition 4 (Common attributes vector). Suppose that the file owner A can be identified by attributes vector $\overrightarrow{\mu_A} = (\mu_{A,0}, \mu_{A,1}, \ldots, \mu_{A,n-1})$, and all users in the same group have $Q(n)$ common attributes denoted by $\overrightarrow{\mu_{\mathrm{all}}} = (\mu_{\mathrm{all},0}, \ldots, \mu_{\mathrm{all},n-1}, \ldots, \mu_{\mathrm{all},Q(n)})$. Then, the common attributes vector $\overrightarrow{\mu_{\mathrm{team}}} = (\mu_{\mathrm{team},0}, \ldots, \mu_{\mathrm{team},n-1})$ can be defined as a partial ordering relation that $\mu_{\mathrm{team},i} = \mu_{A,i}$ if $\mu_{A,i} \in \{\mu_{\mathrm{tall},j} | j = 0, \ldots, Q(n)\}$ and $\mu_{\mathrm{team},i} = 0$, otherwise.

Specifically, the process that the user A constructs the common attribute vector $\overrightarrow{\mu_{\mathrm{team}}} = (\mu_{\mathrm{team},0}, \mu_{\mathrm{team},1}, \ldots, \mu_{\mathrm{team},n-1})$ is detailed in Figures 2 and 3.

As shown in Figures 2 and 3, the user A mainly retains the secret attributes shared by the same group and sets the attributes which are distinct in the user group as 0. Finally, he outputs a common attribute vector $\overrightarrow{\mu_{\mathrm{team}}}$.

6.1. Proof of Ownership. The user A performs the following steps to realize the PoW and retrieves $(w_{n-3}, w_{n-2}, w_{n-1})$ and y.If the downloading right is valid, the inner product will result in 0, meaning that the user A is authorized to retrieve the file. Therefore, the cloud server returns $C_A(w_{n-3}, w_{n-2}, w_{n-1})$ and y back to him. Similarly, the values of $(w_{n-3}, w_{n-2}, w_{n-1})$ and y can be used to update the downloading right for a group of users. Specifically, the process of PoW is shown in Algorithm 2, which is the same for any valid user even if the updated downloading right is used.

6.2. Update the Downloading Right. To share the file to a group, the downloading right update process can be carried out by the user A as the following. In a clear form, the process that the user A calculates the downloading right for a group of users is shown in Algorithm 3.

6.3. Sharing the Downloading Right. After the previous two stages, the user A can share the vector $\overrightarrow{u'_{\mathrm{team}}} = (u'_{\mathrm{team},n-1}, \ldots, u'_{\mathrm{team},1}, u'_{\mathrm{team},0})$ and the secret key sk_A to all users who are within the same attributes set. In these ways, a group of users are provided with the downloading right, which can be valid if the common attributes vector $\overrightarrow{\mu_{\mathrm{team}}}$ is known.

7. Correctness Proof

The previous section is mainly composed of three parts, namely, the file uploading, the proof of ownership, and the downloading right transfer. To verify the correctness of our design, this section intends to prove that file duplication can be effectively eliminated and only authenticated users can access the file.

Firstly, the correctness for the deduplication label is given by Theorem 1.

Theorem 1 (Correctness of deduplication label). *Suppose that the cloud holds a deduplication certificate $\overrightarrow{s_A}$ which is correspondent to file M_A. After the user B uploaded the deduplication label $\overrightarrow{v_B}$ before outsourcing the same file M_A, the cloud can perform deduplication correctly with negligible false positive.*

Proof. Due to the deduplication certificate $\overrightarrow{s_A} = (\mathrm{PSRG}(h_A), -1) \in \mathbb{Z}_q^{n+1}$ stored on the cloud, where $h_A = H(M_A)$. After the user B uploaded the deduplication label $\overrightarrow{v_B} = (P, b_B) \cdot r_B \in \mathbb{Z}_q^{n+1}$ of the same file M_A to the cloud for $b_B = [\mathrm{PSRG}(h_B) + e_B]$, the cloud executes the following calculation on each deduplication certificate. Once $\overrightarrow{s_A}$ is met, the inner product can be carried out as follows:

$$
\begin{aligned}
<\overrightarrow{s_A}, \overrightarrow{v_B}> &= \left(\left(\underbrace{\mathrm{PSRG}(h_A)}_{n\,\mathrm{bits}}, -1 \right) \cdot \left(\begin{bmatrix} \underbrace{P}_{n\times m} \\ \underbrace{b_B}_{1\times m} \end{bmatrix}^{(n+1)\times m} \cdot r_B \right) \right) \\
&= \left(\left(\underbrace{\mathrm{PSRG}(h_A)}_{1\times n}, -1 \right) \cdot \begin{bmatrix} \underbrace{P}_{n\times m} \\ \underbrace{b_B}_{1\times m} \end{bmatrix}^{(n+1)\times m} \right) \cdot r_B \\
&= (\mathrm{PSRG}(h_A) \cdot P - b_B) \cdot r_B \\
&= <\mathrm{PSRG}(h_A) \cdot P - (\mathrm{PSRG}(h_B) \cdot P + e_B), r_B> \\
&= <\mathrm{PSRG}(h_A) \cdot P - \mathrm{PSRG}(h_B) \cdot P - e_B, r_B>.
\end{aligned}
\tag{17}
$$

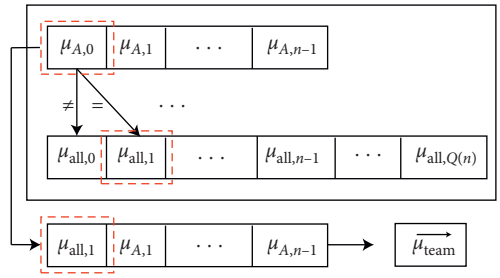


Figure 2: Common attributes.

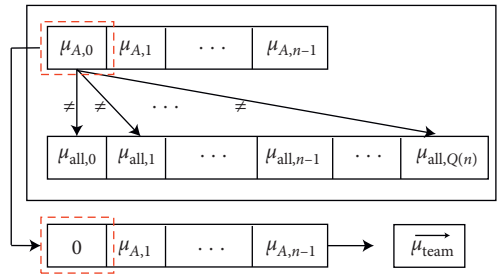


Figure 3: Noncommon attributes.

Since $\text{PSRG}(\cdot)$ is a deterministic algorithm, when $h_A = h_B$,$\text{PSRG}(h_A) = \text{PSRG}(h_B)$. Meanwhile, according to the common matrix P, it is obvious that $\text{PSRG}(h_A) \cdot P = \text{PSRG}(h_B) \cdot P$. Thus, we can easily see that $\langle \overrightarrow{s_A}, \overrightarrow{v_B} \rangle = \langle -e_B, r_B \rangle$ from equation (17). Because the inner product of $< -e_B, r_B > \leq (1/Q(m))$ is definite, the inner product of $< \overrightarrow{s_A}, \overrightarrow{v_B} > \leq (1/Q(m))$ can also be guaranteed, meaning that duplication can be detected with 100% probability.

Theorem 2 (Correctness of download right). *Suppose that the cloud possesses a downloading certificate $\overrightarrow{w}$ corresponding to file M_A, then any legal user can correctly pass the procedure of PoW in terms of his downloading right.*

Proof. For user A, who uploads the original file M_A to the cloud, he rotates $\overrightarrow{u_A} = (u_{A,n-1}, \ldots, u_{A,1}, u_{A,0})$ to right by one bit to get $\overrightarrow{u_A'} = (u_{A,0}, u_{A,n-1}, \ldots, u_{A,1})$ and uses it to reconstruct the cyclic matrix U_A. Then, user A calculates download right

$$\overrightarrow{X_A} = U_A \cdot \overrightarrow{\mu_A} = \left(x_{A,0}, x_{A,1}, \ldots x_{A,n-1}\right)^T, \tag{18}$$

where $\overrightarrow{\mu_A} = (\mu_{A,0}, \mu_{A,1}, \ldots, \mu_{A,n-1})$ are the attributes of the user A. After which the user A sends the download right $\overrightarrow{X_A}$ to the cloud. Finally, the cloud calculates the inner product of

$$< \overrightarrow{w}, \overrightarrow{X_A} > = \sum_{i=0}^{n-1} w_i \cdot x_{A,i} = \sum_{i=0}^{n-2} w_i \cdot x_{A,i} + w_{n-1} \cdot x_{A,n-1} = \sum_{i=0}^{n-2} w_i \cdot x_{A,i} + \left(-\sum_{i=0}^{n-2} w_i \cdot x_{A,i} \right) = 0. \tag{19}$$

Based on the last element of download certificate $\overrightarrow{w}$ is $w_{n-1} = -((\sum_{i=0}^{n-2} x_{A,i} w_i)/(x_{n-1})) \bmod q$, so that the result of $w_{n-1} x_{A,n-1}$ can transfer as $(-\sum_{i=0}^{n-2} w_i x_{A,i})$. Therefore, the inner product of $\langle \overrightarrow{X_A}, t\overrightarrow{w} \rangle$ is zero. For the repeated file user B, the first two steps are the same for the user B.

Then, he also gets the result of download right $\overrightarrow{X_B}$ and sends it to the cloud. Moreover, the inner product of $\langle \vec{w}, t\overrightarrow{X_B'} \rangle$ calculates the process as follows:

$$\begin{aligned} < \vec{w}, \overrightarrow{X_B'} > &= \vec{w} \cdot \left(x_{B,0}, x_{B,0}, \ldots, x_{B,n-3}', x_{B,n-2}', x_{B,n-1}' \right)^T \\ &= \sum_{i=0}^{n-4} w_i x_{B,i} + w_{n-3} x_{B,n-3}' + \left(w_{n-2} x_{B,n-2}' + w_{n-1} x_{B,n-1}' \right) \\ &= \sum_{i=0}^{n-4} w_i x_{B,i} + w_{n-3} x_{B,n-3}' + y \\ r &= \left[\left(\sum_{i=0}^{n-4} w_i x_{B,i} \right) + \left(w_{n-3} x_{B,n-3} + \left(y' - y \right) \right) \right] + y \\ &= \left[-y' + \left(y' - y \right) \right] + y = (-y) + y = 0. \end{aligned} \tag{20}$$

In a word, all legal users who hold the download right corresponding to file M_A can pass the PoW. □

8. Security Analysis

This part will prove that the deduplication label is indistinguishable except for duplication check process, and the downloading right is resistant to forgery. To begin with, the security about deduplication label is given in Theorem 3.

Theorem 3 (Security of deduplication label). *For legitimate users, whether uploading the same or different files to perform deduplication, the deduplication labels are only distinguishable to the duplication check process.*

Proof. The following analysis will be divided into two cases, with respect to the deduplication labels corresponding to same files and different files.

Case 1. Supposing user A and user B possess the same file. They have the same hash value $h_A = h_B$ of two identical files, and their deduplication labels are

$$\begin{aligned} \overrightarrow{v_A^{(1)}} &= \left(\mathrm{PSRG}(h_A) \cdot P + e_A^{(1)} \right) \cdot r_A^{(1)}, \\ \overrightarrow{v_B^{(1)}} &= \left(\mathrm{PSRG}(h_B) \cdot P + e_B^{(1)} \right) \cdot r_B^{(1)}. \end{aligned} \tag{21}$$

According to the deterministic algorithm $\mathrm{PSRG}(\cdot)$, we can see $\mathrm{PSRG}(h_A) = \mathrm{PSRG}(h_B)$. Moreover, for the common matrix P, it is obvious that $\mathrm{PSRG}(h_A) \cdot P = \mathrm{PSRG}(h_B) \cdot P$. However, e_A, e_B and r_A, r_B are randomly sampled from χ_q^m and $\{-1, 0, 1\}^m$, respectively. The probability that the deduplication labels are identical is $(1/(3q)^m) < (1/Q(m))$, which is negligible. Therefore, we claim that the results $\overrightarrow{v_A^{(1)}} = \overrightarrow{v_B^{(1)}}$ is almost impossible, which means $\overrightarrow{v_A^{(1)}}$ and $\overrightarrow{v_B^{(1)}}$ satisfy semantic security.

Case 2. Supposing user A and user B possess different files. That is to say, they have different file hash values that $h_A \neq h_B$, and the deduplication labels are

$$\begin{aligned} \overrightarrow{v_A^{(2)}} &= \left(\mathrm{PSRG}(h_A) \cdot P + e_A^{(2)} \right) \cdot r_A^{(2)}, \\ \overrightarrow{v_B^{(2)}} &= \left(\mathrm{PSRG}(h_B) \cdot P + e_B^{(2)} \right) \cdot r_B^{(2)}. \end{aligned} \tag{22}$$

Similarly, since $\mathrm{PSRG}(h_A) \neq \mathrm{PSRG}(h_B)$, the probability that deduplication labels are the same is $(1/(n+1)(3q)^m) < (1/Q(m))$, which is indistinguishable from the distribution of Case 1.

Therefore, we can conclude that, since the deduplication labels of the same file are different, Case1 is of the same distribution indistinguishable from Case2, and the deduplication labels are semantic secure. In summary, the deduplication tags corresponding to the same file and different files are indistinguishable. □

Theorem 4 (Security proof of downloading right). *None of the users can forge a valid downloading right $\overrightarrow{X_A}$ which can deceive access control.*

Specifically, the security analysis of the downloading right can be guided by Lemmas 1 and 2.

Lemma 1 *After the original uploader A outsourced the file M_A to the cloud, the entire download certificate $\overrightarrow{w}$ is known only by the cloud.*

Proof. According to inner product predicate, the user A's downloading right $\overrightarrow{X_A}$ can make the inner products $\langle \overrightarrow{X_A}, t\overrightarrow{w} \rangle$ output 0.

However, the download certificate $\overrightarrow{w}$ is calculated by the user A who samples $w_i (i = 0, \ldots, n-2)$ and sets the last element w_{n-1} of the download certificate $\overrightarrow{w}$ to be $w_{n-1} = -((\sum_{i=0}^{n-2} x_{A,i} w_i)/(x_{n-1})) \bmod q$.

Then, when the user A uploads for the first time, the cloud obtains the completed download certificate $\overrightarrow{w}$ corresponding to A's secret attributes. For now, if there is an illegal user who tries to falsify the download certificate $\overrightarrow{w} \xleftarrow{\$} \mathbb{Z}_q^n$ to cheat the PoW system, his advantage is

$$\Pr\left[< \overrightarrow{X_A}, \overrightarrow{w'} > = 0 \right] = \frac{1}{q^{(n-1)}} \leq \frac{1}{Q(n)}, \tag{23}$$

which is negligible. □

Lemma 2. *For repeated file uploaders, they do not know the remaining elements of the download certificate $\overrightarrow{w}$ except for $(w_{n-3}, w_{n-2}, w_{n-1})$.*

Proof. Take a repeated file uploader B as an example, he uses $(w_{n-3}, w_{n-2}, w_{n-1})$ to update the last three elements of download right $\overrightarrow{X_B}$ into $(x'_{B,n-3}, x'_{B,n-2}, x'_{B,n-1})$.

In detail,

$$x'_{B,n-3} = x_{B,n-3} - \frac{(y - y')}{w_{n-3}}, \tag{24}$$

$$w_{n-2} x'_{B,n-2} + w_{n-1} x'_{B,n-1} = y. \tag{25}$$

Since the value of w_{n-3} is known, the result of $x'_{B,n-3}$ can be calculated. However, because the rank of formula (25) is equal to 1 and $w_{n-2} x'_{B,n-2} + w_{n-1} x'_{B,n-1} = y$, the formula of (25) contains two unknowns variables. Thus, the results of $x'_{B,n-2}$ and $x'_{B,n-1}$ are infinite. Therefore, when the user B calculates the downloading right, he does not know the remaining elements of the download certificate $\overrightarrow{w}$ except for $(w_{n-3}, w_{n-2}, w_{n-1})$.

Considering that the solutions of formula (25) are infinite, the security of downloading right can be effectively protected, namely, $\overrightarrow{X_B}$ of the user B. Thus, it also guarantees the confidentiality of legal users' attributes. If an illegal user attempts to forge the remaining $n-3$ elements of $\overrightarrow{w}$ to get the new download certificate $\overrightarrow{w''} = (w''_0, \ldots, w''_{n-4}, w_{n-3} \ldots, w_{n-1})$, his advantage is just

$$\Pr\left[< \overrightarrow{X'_B}, \overrightarrow{w''} > = 0 \right] = \frac{1}{q^{(n-3)}} \leq \frac{1}{Q(n)}, \tag{26}$$

which is negligible. Therefore, our scheme will not expose the remaining elements of the download certificate $\overrightarrow{w}$.

In terms of Lemmas 1 and 2, it can be seemed that no user can forge a valid downloading right since the complete download certificate and the attributes vector $\overrightarrow{\mu_A}$ will not be exposed. □

9. Performance Analysis

Then, the performance of our schemes will be analysed comparing with other main technologies. The notation of symbols can be found in Table 1, as for functions, such as the necessity of third-party, deduplication level, participants, and the necessity of key fusion, and the comparison can be found in Table 2.

Compared with the schemes from [2, 3, 9], our scheme does not require any third-party, which effectively avoided extra trusting relationships and can save numerous computation/communication resources. Moreover, our scheme executes file deduplication amongst multiple users, implying that it is more flexible and more adaptive to various cloud environment. From the perspective of key fusion, when compared with the literature from [2, 3, 8, 9], any key fusion process is unnecessary in our scheme, so that it can be applied even if the user resources are limited.

Then, we compare the computation overheads for deduplication taken by the client, third-party, and cloud in the above schemes. The details are given in Table 3.

Compared with the cost on client side in scheme [3], that of our scheme is $O(f)$Hash + $O(1)$PSRG, where a pseudorandom number sequence is generated instead of N convergence keys. In fact, it means that our scheme is more efficient since PSRG can be iterated generated via small numbers, not saying that our scheme if free of any third-party. Moreover, the hash value of file can be secretly used as the encryption key in this paper. Therefore, there is no need for multiple users to reconstruct the convergence key, which further outperformed the scheme of [3] by avoiding the consumption of key distribution and fusion.

Compared with the schemes in [8, 9], our method does not need to construct Bloom filter or attribute binary tree on client side, so the computational cost is slightly advantageous. In addition, since our scheme does not involve any third-party, the computational cost of TTP can be neglected. As for the overhead on cloud side, our scheme does not have to initialize any ownership data structure compared with that of schemes [8, 9]. Therefore, the calculation is deduced to $O(g)$ since it is not related to the file size but only to the length hash value.

Now, we compare the computational overhead for PoW, respectively on client, third-party and cloud side. The results are shown in Table 4.

It can be seen from Table 4 that users have to preserve and search the Bloom filter or attribute binary tree to accomplish PoW in [2, 3, 8, 9]. So, there is an additional cost $O(kL)$ or $O(N \log N)$ on the client side. However, our scheme does not require this process, so the calculation cost is only $O(f)$Hash + $O(n)$Add, where the second term is just n times of add operation. Comparing the cost on cloud side, our scheme dose also outperformed that of [2, 3, 8, 9], which

User B

Input: $((w_{n-3}, w_{n-2}, w_{n-1}), y, \overrightarrow{\mu_B})$

(1) Samples $\overrightarrow{u_B} = (u_{B,0}, u_{B,n-1}, u_{B,n-2}, \ldots, u_{B,1})$, where $u_{B,i}$ is irreversible and cofficients belong to $\mathbb{Z}_q$

(2) Computes $f(\overrightarrow{\mu_B}) \leftarrow \overrightarrow{\mu_B}$, $g(\overrightarrow{u_B}) \leftarrow \overrightarrow{u_B}$, $x_{B,0} \leftarrow g(\overrightarrow{u_B}) \cdot f(\overrightarrow{\mu_B})$
for all $k \in \{1, \ldots, n-1\}$
$g(\overrightarrow{u_{B,k}}) \leftarrow g(\overrightarrow{u_B}) \cdot x \bmod x^n$, $x_{B,k} \leftarrow g(\overrightarrow{u_{B,k}}) \cdot f(\overrightarrow{\mu_B})$, $\overrightarrow{u_B} \leftarrow \overrightarrow{u_{B,k}}$, $k \leftarrow k+1$

(3) Computes $y' \leftarrow w_{n-2} \cdot x_{B,n-2} + w_{n-1} \cdot x_{B,n-1}$, $x_{B,n-3}' \leftarrow x_{B,n-3} - (y - y')/w_{n-3}$

(4) Samples $x_{B,n-2}' \xleftarrow{\$} \mathbb{Z}_q^n$, $x_{B,n-2} \leftarrow x_{B,n-2}'$; and computes $x_{B,n-1} \leftarrow ((y - w_{n-2} \cdot x_{B,n-2}')/(w_{n-1}))$

(5) For all $k \in \{1, \ldots, n-1\}$

(6) $u_{B,i}\mu_{B,0} + u_{B,[i+(n-1)] \bmod n}\mu_{B,1} + \cdots + u_{B,(i+1) \bmod n}\mu_{B,n-1} \leftarrow x_{B,i}$

Output: $\left\{\overrightarrow{u_B} = t(u_{B,n-1}, \ldots, u_{B,1}, u_{B,0})\right\}$

Algorithm 1: Calculation process chart of repeated file.

User A

Input: $(\overrightarrow{u_A}, \overrightarrow{\mu_A}, n, \mathrm{sk}_A)$

(1) Computes $f(\overrightarrow{\mu_A}) \leftarrow \overrightarrow{\mu_A}$, $g(\overrightarrow{u_A'}) \leftarrow g(\overrightarrow{u_A}) \cdot x \bmod x^n$, $\overrightarrow{u_A} \leftarrow u_A'$, $x_{A,0} \leftarrow g(\overrightarrow{u_A}) \cdot f(\overrightarrow{\mu_A})$

(2) Computes $f(\overrightarrow{\mu_B}) \leftarrow \overrightarrow{\mu_B}$, $g(\overrightarrow{u_B}) \leftarrow \overrightarrow{u_B}$;
$x_{B,0} \leftarrow g(\overrightarrow{u_B}) \cdot f(\overrightarrow{\mu_B})$

(3) For all $k \in \{1, \ldots, n-1\}$
$g(\overrightarrow{u_{A,k}}) \leftarrow g(\overrightarrow{u_A}) \cdot x \bmod x^n$, $x_{A,k} \leftarrow g(\overrightarrow{u_{A,k}}) \cdot f(\overrightarrow{\mu_A})$
$\overrightarrow{u_A} \leftarrow \overrightarrow{u_{A,k}}$, $k \leftarrow k+1$

(4) Sends $\left\{\overrightarrow{X_A} = (x_{A,0}, x_{A,1}, \ldots, x_{A,n-1})\right\}$ to Cloud.

Cloud

Input: $(\overrightarrow{w}, \overrightarrow{X_A}, C_A, \alpha, q)$

(5) Computes $<\overrightarrow{w}, \overrightarrow{X_A}> \leftarrow \sum_{i=0}^{n-1} w_i \cdot x_{A,i}$

(6) If $<\overrightarrow{w}, \overrightarrow{X_A}> = 0$
Output: $\{C_A, (w_{n-3}, w_{n-2}, w_{n-1}), y\}$; Otherwise
Output: NULL.

(7) Sends $\{C_A, (w_{n-3}, w_{n-2}, w_{n-1}), y\}$ to User A

Algorithm 2: Process chart of PoW for the original file user.

User A

Input: $(\overrightarrow{\mu_{\text{team}}}, n, (w_{n-3}, w_{n-2}, w_{n-1}), y)$

(1) Samples $\overrightarrow{u_{\text{team}}} = (u_{\text{team},0}, u_{\text{team},1}, \ldots, u_{\text{team},n-1})$, where $u_{\text{team},0}$ is irreversible with cofficients belong to $\mathbb{Z}_q$

(2) Computes $f(\overrightarrow{\mu_{\text{team}}}) \leftarrow \overrightarrow{\mu_{\text{team}}}$
for all $k \in \{1, \ldots, n-1\}$
$g(\overrightarrow{u_{\text{team},k}}) \leftarrow g(\overrightarrow{u_{\text{team}}}) \cdot x \bmod x^{n+1}$; $x_{\text{team},(n-k-1)} \leftarrow g(\overrightarrow{u_{\text{team},k}}) \cdot f(\overrightarrow{\mu_{\text{team}}})$; $\overrightarrow{u_{\text{team}}} \leftarrow \overrightarrow{u_{\text{team},k}}$; $k \leftarrow k+1$

(3) Computes $y' \leftarrow w_{n-2} \cdot x_{\text{team},n-2} + w_{n-1} \cdot x_{\text{team},n-1}$

(4) Computes $x'_{\text{team},n-3} \leftarrow x_{\text{team},n-3} - ((y - y')/(w_{n-3}))$

(5) Samples $x'_{\text{team},n-2} \xleftarrow{\$} \mathbb{Z}_q^n$, $x_{\text{team},n-2} \leftarrow x'_{\text{team},n-2}$, and Computes $x_{\text{team},n-1} \leftarrow (y - w_{n-2} \cdot x'_{\text{team},n-2})/w_{n-1}$

(6) For all $k \in \{1, \ldots, n-1\}$
$u_{\text{team},i} \cdot \mu_{\text{team},0} + u_{\text{team},[i+(n-1)] \bmod n} \cdot \mu_{\text{team},1} + \cdots + u_{\text{team},(i+1) \bmod n} \cdot \mu_{\text{team},n-1} \leftarrow x_{\text{team},i}$

Output: $\left\{\overrightarrow{u_{\text{team}}} = (u_{\text{team},0}, u_{\text{team},n-1}, \ldots, u_{\text{team},1})\right\}$

Algorithm 3: Calculation process chart of common downloading right.

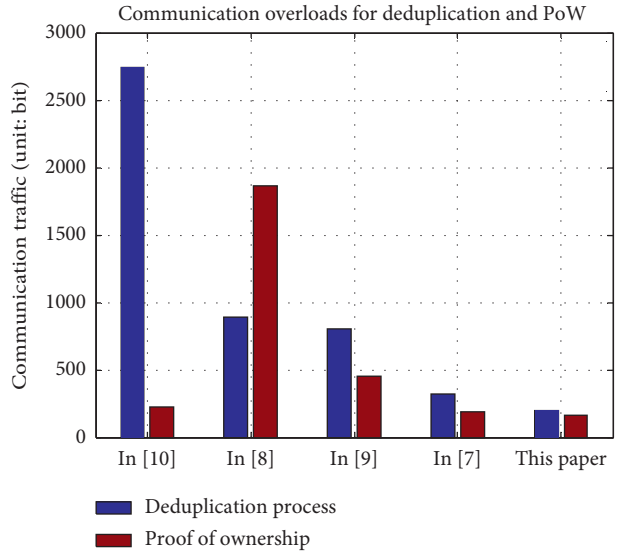


Figure 4: Histogram of communication cost of similar schemes.

Table 2: Function comparison between main data deduplication schemes.

Schemes	Technology	TTP	Level	Object	Key fusion
[8]	BL-MLE + PoW	—	Block	Single user	Yes
[3]	Threshold blind signature + verifiable secret sharing	Key servers	File	Single user	Yes
[2]	Authentication protocol + authorization detection	Cloud server	File	Multiple users	Yes
[9]	Attribute encryption + random sampling	Attribute center	Block	Multiple users	Yes
This paper	Attribute access policy + inner product predicate	—	File	Multiple users	No

Table 3: Computation overheads for deduplication.

Schemes	Client	TTP	Cloud
[8]	$O(b)$Hash · Hash	—	$O(b)$PoW
[3]	$O(f)$Hash + $O(N)$CE_K	$O(f)$Hash	$O(g)$
[2]	$O(f)$	$O(f)$	—
[9]	$O(f)$Hash + $O(f)$PoW	$O(f)$Hash	$O(f)$PoW
This paper	$O(f)$Hash + $O(1)$PSRG	—	$O(g)$

Table 4: Computation overheads for PoW.

Schemes	Client	TTP	Cloud
[8]	$O(b)$Hash + $O(b)$	—	$O(f)$Add
[3]	$O(f)$Hash + $O(kL)$	$O(N)$CE_K	$O(kLf)$Add
[2]	$O(f) + O(kL)$	—	$O(kLf)$Add
[9]	$O(f)$Hash + $O(N \log N)$	$O(N)$CE_K	$O(kLf)$Add
This paper	$O(f)$Hash + $O(n)$Add	—	$O(n)$Add

is $O(n)$Add. The reason is similar that the calculation cost on cloud side has nothing to do with the file size but only the number of attributes.

Finally, taking the file of 256 bits as an example, we compare the communication overhead for deduplication and PoW amongst the same set of schemes. The details are shown in Figure 4.

According to Figure 4, our scheme has obvious advantage on communication overheads compared with other schemes. Our solution can effectively reduce the usage of bandwidth as well as time delay. Moreover, since all deduplication check and ownership proof processes are independent, our scheme is capable of parallel processing, which is more fit for batch implementation.

10. Conclusions

This paper proposed a novel deduplication scheme based on LWE and FE to balance the conflict between the accessibility and the indistinguishability of data. Focusing on the purpose of deduplication check, LWE is exploited to construct deduplication labels which are distinguishable only if their deduplication certificates are known. To realize more efficient and flexible access control, inner product predicate is used that data can be retrieved only if both users downloading right and attributes vector are possessed. Thanks to the separation of downloading right and user's attributes, the downloading right can be recalculated for repeated uploading and authorization transfer without changing the corresponding deduplication label or download certificate over cloud. Correctness and security analyses proved that deduplication can be accomplished only by the duplication check process with negligible false positive, and it is almost impossible for any adversaries to fabricate a legal downloading right. Compared with other main technologies, our scheme is more applicable to multiuser environment and freed from trusted third-party. Since both duplication check and ownership proof are realized by inner product, the performances of computation and communication are more advantageous in our method, not mentioning its capacity of batch processing due to parallelism.

Data Availability

The data set was obtained from Chongqing Tongnan Electric Power Co., Ltd (telephone: 023-44559308; official website: http://www.12398.gov.cn/html/information/753078881/753078881201200006.shtml).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Bo Mi and Fengtian Kuang for their comments and suggestions. This work was supported in part by the National Natural Science Foundation of P.R. China (Grant nos. 61573076, 61703063, and 61903053); the Science and Technology Research Project of the Chongqing Municipal Education Commission of P.R. China (Grant nos. KJZD-K201800701, KJ1705121, and KJ1705139); and the Program of Chongqing Innovation and Entrepreneurship for Returned Overseas Scholars of P.R. China (Grant no. cx2018110).

References

[1] M. Wen, K. Ota, H. Li, and J. Lei, "Secure data deduplication with reliable key management for dynamic updates in cpss," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 4, pp. 1–11, 2016.

[2] J. Li, Y. K. Li, X. Chen, P. P. C. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 5, pp. 1206–1216, 2015.

[3] M. Miao, J. Wang, H. Li, and X. Chen, "Secure multi-server-aided data deduplication in cloud computing," *Pervasive and Mobile Computing*, vol. 24, pp. 129–137, 2015.

[4] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: deduplication in cloud storage," *IEEE Security & Privacy Magazine*, vol. 8, no. 6, pp. 40–47, 2010.

[5] B. Mi, D. Huang, S. Wan, L. Mi, and J. Cao, "Oblivious transfer based on NTRUEncrypt," *IEEE ACCESS*, vol. 6, pp. 35283–35291, 2018.

[6] L. Li, X. Chen, X. Huang et al., "Secure distributed deduplication systems with improved reliability," *IEEE Transactions on Computers*, vol. 64, no. 12, pp. 356-357, 2015.

[7] P. Puzio, R. Molva, M. Önen et al., "ClouDedup: secure deduplication with encrypted data for cloud storage," *5th International Conference on Cloud Computing Technology and Science (Colud-Com)*, vol. 1, pp. 363–370, 2013.

[8] R. Chen, Y. Mu, G. Yang, et al.,F. Guo, "BL-MLE: block-level message-locked encryption for secure large file deduplication," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2643–2652, 2015.

[9] D. Boneh, A. Sahai, and B. Waters, "Functional encryption: definitions and challenges," *Tcc*, vol. 2010, pp. 253–273, 2010.

[10] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," 2005.

[11] A. Zhou, S. Wang, S. Wan et al., "LMM: latency-aware microservice mashup in mobile edge computing environment," *Neural Computing and Applications*, vol. 4, no. 5, pp. 1–15, 2020.

[12] L. González-Manzano, J. M. D. Fuentes, and K. K. R. Choo, "Ase-PoW: a proof of ownership mechanism for cloud deduplication in hierarchical environments," 2016.

[13] Y. Chen, C. L. Li, J. L. Lan et al., "Secure sensitive data deduplication schemes based on deterministic/probabilistic proof of file ownership," *Journal on Communications*, vol. 36, no. 9, pp. 1–12, 2015.

[14] S. Wan, Y. Xia, L. Qi et al., "Automated colorization of a grayscale image with seed points propagation," *IEEE Transactions on Multimedia*, vol. 99, pp. 1–12, 2020.

[15] M. M. Xie, X. F. Liao, and Q. Zhou, "Generalized oblivious transfer protocol in distributed setting based on secret sharing," *Computer Engineering*, vol. 40, no. 3, pp. 184–187, 2014.

[16] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing*, vol. 398, pp. 520–530, 2020.

[17] J. Li, X. Chen, M. Li, J. Li, P. P. C. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 6, pp. 1615–1625, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8898062, 11 pages
https://doi.org/10.1155/2020/8898062

Research Article

Travel Time Reliability-Based Signal Timing Optimization for Urban Road Traffic Network Control

Zhengfeng Ma,[1,2] Darong Huang,[3] Changguang Li,[4] and Jianhua Guo[4]

[1]*School of Traffic & Transportation, Chongqing Jiaotong University, Chongqing 400074, China*
[2]*School of Civil & Transportation Engineering, Qinghai Nationalities University, Xining 810007, China*
[3]*School of Information Science & Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[4]*Intelligent Transportation System Research Center, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Darong Huang; drhuang@cqjtu.edu.cn and Jianhua Guo; gjh@seu.edu.cn

Received 14 September 2020; Revised 29 October 2020; Accepted 12 November 2020; Published 2 December 2020

Academic Editor: Esam Hafez Abdelhameed

Due to increasing traffic demand, many metropolitan areas are experiencing extensive traffic congestion, which demands for efficient traffic signal timing and optimization. However, conventional efficiency measure-based signal optimization cannot handle the ubiquitous uncertainty in the road networks, demanding for the incorporation of reliability measures into signal optimization, which is still in its early stage. Therefore, targeting this issue, based on the recent studies on recognizing travel time reliability (TRR) as an important reliability measure of road networks, a travel time reliability-based urban road traffic network signal timing optimization model is proposed in this paper, with the objective function to optimize a TTR measure, i.e., buffer time index. The proposed optimization model is solved using the heuristic particle swarm optimization approach. A case study is conducted using microscopic traffic simulation for a road network in the City of Nanjing, China. Results demonstrate that the proposed optimization model can improve travel time reliability of the road traffic network and the efficiency of the road traffic network as well. Future studies are recommended to expand the integration of travel time reliability into traffic signal timing optimization.

1. Introduction

Due to the increasing motorization and urbanization around the globe, congestion has become a pronounced phenomenon for many metropolitan areas (Huang et al. 2017) [1]. Consequently, many measures have been adopted to battle the worsening traffic congestion, with the traffic signal timing optimization as one of the most direct and effective strategies. However, due to many factors in the context of metropolitan areas, uncertainty is ubiquitous in urban transportation systems, and hence in addition to the conventional efficiency measures, the incorporation of reliability measures into traffic signal optimization to deal with traffic uncertainty is gaining increasing attention from different perspectives of the society.

Travel time reliability (TTR) is an important reliability measure, relating heavily to the variability of travel time. It is an important indicator for measuring the reliability of traffic systems. Travel time reliability is in general defined as the probability of a vehicle reaching the destination from the origin within a specified time. It can also be defined as the maximum time a traveler needs to arrive at the destination on time with a certain probability. According to different objectives, travel time reliability can be measured in terms of road segment travel time reliability, path travel time reliability, or road network travel time reliability. In this end, road segment travel time reliability refers to the probability a traveler completes the travel on a given road segment within a given period of time, path travel time reliability takes into account all the road segment travel time reliability in the path, and road network travel time reliability will incorporate the reliability of travel time over all OD pairs.

Even though the importance of travel time reliability has been acknowledged by traffic signal control practitioners or

scholars, the incorporation of travel time reliability into urban traffic signal timing optimization and control is still in its infancy. Currently, few studies have paid attention to optimize traffic signal timing based on travel time reliability, both for isolated intersections or network level signal timing optimization. Therefore, the objective of this paper is to propose an urban road traffic network signal timing optimization model based on optimizing travel time reliability of the road network. The heuristic approach of particle swarm optimization is applied to solve the proposed model, and microscopic simulation is used in a case study to implement and validate the proposed model for a road network in the City of Nanjing, China, as an example.

The rest of the paper is organized as follows. First, Section 2 provides a brief review on travel time reliability measures and travel time reliability-based signal optimization studies. Then, Section 3 presents the proposed travel time reliability based signal timing optimization model, together with the solution approach based on particle swarm optimization. Afterwards, a case study is conducted to implement and validate the proposed model, together with a comparison of the proposed model with the conventional travel time-based optimization model. Finally, the paper concludes with summaries and recommendations on future research.

2. Literature Review

In this section, travel time reliability measures are summarized, together with a brief review on travel time reliability-based signal timing optimization studies.

2.1. Travel Time Reliability Measures. Many travel time reliability measures have been proposed in the literature. In this end, commonly used travel time reliability measures include in general probabilistic indicators (Asakura (1996) [2]; Lo et al. (1999) [3]; Levinson and Zhang (2001) [4]), statistical indicators (Booz-Allen (1998) [5]; Recker et al. (2005) [6]; Sisiopiku and Rouphail (1994) [7]; Petty et al. (1998) [8]), buffer time indicators (Lomax et al. (2001) [9]; Chen et al. (2003) [10]; Lo (2002) [11]; Lo and Tung (2003) [12]; Lo and Luo (2004) [13]; Lo et al. (2006) [14]; Luo (2004) [15]; Siu and Lo (2008) [16]; Shao et al. (2006) [17]; Shao et al. (1985) [18]; Shao et al. (2008) [19]; Lam et al. (2008) [20]), and delay indicators (Lomax et al. (2003) [21]). In practice, probability indicators could be the distribution of travel time or percentile travel time, statistical indicators could be the average, median, or standard deviation of travel time, buffer time indicators could be buffer time or buffer time index of travel time, and delay indicators could be delay time or delay time index. As is clear from above descriptions, all these reliability indicators are helpful for transportation system managers to estimate the performance of the road network, and all these indicators can be tailored to accommodate the purpose of specific transportation applications.

2.2. TTR-Based Signal Timing Optimization. Currently, reliability-based traffic signal control is limited with insufficient applications. Heydecker modified the equation of control delay to show the randomness of traffic, and the randomness of control effectiveness is reflected by the correction of delay equation [22]. Although the concept is relatively easy, the steady state at the intersection is difficult to achieve at each cycle at higher saturation level. Kamarajugadda and Park used delay variance and average delay as optimization objectives to consider reliability in traffic signal control optimization, while delay variance is obtained by assuming a given delay distribution and the selected normal distribution needs further justification [23]. Hong studied the reliability of signalized intersections and used the randomness of intersection traffic signal control to characterize its reliability [24]. Using the phase clearance reliability (PCR) as the starting point, single-layer and multi-layer signal control models are adopted. Simulation results show that, under low saturation level, PCR can be greatly improved by increasing the traffic signal control cycle. Lu and Niu proposed a signal timing optimization model based on PCR [25]. According to the definition of PCR and the stochastic characteristics of arrival rate, quantitative relationship between PCR and parameters at each intersection is studied, and the equation of cycle and green time under given PCR can be derived. Application results showed that the randomness of queuing length at an intersection has great influence on signal parameters. Lu and Niu studied the influence of traffic flow randomness on traffic signal timing optimization at the intersection level [26]. PCR is expressed by expected offset of each phase, and traffic signal timing optimization model is established with the goal of minimizing the sum of all expected offsets. Example studies showed that longer green time is required for larger phase variance at the intersection under a given reliability level.

2.3. Summary. In summary, the study on reliability-based urban traffic road network signal timing and control is still in its infancy. First, studies in this field mostly focus on the framework and definition of the concept with limited applicable models and methods. In addition, current studies are mostly directed at isolated intersection with limited studies on reliability of network traffic signal control. Therefore, this paper proposes a travel time reliability based signal timing optimization model for urban road network signal timing optimization and control.

3. Proposed TTR-Based Signal Timing Model

In this section, the selected travel time reliability measure is described, and the proposed TTR-based signal timing model is presented together with the solution approach based on particle swarm optimization.

3.1. Travel Time Reliability Measure Selection. Travel time reliability measure is fundamental in the field of transportation system reliability optimization. In this end, considering the importance of travel time in measuring the performance of transportation systems, buffer time is defined as the extra travel time within a reasonable range to ensure an on-time arrival at the destination under uncertain

traffic conditions. In this sense, buffer time measures the reliability of road network from the perspective of travelers and hence can effectively assist the travelers in making reasonable travel plans to tackle traffic uncertainty. In addition, since travel time is closely related to the traveling distance, in order to measure the reliability of travel time consistently across the road network, buffer time index is developed through normalizing the buffer time with respect to the traveling distance. Therefore, in this paper, the buffer time index is used as the reliability indicator with its calculation as follows:

$$\mathrm{BTI} = \frac{T_{90} - \overline{T}}{\overline{T}}, \tag{1}$$

where T_{90} is the 90% percentile value of the travel time in the sample data and $\overline{T}$ is the average travel time. It should be noted that there is a balance between the selected percentile value and the efficiency performance of the optimized signal control system. In general, it is conjectured that the higher percentile value will introduce higher network reliability with reduced efficiency performance. Therefore, in order to ensure a preferable integrated system performance in terms of reliability and efficiency, 90% percentile travel time is selected in this paper when calculating the buffer time index.

Buffer time has many ramifications in transportation field, relating to factors such as purpose of traveler, travel mode, and psychological factors of traveler. Buffer time can be used for the comparison of the same road segment at different times and different road segments at the same time as well. Buffer time can reflect the changes in the accessibility and convenience of travel at different stages, and smaller buffer time indicates higher level of travel convenience and accessibility. Buffer time can also be used to determine the level of sustainable urban transport development for further road network optimization.

3.2. Proposed TTR-Based Optimization Model. The performance of traffic signal control system manifests the state of traffic flow movement under the control of a certain timing plan. The essence of establishing a traffic signal control model is to use mathematical or analytical methods to simulate the traffic flow movement on the road network and study the influence of changes in signal timing parameters on the movement of traffic, so as to objectively develop an optimized signal timing plan. The traffic model should be able to reliably assess the traffic movement parameters under the control of different traffic timing schemes.

In the abovementioned signal optimization process, delay is conventionally selected as efficiency measure for signal timing optimization. Delay is closely related to travel time. However, vehicle travel time is a random variable, and average travel time cannot reflect the actual traffic condition. For example, for heavily uncertain traffic, average travel time cannot accurately reflect the reliability of road network. Therefore, as discussed previously, reliability measure should be incorporated into signal optimization. As a typical travel time reliability measure, buffer time can be incorporated to develop a regional traffic signal timing optimization model. In this model, the average buffer time index of road segments in the road network can be minimized to improve road network reliability. In this end, the objective function of the model is defined as

$$y = \frac{1}{n} \sum_{i=1}^{n} \mathrm{BTI}_i, \tag{2}$$

where n represents the number of station pairs in the road network, BTI_i represents the buffer time index of the road segment for the i^{th} station pair in the road network, and y represents the optimization objective function. It should be emphasized that station pairs are counted according to adjacent intersections and directions are considered. For example, a road section can be counted as two station pairs according to different directions, and the buffer time index should be calculated separately in the model.

Next, constraints are set for the major signal control parameters, including offset, green time, and signal cycle. First, effective green time cannot be negative. Therefore, following constraints are listed as

$$\begin{aligned} g_{i,k} &\geq 0, \\ g_{i,\min} \leq g_{i,k} &\leq g_{i,\max}, \end{aligned} \tag{3}$$

where i denotes the intersection number in the road network, k denotes the phase number of the intersection, $g_{i,k}$ denotes the effective green time of the k^{th} phase of the i^{th} intersection, $g_{i,\min}$ denotes the lower limit of effective green time for the i^{th} intersection, and $g_{i-\max}$ represents the upper limit of the effective green time for the i^{th} intersection.

Second, the traffic signal cycle of an intersection cannot be negative. Therefore, following constraints are listed as

$$\begin{aligned} \sum_{k=1}^{m} g_{i,k} + L_i &= C_i, \\ C_{i,\min} \leq C_{i,k} &\leq C_{i,\max}, \end{aligned} \tag{4}$$

where m denotes the total number of phases for the intersection, L_i denotes the total loss time in the signal cycle of the i^{th} intersection, C_i denotes the traffic signal control cycle of the i^{th} intersection, $C_{i,\min}$ indicates the lower limit of the cycle for the i^{th} intersection, and $C_{i-\max}$ represents the upper limit of the cycle for the i^{th} intersection.

Similarly, phase offset in signal control cannot be negative, with the constraint listed as

$$\phi \geq 0, \tag{5}$$

where ϕ represents the phase offset between two intersections.

In summary, the travel time reliability-based urban road network traffic signal timing optimization model can be established as follows:

$$Z = \min y = \min \frac{1}{n} \sum_{i=1}^{n} \mathrm{BTI}_i$$

$$\text{s.t} \begin{cases} g_{i,k} \geq 0, \\ g_{i,\min} \leq g_{i,k} \leq g_{i,\max}, \\ \sum_{k=1}^{m} g_{i,k} + L_i = C_i, \\ C_{i,\min} \leq C_{i,k} \leq C_{i,\max}, \\ \phi \geq 0, \end{cases} \tag{6}$$

where $g_{i,\min}$ is set as 0 for both off-peak and peak hours, $g_{i,\max}$ is set as 50 seconds or 60 seconds for off-peak hours or peak hours, respectively, $C_{i,\min}$ is set as 0 for both off-peak and peak hours, and $C_{i,\max}$ is set as 150 seconds or 180 seconds for off-peak or peak hours, respectively.

3.3. Particle Swarm Optimization (PSO) Procedure. Particle swarm optimization (PSO) procedure is adopted in this paper to solve the proposed optimization model. Particle swarm algorithm originated from the foraging process of biological population or group. Each individual in the group is termed as a particle, and the space where the particle is located is termed as a D-dimensional space. The D-dimensional space represents the solution space of the optimization problem, and the position of each particle represents a solution. In order to move the particles in the D-dimensional space, i.e., to search the solution space, each particle is given a certain initial flight speed. In order to evaluate the location of a particle, that is, to evaluate the solution in the solution space, a fitness function must be defined. For PSO, through a sharing mechanism, the search information is shared from a global scope, and each particle changes the direction of advancement according to its own moving experience so that the entire population moves toward the global optimum value. In addition, particle swarm optimization uses the uncertainty of random factors and inertia weight to expand the search space and ensures the global convergence of the optimization algorithm.

During the movement of each particle in the D-dimensional space, the fitness function of its position is calculated and the maximum value of the fitness function of the particle in its own flight path is recorded as the optimal fitness value. The particle position corresponding to the optimal fitness value is recorded as the individual optimal value. For the entire group, there is only one location that attracts all particles. The optimal fitness values of all particles are compared and the largest fitness value is regarded as the global optimum fitness value. The particle position corresponding to the global optimal fitness value is recorded as the global optimal value, i.e., the solution to the optimization problem.

The flying speed of each particle is not fixed. After each population movement, the flying speed of each particle is updated using the velocity equation. Clerc and Kennedy improved the basic particle swarm algorithm and introduced a shrinkage factor in the velocity equation as below to ensure the convergence of the optimization process [27]:

$$\begin{aligned} V_{\mathrm{id}}^{t+1} &= K\left(V_{id}^{t} + c_1 \cdot \mathrm{rand}_1 \cdot \left(p\mathrm{best}_{\mathrm{id}}^{t} - V_{id}^{t}\right) + c_2 \cdot \mathrm{rand}_2\left(g\mathrm{best}_{d}^{t} - V_{id}^{t}\right)\right), \\ \theta &= c_1 + c_2, \\ K &= \frac{2}{\left|2 - \theta - \sqrt{\theta^2 - 4\theta}\right|}, \end{aligned} \tag{7}$$

where $p\mathrm{best}_{\mathrm{id}}^{t}$ is the d th dimensional element of the i^{th} particle in generation t; $g\mathrm{best}_{\mathrm{id}}^{t}$ is the best d^{th} dimensional element for all particles in generation t; rand_1 and rand_2 are uniformly distributed random numbers within [0, 1]; V_{id}^{t} is the i^{th} dimensional element representing particle speed; c_1 and c_2 are the accelerating factors with $c_1 = c_2 = 2.005$; and K is the shrinking factor.

In summary, given proper design of the particle swarm optimization problem, the general flowchart of implementing the particle swarm optimization is shown in Figure 1.

3.4. Particle Swarm Optimization Design. To solve the regional signal timing optimization issue using the particle swarm optimization procedure, mainly two aspects should be designed first. The first aspect is the parameters setting of the optimization algorithm, and the second aspect is determination of the fitness function.

3.4.1. Parameter Settings. According to the proposed optimization model, the traffic signal control parameters to be optimized include mainly intersection phase offset and green time of each phase. Therefore, each particle in the population must express green time for each phase and phase offset. Note that signal cycle for each intersection can be computed by summing up the green times for the corresponding signal phases. In summary, the structure of each particle is described in Figure 2, with the dimension of each particle set to 65.

In addition, the number of particle populations is set to 30. The corresponding velocity vector for each particle has a dimension of 65, and the total number of velocity vectors for all particles is 30. The maximum evolution generation is set to 100.

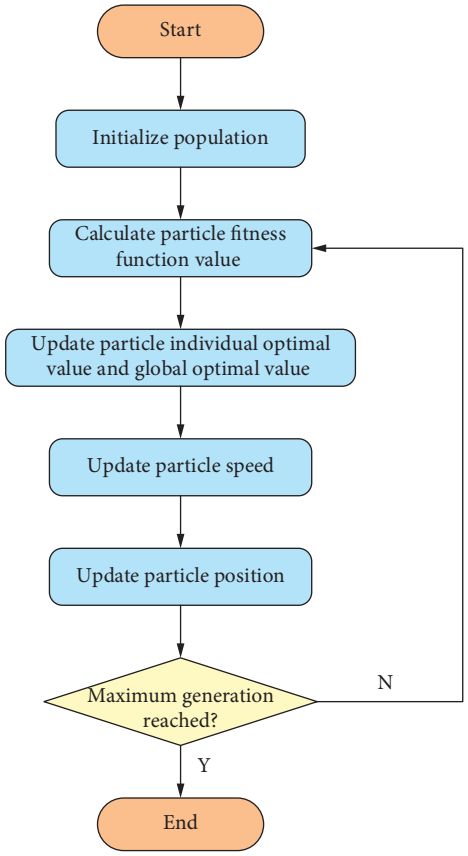


Figure 1: Flow chart of particle swarm algorithm optimization.

3.4.2. Fitness Function. Based on the proposed urban road network traffic signal timing optimization model, the average buffer time index of all road segments in the road network is used as the fitness function to evaluate the signal timing plan represented by each particle in the particle group. After running simulation, the buffer time index of all road segments in the road network can be calculated. Smaller average buffer time index shows that travelers do not need to reserve excessive extra time and the travel time of the road network is reliable. Therefore, the traffic signal timing plan can increase the reliability of the travel time of the road network. The equation for the fitness function is as follows:

$$\text{fitness}_2 = \frac{n}{\sum_{i=1}^{n} \text{BTI}_i}, \tag{8}$$

where n represents the number of station pairs in the road network and BTI_i represents the buffer time index of the segment along the i^{th} station pair of the road network.

4. Case Study

This paper proposed an urban road network traffic signal timing optimization model, which can be solved using the heuristic particle swarm optimization procedure. In this section, the proposed model is implemented and validated in a microscopic simulation environment for a real-world urban road network. Note that microscopic traffic simulation software Paramics is selected in this study due to its flexible programing ability provided through abundant Application Programming Interfaces (APIs).

4.1. Study Area and Data Collection. The study area selected in this paper is a region in the city of Nanjing, China. For this road network, 22 radio frequency identification (RFID) base stations are installed, collecting individual vehicle passing records continuously. These base stations are in general located along Zhujiang Road, East Zhongshan Road, Ruijin Road, Middle Longpan Road, and Yu Dao Street. The selected road network and the locations of the RFID base stations are shown in Figure 3, and the overview of the intersections within this road network is shown in Table 1.

RFID is a noncontact automatic identification technology. Noncontact two-way radio communication is employed to automatically recognize target objects, and therefore, for each vehicle equipped with a RFID tag passing a certain RFID base station, a vehicle passing record will be generated, with collected information primarily including base station number, passing time, and vehicle license plate number. From these vehicle passing records, travel time between base station pairs can be obtained by matching the recorded information at the starting base station and the destination base station. The RFID base station pairs are listed in Table 2 for the selected road network. For more information on processing RFID data, readers can refer to [28].

In addition, for signal timing parameters of this road network, primarily the turning information and the phase setting information are collected manually for both peak hours and off-peak hours. For these intersections, 7 intersections have 5 phases, 2 intersections have 7 phases, and 1 intersection has 6 phases. The overview of the signal timing setting is shown in Tables 3 and 4 for peak hours and off-peak hours, respectively.

4.2. Comparison Models and Performance Measures. Three models will be implemented and compared in this case study, as listed in Table 5. Original_Plan indicates the current timing plan without optimization. TTR_Plan is the timing plan generated using the proposed travel time reliability based optimization model. TT_Plan is the timing plan generated using minimum mean travel time as the optimization objective with the objective function defined as

$$Z = \min \frac{1}{n} \sum_{i=1}^{n} \overline{TT}_i, \tag{9}$$

where $\overline{TT}_i$ is the average travel time of the road segment i in the road network. Note that TT_Plan is also solved using the particle swarm optimization technique, and the fitness function for PSO is as follows:

$$\text{fitness}_3 = \frac{n}{\sum_{i=1}^{n} \overline{TT}_i}. \tag{10}$$

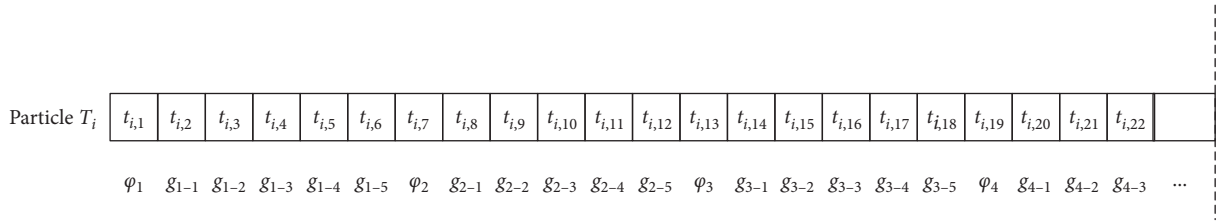


Figure 2: Particle structure.

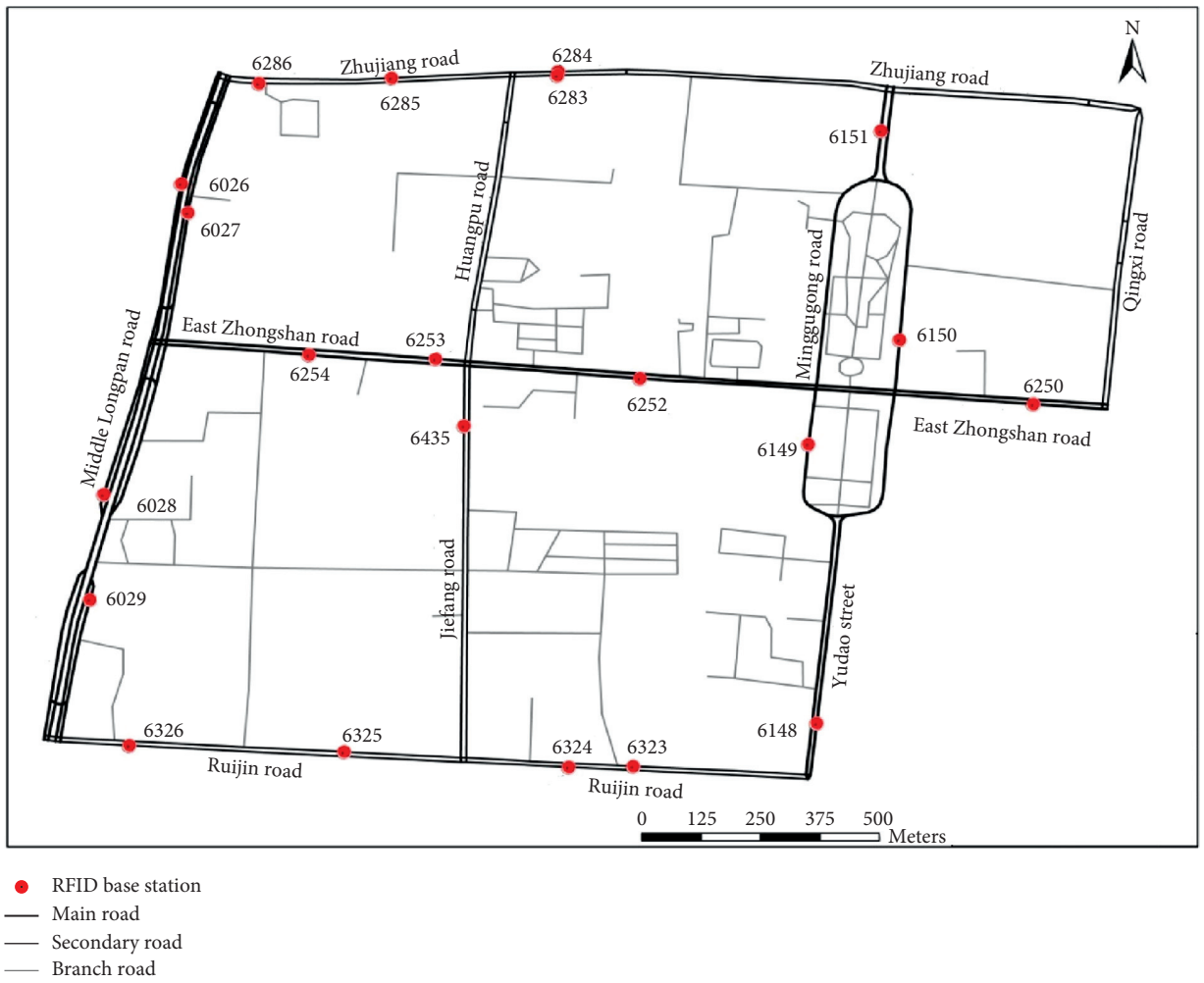


Figure 3: Selected road network.

In order to compare these three models, four performance measures are selected, including travel time of road network (NTT), buffer time index (BTI), queue length at the intersection (QL), and delay of the road (DR). Note that these performance measures are calculated for the simulated road network, which can provide detailed traffic network condition data for computing these measures.

4.3. Simulation Model Calibration. Before model validation and comparison, it is necessary to calibrate the simulated road network in the simulation software, i.e., to adjust the traffic volume for each OD pair in the simulated network, so that the simulated road network will reflect truthfully the real world road network. For this purpose, the particle swarm optimization technique is used for simulation model calibration, as presented below.

4.3.1. Parameter Settings of OD Calibration Algorithm. In the road network, there are 144 OD pairs to be calibrated. Therefore, each particle in the defined particle group will

Table 1: Road network intersection overview.

No.	Intersection location
J_1	Zhujiang rd./Middle longpan rd.
J_2	Huangpu rd./Zhujiang road rd.
J_3	Zhujiang rd./Beianmen bridge rd
J_4	Middle longpan rd./East Zhongshan rd.
J_5	East Zhongshan rd./Huangpu rd.
J_6	East Zhongshan rd./Minggugong rd.
J_7	East Zhongshan rd./Minggugong rd.
J_8	Middle longpan rd./Ruijin rd.
J_9	Ruijin rd./Jiefang rd.
J_{10}	Ruijin rd./Yudao st.

Table 2: RFID base station pair overview.

No.	Start station	End station
1	6148	6251
2	6150	6283
3	6151	6250
4	6026	6254
5	6028	6326
6	6251	6435
7	6252	6150
8	6253	6028
9	6283	6435
10	6285	6026
11	6324	6148
12	6435	6324
13	6148	6250
14	6149	6323
15	6151	6251
16	6027	6286
17	6029	6254
18	6252	6149
19	6253	6027
20	6254	6435
21	6284	6151
22	6286	6435
23	6325	6029
24	6435	6325
25	6286	6284
26	6283	6285
27	6026	6028
28	6029	6027
29	6254	6252
30	6252	6250
31	6251	6253
32	6326	6324
33	6323	6325
34	6148	6150
35	6151	6149

have 144 elements, each of which corresponds to a volume of an OD pair, with the structure of the particle shown in Figure 4.

In addition to the structure design of the particle, the number of particles is set to 30. The corresponding velocity vector of each particle has a dimension of 144, and there are 30 particle velocity vectors. The maximal evolution generation is set to 100.

Table 3: Original signal timing parameters for peak hours (unit: second).

Timing parameter	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
G_{i-1}	48	15	13	15	16	20	20	13	55	21
G_{i-2}	40	25	28	40	17	33	35	11	20	33
G_{i-3}	12	35	34	20	48	58	56	20	25	21
G_{i-4}	21	30	20	22	15	20	20	40	12	24
G_{i-5}	44	20	40	10	19	24	24	10	23	40
G_{i-6}	—	—	—	—	—	—	—	—	—	13
G_{i-7}	—	—	—	—	—	—	—	25	—	—
Y_i	3	3	3	3	3	3	3	3	3	3
C_i	180	140	150	155	130	170	170	170	150	170

Table 4: Original signal timing parameters for off-peak hours (unit: second).

Timing parameter	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
G_{i-1}	38	12	10	12	15	18	23	12	40	15
G_{i-2}	33	20	20	28	12	30	27	10	15	22
G_{i-3}	12	28	25	15	38	45	42	18	20	20
G_{i-4}	22	25	20	17	10	15	18	27	10	18
G_{i-5}	30	15	30	12	15	22	15	10	20	27
G_{i-6}	—	—	—	15	—	—	—	22	—	10
G_{i-7}	—	—	—	10	—	—	—	20	—	—
Y_i	3	3	3	3	3	3	3	3	3	3
C_i	150	115	120	130	105	145	140	140	120	130

Table 5: Comparative models.

Model abbreviations	Description
Original_Plan	Original timing plan
TTR_Plan	Optimized plan based on maximum TTR
TT_Plan	Optimized plan based on minimum TT

4.3.2. Fitness Function. According to the positions of RFID base stations in the road network, vehicle detectors are set in the simulated road network, counting number of vehicles passing the detectors during the simulation. Note that the difference between simulated traffic volume and real world traffic volume indicates the closeness of the simulated network to the real world network. Consequently, this difference is used to build the fitness function of the particle swarm optimization algorithm, as follows:

$$\text{fitness}_1 = \frac{1}{\left(\text{sum}\left(\left|\text{rfid}_i - \text{vde}_i\right|\right)/22\right) + 1}, \tag{11}$$

where rfid_i denotes the real world traffic volume detected by the i^{th} RFID base station and vde_i denotes the simulated traffic volume detected by the i^{th} vehicle detector.

4.3.3. Calibration Result. Using the designed particle swarm optimization algorithm, the simulated road network was calibrated for both peak hours and off-peak hours, with the pattern of the fitness function values shown in Figures 5 and 6, respectively. Clearly, with the progress of optimization,

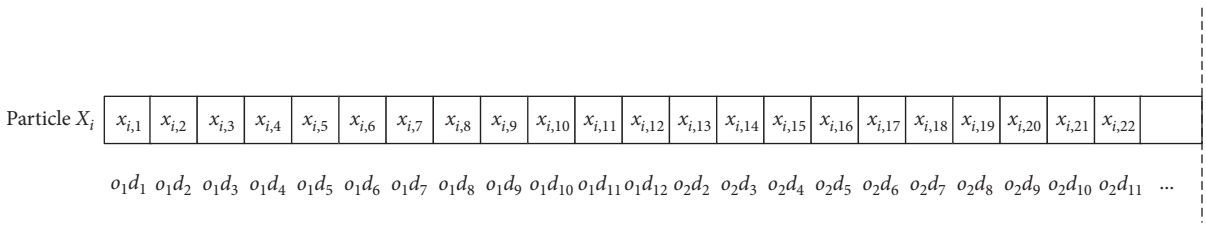


Figure 4: Structure of OD pair volume particles.

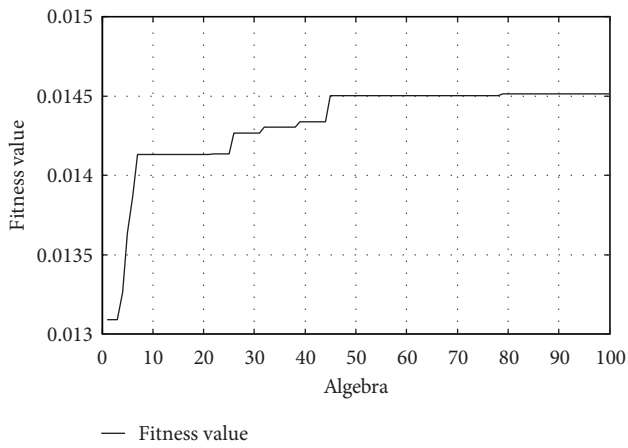


Figure 5: Calibration fitness function pattern for peak hours.

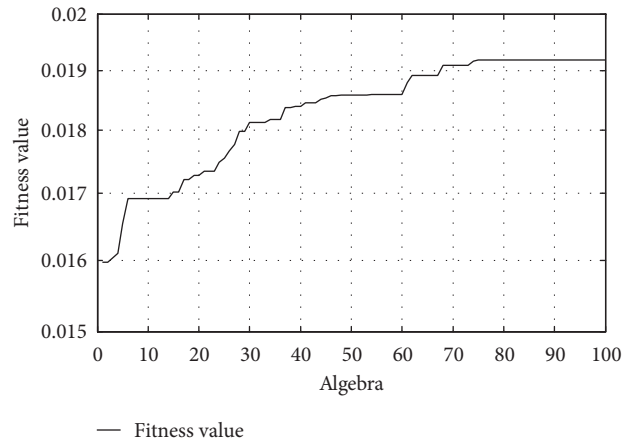


Figure 6: Calibration fitness function pattern for off-peak hours.

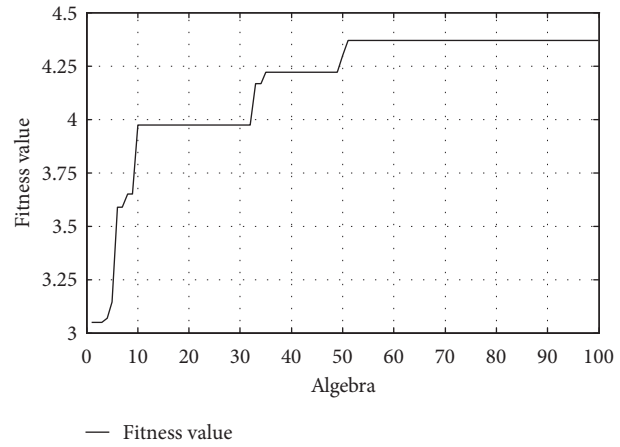


Figure 7: TTR_Plan fitness function pattern for peak hours.

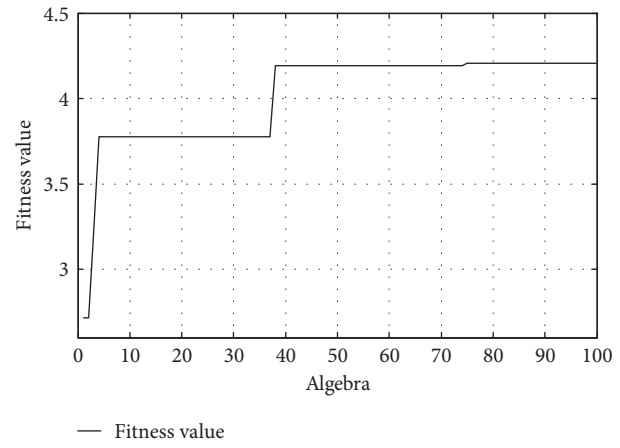


Figure 8: TTR_Plan fitness function pattern for off-peak hours.

the difference between the simulated and real world traffic volumes decreases continuously, and to the end of the optimization, the differences remain stable, indicating the convergence of the calibration process, for both peak hours and off-peak hours.

4.4. TTR_Plan Result. Using the calibrated road network, TTR_Plan was implemented. Figures 7 and 8 show the pattern of fitness function values of TTR_Plan during peak and off-peak hours, respectively. It can be seen that for both peak hours and off-peak hours, the fitness function value gradually increases as the optimization iteration proceeds, indicating a continuous decrease of average buffer time index of the road network, i.e., a continuous improvement of the reliability of travel time in the road network.

The optimized signal timing settings are shown in Tables 6 and 7 for peak hours and off-peak hours, respectively. Clearly, the proposed model adjusted the signal timing settings for all the intersections, compared with the signal timing settings in Original_Plan.

4.5. TT_Plan Result. Using the calibrated road network, TT_Plan was implemented. Figures 9 and 10 show the pattern of fitness function values of TT_Plan during peak and off-peak hours, respectively. It can be seen that TT_Plan

Table 6: TTR_Plan signal timing parameters for peak hours (unit: second).

Timing parameters	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
G_{i-1}	28	9	21	30	30	53	42	11	24	16
G_{i-2}	25	38	33	33	34	16	28	8	32	28
G_{i-3}	40	58	27	1	51	4	34	21	42	46
G_{i-4}	40	30	24	51	30	47	37	16	14	23
G_{i-5}	32	30	60	4	20	45	23	26	53	30
G_{i-6}	—	—	—	8	—	—	—	29	—	18
G_{i-7}	—	—	—	32	—	—	—	47	—	—
φ_i	70	123	64	91	117	31	65	59	110	49
Y_i	3	3	3	3	3	3	3	3	3	3
C_i	180	180	180	180	180	180	179	179	180	179

Table 7: TTR_Plan signal timing parameters for off-peak hours (unit: second).

Timing parameters	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
G_{i-1}	44	20	35	23	16	24	28	20	36	23
G_{i-2}	29	18	37	14	13	20	28	26	42	13
G_{i-3}	16	9	32	12	50	30	38	23	21	13
G_{i-4}	19	43	16	13	11	38	25	17	27	28
G_{i-5}	27	44	15	22	16	23	16	15	9	24
G_{i-6}	—	—	—	13	—	—	—	14	—	30
G_{i-7}	—	—	—	32	—	—	—	14	—	—
Φi	41	50	62	14	77	90	23	80	87	90
Y_i	3	3	3	3	3	3	3	3	3	3
C_i	150	149	150	150	121	150	150	150	150	149

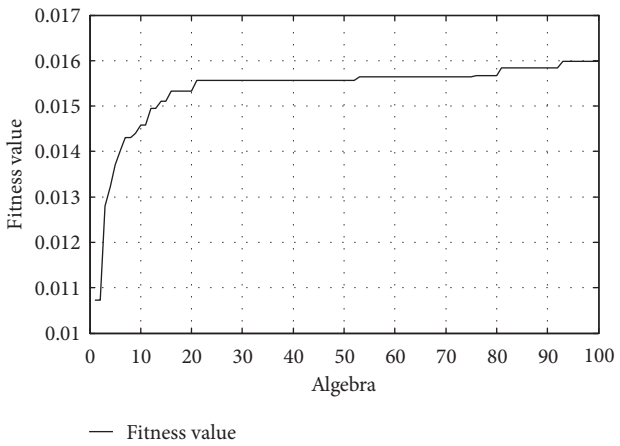


Figure 9: TT_Plan fitness function pattern for peak hours.

shows the same pattern as TTR_Plan, for both peak hours and off-peak hours. This indicates that TT_Plan improves network performance in terms of average travel time. However, no inference on the reliability of travel time can be drawn as travel time reliability measure is not incorporated in the optimization process.

Similarly, the optimized signal timing settings are shown in Tables 8 and 9 for peak hours and off-peak hours, respectively. Clearly, TT_Plan also adjusted differently the

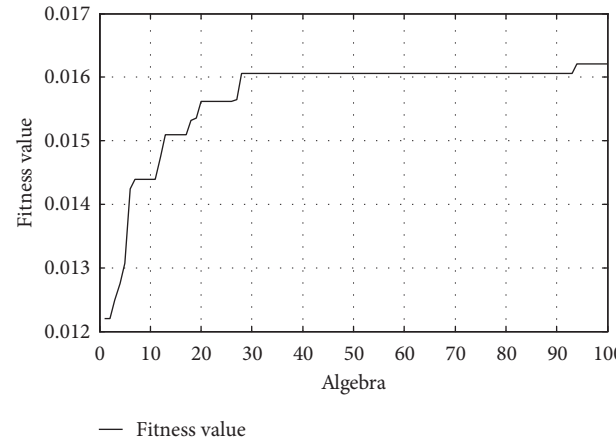


Figure 10: TT_Plan fitness function pattern for off-peak hours.

Table 8: TT_Plan signal timing parameters for peak hours (unit: second).

Timing parameter	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
G_{i-1}	9	31	20	25	16	38	34	34	34	36
G_{i-2}	52	28	21	22	17	22	24	6	32	21
G_{i-3}	30	50	43	22	60	37	36	22	23	25
G_{i-4}	43	30	23	15	15	32	34	30	30	35
G_{i-5}	31	26	57	16	19	36	37	12	46	27
G_{i-6}	—	—	—	30	—	—	—	37	—	17
G_{i-7}	—	—	—	28	—	—	—	17	—	—
φ_i	57	107	102	126	51	95	49	75	52	85
Y_i	3	3	3	3	3	3	3	3	3	3
C_i	180	180	179	179	142	180	180	179	180	179

Table 9: TT_Plan signal timing parameters for off-peak hours (unit: second).

Timing parameter	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
G_{i-1}	47	4	29	8	16	18	30	30	25	20
G_{i-2}	24	26	30	32	12	33	38	3	36	23
G_{i-3}	15	39	30	26	50	27	14	11	6	3
G_{i-4}	20	31	34	20	11	29	19	24	32	39
G_{i-5}	30	35	12	6	16	28	35	22	36	27
G_{i-6}	—	—	—	21	—	—	—	14	—	20
G_{i-7}	—	—	—	14	—	—	—	24	—	—
φ_i	51	47	93	64	69	64	60	50	61	64
Y_i	3	3	3	3	3	3	3	3	3	3
C_i	151	150	150	148	120	150	151	149	150	150

signal timing settings for all the intersections, compared with the signal timing settings in Original_Plan.

4.6. Performance Comparisons. Using optimized timing plans given above for TTR_Plan and TT_Plan, the performances of the three models can be compared quantitatively in terms of four performance measures, i.e., travel time of road network (NTT), buffer time index (BTI), queue length at the intersection (QL), and delay of the road (DR).

Table 10: Performance comparison for peak hours.

Timing plan	NTT (s)	QL (m)	DR (s)	BTI
Original_Plan	440.89	64.63	75.76	0.5245
TTR_Plan	437.67 (−0.73%)	48.18 (−25.45%)	28.39 (−62.53%)	0.2336 (−55.46%)
TT_Plan	357.22 (−18.98%)	41.13 (−36.36%)	45.99 (−39.30%)	0.3733 (−28.83%)

Table 11: Performance comparison for off-peak hour.

Timing plan	NTT (s)	QL (m)	DR (s)	BTI
Original_Plan	362.22	49.58	56.12	0.5269
TTR_Plan	318.44 (−12.09%)	36.93 (−25.51%)	28.39 (−49.41%)	0.3495 (−33.67%)
TT_Plan	345.67 (−4.57%)	39.93 (−19.46%)	41.09 (−26.78%)	0.4583 (−13.02%)

Table 10 lists the performance measures of the three models for peak hours. On observing Table 10, first, it can be seen that compared with Original_Plan, both TTR_Plan and TT_Plan show significant improvement in terms of all the measures. This indicates a significant margin of improving the original timing through applying optimization technique. Second, TTR_Plan outperforms TT_Plan in terms of BTI, while TT_Plan outperforms TTR_Plan in terms of NTT, which is in alignment with the optimization objective of these two models. Third, in terms of QL and DR, the performances of TTR_Plan and TT_Plan are mixed, indicating comparable performance of TTR_Plan and TT_Plan. In summary, both signal timing optimization plans can improve the network performance over the original timing plans, while the two optimization plans show comparable performances. On reflection, this might be caused by high traffic level where there might be less room left for optimization.

Table 11 lists the performance measures of the three models for off-peak hours. On observing Table 11, first, it is clear that both TTR_Plan and TT_Plan outperform significantly the Originial_Plan, which indicates that optimization technique can improve the performance of signal timing for off-peak hours. Second, different from the results for peak hours, for all the four performance measures, TTR_Plan consistently outperforms TT_Plan. This is an interesting finding, indicating that minimizing travel time reliability might at the same time minimize travel time for off-peak hours. On reflection, this might be caused by the existence of excessive room left for off-peak traffic levels to first balance travel time and then reduce the level of average travel time before reaching the minimized average buffer time index.

In summary, it is clear that for both peak hours and off-peak hours, the proposed travel time reliability-based signal optimization model can improve the performance of urban road network, in terms of both efficiency and reliability. In particular, for off-peak hours, the proposed model shows a consistent improvement of network efficiency and network reliability over models of minimizing average travel time only.

5. Conclusions

Recently, travel time reliability has become an important performance measure of the urban traffic network. However, in urban traffic signal control systems, travel time reliability has not been sufficiently investigated. Therefore, more research is needed to understand the performance of network traffic signal control with the objective to optimize travel time reliability. To this end, an urban traffic network signal timing optimization model is proposed in this paper to optimize the average buffer time index of all road segments in the network. Particle swarm algorithm is adopted to solve the optimization models for both peak and off-peak hours. A case study is conducted for a road network in Nanjing city. The results show that the proposed travel time reliability-based signal timing optimization model can significantly improve the reliability of traffic network and efficiency as well, in particular for off-peak hours when excessive room is available for traffic signal optimization.

Considering the importance of incorporating reliability measures into real-world traffic management and control, future research is recommended as follows. First, more travel time reliability measures can be investigated in urban traffic network signal timing optimization. In particular, the effect of travel time percentile can be further investigated to show its effect on the reliability of the optimized system. Second, studies should be conducted to relate travel time reliability measures to the traffic condition uncertainty models as developed by Guo et al. (2008, 2012, 2014) [29–31] and Shi et al. [32], so that signal timing could be directly related to uncertain traffic conditions. Third, more advanced traffic signal optimization methods such as reinforced learning approach could be investigated together with the reliability measures. Finally, and most importantly, online methods are expected to be developed to meet the real world requirement of urban traffic signal optimization and control.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of P. R. China, under Grants 61573076, 61903053, and 61703063; the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJZD-K201800701; and the Program of Chongqing Innovation and Entrepreneurship for Returned Overseas Scholars of P.R. China under Grant cx2018110.

References

[1] W. Huang, Y. Wei, J. Guo, and J. Cao, "Next-generation innovation and development of intelligent transportation system in China," *Science China Information Sciences*, vol. 60, no. 11, 2017.

[2] Y. Asakura, "Reliability measures of an origin and destination pair in a deteriorated road network with variable flows," in *Proceeding of the 4th Meeting of the EURO*, pp. 273–287, Newcastle, UK, June 1996.

[3] K. Lo, H. Yang, and W. Tang, "Combining performance measure of a road network," in *Proceedings of the Hong Kong Society for Transportation Studies. Travel Time Proceedings Hong Kong, and Capacity Reliability for of the 4th Conference of 1999*, Hong Kong, December 1999.

[4] D. Levinson and L. Zhang, "Travel time variability after a shock: the case of the twin cities ramp meter shut off," in *Proceedings of the the First International Symposium on Transportation Network Reliability, IN STR 2001*, pp. 1–20, Kyoto, Japan, 2001.

[5] Booz-Allen, "California transportation plan: transportation system performance measures," Final Report, California Department of Transportation, Transportation System Information Program, Sacramento, CA, USA, 1998.

[6] W. Recker, Y. Chung, J. Park et al., "Considering risk-taking behavior in travel time reliability," California Partners for Advanced Transit and Highways (PATH), Richmond, CA, USA, Paper UCB-ITS-PRR-2005-3, 2005.

[7] V. Sisiopiku and N. Rouphail, "Towards the use of detector output for arterial link travel time estimation: a literature review," *Transportation Research Record*, no. 1457, pp. 158–165, 1994.

[8] K. F. Petty, P. Bickel, M. Ostland et al., "Accurate estimation of travel times from single-loop detectors," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 1, p. 1, 1998.

[9] T. Lomax, S. Turner, and R. Margiotta, "Monitoring urban roadways in 2000: using archived operations data for reliability and mobility measurement," Texas Transportation Institute, the Texas A & M University System, College Station, TX, USA, FHWA-OP-02-029, 2001.

[10] C. Chen, A. Skabardonis, and P. Varaiya, "Travel-time reliability as a measure of service," *Transportation Research Record*, no. 1855, pp. 74–79, 2003.

[11] H. Lo, "Trip travel time reliability in degradable transport networks," in *Proceedings of the 15th International Symposium on Transportation and Traffic Theory, ISTTT*, pp. 541–560, Adelaide, Australia, July 2002.

[12] H. K. Lo and Y.-K. Tung, "Network with degradable links: capacity analysis and design," *Transportation Research Part B: Methodological*, vol. 37, no. 4, p. 345, 2003.

[13] H. Lo and X. Luo, "Route choice behavior in degradable transport networks," in *Proceedings of the 8th International Conference on Applications of Advanced Technologies in Transportation Engineering*, pp. 61–65, Beijing, China, May 2004.

[14] H. K. Lo, X. W. Luo, and B. W. Y. Siu, "Degradable transport network: travel time budget of travelers with heterogeneous risk aversion," *Transportation Research Part B: Methodological*, vol. 40, no. 9, p. 792, 2006.

[15] X. Luo, "Transport network with degradable links and stochastic demands," M.Phil. Dissertation, Hong Kong University of Science and Technology, Hong Kong, 2004.

[16] B. W. Y. Siu and H. K. Lo, "Doubly uncertain transportation network: degradable capacity and stochastic demand," *European Journal of Operational Research*, vol. 191, no. 1, pp. 166–181, 2008.

[17] H. Shao, W. H. K. Lam, and M. L. Tam, "A reliability-based stochastic traffic assignment model for network with multiple user classes under uncertainty in demand," *Networks and Spatial Economics*, vol. 6, no. 3-4, pp. 173–204, 2006.

[18] H. Shao, W. H. K. Lam, Q. Meng, and M. L. Tam, "Demand-driven traffic assignment problem based on travel time reliability," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1985, no. 1, pp. 220–230, 2006.

[19] H. Shao, W. H. K. Lam, M. L. Tam, and X.-M. Yuan, "Modelling rain effects on risk-taking behaviours of multi-user classes in road networks with uncertainty," *Journal of Advanced Transportation*, vol. 42, no. 3, pp. 265–290, 2008.

[20] W. H. K. Lam, H. Shao, and A. Sumalee, "Modeling impacts of adverse weather conditions on a road network with uncertainties in demand and supply," *Transportation Research Part B: Methodological*, vol. 42, no. 10, pp. 890–910, 2008.

[21] T. Lomax, D. Schrank, S. Turner, and R. Margiotta, *Selecting Travel Reliability Measures*, Texas Transportation Institute, Cambridge Systematics, Inc., Cambridge, MA, USA, 2003.

[22] B. Heydecker, "Treatment of random variability in traffic modeling," in *Proceedings of the Workshop Traffic Granular Flow*, Julich, Germany, October 1995.

[23] A. Kamarajugadda and B. Park, "Stochastic traffic signal timing optimization," Ph.D. dissertation, University of Virginia, Charlottesville, VA, USA, 2003.

[24] K. L. Hong, "A reliability framework for traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 45–50, 2006.

[25] B. Lu and H. Niu, "Modeling and Simulation of the reliability of intersection traffic signal control," *Transportation System Engineering and Information*, vol. 6, pp. 45–50, 2011.

[26] B. Lu and H. Niu, "Single-point intersection traffic signal timing optimization under random conditions," *Journal of Transportation Engineering*, vol. 6, pp. 116–120, 2010.

[27] M. Clerc and J. Kennedy, "The particle swarm—explosion, stability, and convergence in a multidimensional complex space," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, 2002.

[28] J. Guo, C. G. Li, X. Qin et al., "Analyzing distributions for travel time data collected using radio frequency identification technique in urban road networks," *Science China Technological Sciences*, vol. 62, pp. 106–120, 2019.

[29] J. Guo, B. Williams, and B. Smith, "Data collection time intervals for stochastic short-term traffic flow forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2024, no. 1, pp. 18–26, 2007.

[30] J. Guo, W. Huang, and B. M. Williams, "Integrated heteroscedasticity test for vehicular traffic condition series," *Journal of Transportation Engineering*, vol. 138, no. 9, pp. 1161–1170, 2012.

[31] J. Guo, W. Huang, and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[32] G. Shi, J. Guo, W. Huang, and B. Williams, "Modeling seasonal heteroscedasticity in vehicular traffic condition series using seasonal adjustment approach," *ASCE Journal of Transportation Engineering*, vol. 140, no. 5, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6617291, 9 pages
https://doi.org/10.1155/2020/6617291

Research Article

Analysis of Vibration and Noise for the Powertrain System of Electric Vehicles under Speed-Varying Operating Conditions

Chenghao Deng,[1,2] **Qingpeng Deng**,[2] **Weiguo Liu**,[2] **Cheng Yu**,[2] **Jianjun Hu**,[1] **and Xiaofeng Li**[3]

[1]*School of Automotive Engineering, Chongqing University, Chongqing 400040, China*
[2]*Chongqing Chang'an New Energy Automobile Technology Co. Ltd., Chongqing 401120, China*
[3]*Chongqing Deyin Technology Co. Ltd., Chongqing 400050, China*

Correspondence should be addressed to Chenghao Deng; dengch@changan.com.cn and Qingpeng Deng; dqp_2017@163.com

Received 10 October 2020; Revised 1 November 2020; Accepted 5 November 2020; Published 23 November 2020

Academic Editor: Yong Chen

Whine noise from the electric powertrain system of electric vehicles, including electromagnetic noise and gear-meshing noise, significantly affects vehicle comfort and has been getting growing concern. In order to identify and avoid whine problems as early as possible in the powertrain development process, this paper presents a vibration and noise simulation methodology for the electric powertrain system of vehicles under speed-varying operating conditions. The electromagnetic forces on the stator teeth of the motor and the bearing forces on the gearbox for several constant-speed operating conditions are obtained first by electromagnetic field simulation and multi-body dynamic simulation, respectively. Order forces for the speed-varying operating condition are generated by interpolation between the obtained forces, before they are applied on the mechanical model whose natural modes have been calibrated in advance by tested modes. The whine noise radiated from the powertrain is then obtained based on acoustic boundary element analysis. The simulated bearing forces indicate that the overlooking of the motor torque ripple does not result in significant loss in simulation accuracy of electromagnetic noise. The simulation results and tested data show good consistency, with the relative frequency deviation of local peaks being less than 8% and the error of the average sound pressure level (SPL) being mostly below 10 dB (A).

1. Introduction

The electric vehicle industry has achieved rapid development in recent years. The high-frequency electromagnetic and gear whine noise emitted from the electric powertrain system could significantly affect driving comfort and has become an important noise, vibration, and harshness (NVH) problem of electric vehicles. The motor and reducer are two main vibration and noise sources of the electric powertrain system. The electromagnetic forces of the motor and the gear-meshing forces of the reducer could cause structural vibration and whine noise that shows obvious order characteristics.

The integration design of the electric drive system has become a technological trend, which means that more functional units, such as the motor, reducer, motor control unit, and power supply, are integrated into just one drive unit, i.e., the electric powertrain. The integration of the electric drive system can significantly reduce the volume, weight, and cost of the powertrain system and hence gain competitive advantages in the market place. As a result, the physical boundaries between components turn increasingly vague, and the mechanical coupling between the component structures become stronger, which introduces challenges to NVH analysis and control [1]. The generation mechanism of powertrain whine noise is consistent with that of the motor and reducer [2]. However, the coupling effect between the components will significantly affect the characteristics of whine noise. For instance, the tangential electromagnetic forces acting on the stator teeth, which can be ignored in the noise analysis of a standalone motor, turn

to be nonnegligible in the NVH simulation of the integrated powertrain, as the tangential electromagnetic forces could excite the reducer housing to vibrate and radiate noise [3]. Fang and Zhang [4] analyzed the vibration characteristics of the electric drive system through simulation and test, revealing that the motor, reducer, and controller, due to the coupling phenomenon, must be considered as an indivisible whole in the NVH analysis. Vibration and sound simulation based on computer-aided engineering (CAE) is an important approach for analyzing and optimizing electric powertrain noise, especially in the early phase of design. Harris et al. [5] introduced a CAE method to optimize the whine problem caused by gear-meshing excitation. The dynamic meshing force at the contact point of gears is reduced by changing the geometry of the rim and web of the gears. Yu et al. [6, 7] established a finite element (FE) model of the electric drive assembly system to predict its vibration. The simulation model reflects part of the frequency characteristics of vibration, but quantitative evaluation on the vibration simulation accuracy is not provided. The key challenge of NVH simulation for the highly integrated electric powertrain system is to efficiently calculate the whine noise caused by various types of excitations, such as electromagnetic forces and gear-meshing forces, within the whole speed range and with good simulation accuracy. However, the NVH simulation method for integrated electric powertrain with both satisfactory accuracy and efficiency is rarely reported.

In this paper, NVH simulation analysis for an integrated electric powertrain system under electromagnetic and gear-meshing excitations is performed. The electromagnetic forces on the stator teeth and gear-meshing forces acting on the bearings in the time domain for several constant-speed operating conditions are obtained first by electromagnetic simulation and multi-body dynamic simulation. Forces in the frequency domain are then obtained by performing fast Fourier transform (FFT). Cubic spline interpolation is utilized to obtain the order forces under the speed-varying condition, which significantly cuts down the time required for multi-condition force simulation. In order to ensure the simulation accuracy, the material parameters of the motor stator are calibrated by performing modal correlation analysis of the tested modes and the simulated modes before the FE model is used for any dynamic simulation, including the multi-body analysis and vibration simulation. The acoustic transfer vector (ATV) from the surface vibration of the powertrain housing to the sound pressure of the observing point is calculated by the acoustic FE analysis. Finally, the radiated sound pressure is calculated by using the obtained housing vibration velocity and ATV. Based on the simulation model, the influence of the motor torque ripple on the whine noise is evaluated. The effectiveness of the simulation method is verified by using tested results. The influence of the stator breathing mode on the 48-order whine noise is revealed. The main contribution of the paper lies in the presentation of an NVH simulation method for the electric powertrain system with satisfactory accuracy and efficiency, which makes it possible to identify and avoid whine problems at early stages of powertrain development.

2. Process of Vibration and Sound Analysis

When the electric vehicle is speeding up or decelerating, harmonic excitation forces with order characteristics in the electric drive system excite the powertrain housing to vibrate and radiate noise into the air. There are two main types of harmonic excitation forces responsible for whine noise, i.e., electromagnetic excitation loads and gear-meshing forces. The electromagnetic loads mainly consist of two parts, including the electromagnetic forces on the stator teeth and the torque ripple on the rotor [8–10]. The former acts on the stator structure directly, while the latter acts on the rotor shaft which transmits the pulsating harmonic load to the powertrain housing through the bearings of the drive system [11]. Gear-meshing forces are the dynamic loads produced by the interaction between meshing gears, which can also be transmitted to the powertrain housing through the bearings. In this paper, the time-domain electromagnetic loads under a couple of constant-speed conditions are obtained by 2-dimensional (2D) electromagnetic field simulation in the software Maxwell. The time-domain forces on the bearings under the constant-speed conditions are obtained through multi-body dynamic simulation. Then, the order loads are generated by the interpolation algorithm after the forces of the constant-speed conditions have been obtained. The normal vibration velocity of the powertrain housing is obtained by using FE analysis by applying the order forces onto the structural model. The acoustic FE module in the commercial software Virtual Lab is used to calculate the acoustic transfer vector (ATV) from the surface vibration velocity of the powertrain housing to sound pressure at the acoustic observing positions. At last, the sound pressure levels (SPLs) of the observing positions can be calculated by the surface vibration velocity of the powertrain housing and the ATV.

NVH simulation error of the electric powertrain system is affected by a number of factors, among which the modeling accuracy of the structural modes is a crucial one. On the one hand, the modeling accuracy of the structural modes determines the accuracy of the multi-body dynamic simulation, which means it will affect the computed results of the bearing forces. On the other hand, the simulated results of forced vibration are significantly affected by the accuracy of structural modes. The key challenge in modeling the powertrain structure lies in the treatment of the stator. The stator core is composed of compacted silicon steel sheets, which shows material anisotropy and parameter uncertainty. These characteristics make it quite difficult to model the core accurately. In addition, the stator windings have similar characteristics. In order to model the stator as accurate as possible, modal correlation analysis for tested modes and simulated modes [12–14] is performed to calibrate the stator parameters before the FE model is used for multi-body dynamic simulation and vibration analysis. The complete process of the NVH simulation analysis is illustrated in Figure 1.

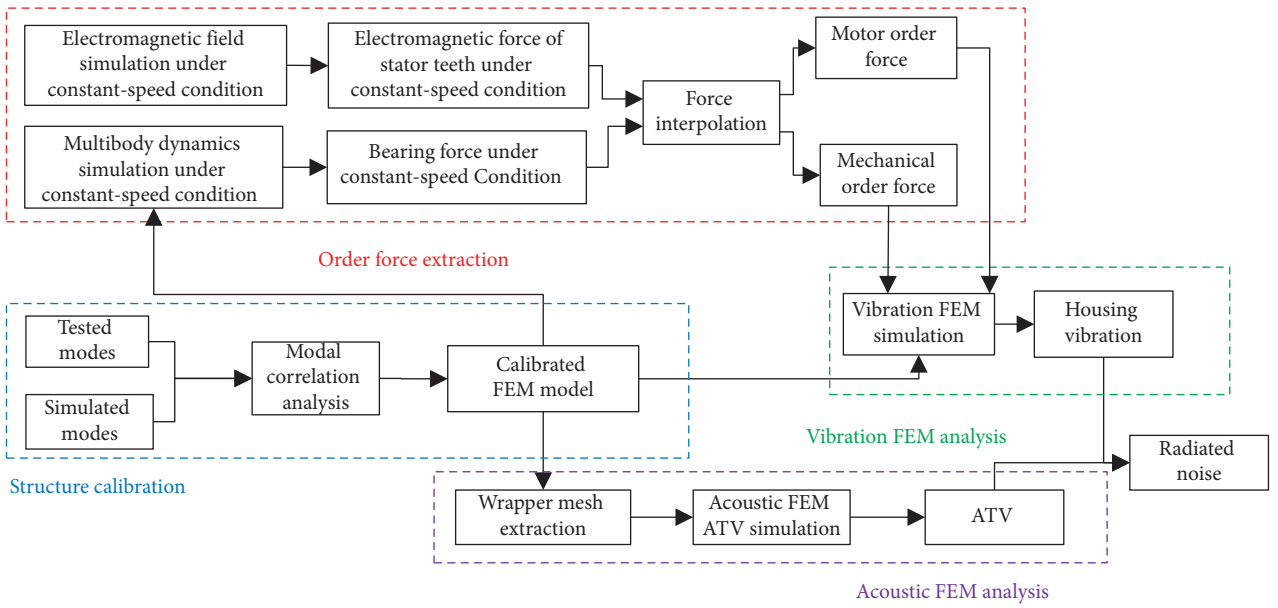


Figure 1: Process of vibration and sound analysis.

3. FE Modeling and Calibration for the Powertrain System

In Figure 2, the 3D structural model of powertrain NVH simulation is illustrated. The powertrain system consists of a motor, a motor controller, and a reducer, and it connects to the vehicle body through the suspensions.

The difficulty of FE modeling lies in the construction of the stator model. Under the constant-speed condition, the air-gap electromagnetic force is a periodic function of time and circumferential angle. As a result, under the speed-varying condition, the air-gap force shows order characteristics in both the frequency domain and the wavenumber domain. The air-gap electromagnetic load can be considered as a superposition of a series of "force patterns" which can be regarded as a set of basis of the electromagnetic load and can be obtained by 2D Fourier transform (FT) to the air-gap electromagnetic load. Each force pattern has a specific spatial distribution and rotates circumferentially under a specific frequency. The magnitude of the electromagnetic noise is heavily dependent on the level of agreement between the "force patterns" and the stator modes. While the rotating frequency of a "force pattern" is close to a modal frequency and the shape of the "force pattern" matches well with the modal shape, the stator will experience strong resonance, radiating intense electromagnetic noise. In order to confirm the accuracy of simulation, an accurate FE model of stator structure is of great importance. Natural modes for the stator structure are tested in advance to conduct parameter calibration.

The calibration of the model is a process of optimizing model parameters, thus making the simulation model represent the actual dynamic characteristics of the structure and leading to good agreement between the simulation results and the test data regarding the modal shapes and frequencies. In general, there are two approaches in terms of

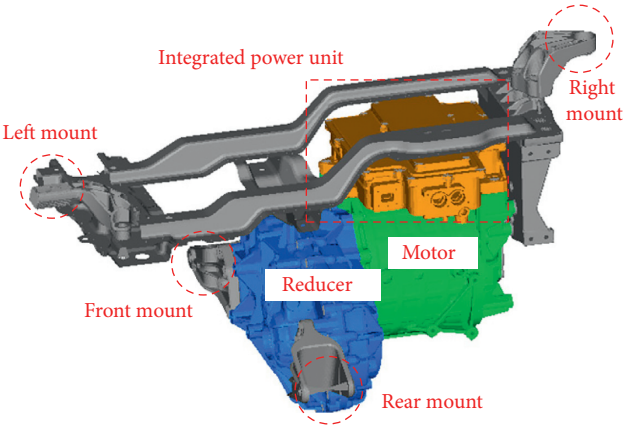


Figure 2: Structural model of the electric powertrain system.

calibration, namely, manual adjustment and modal correlation analysis. Manual adjustment needs to tune each physical parameter in order to achieve good agreement. Hence, this method heavily relies on personal experience and is difficult to implement for a complicated model. Modal correlation analysis can be carried out by using commercial software, for example, the correlation module in Virtual Lab, in which parameters can be optimized automatically after importing the tested modes into the software. In this paper, the latter method has been adopted.

The FE model of the stator is shown in Figure 3. The modeling method of the winding is revealed in the zoom-in image of Figure 3. The winding is considered to be made up of two parts: the equivalent isolation layer and the equivalent winding. The former part is isotropic material with small elasticity modulus, and the latter is anisotropic material. The initial material parameters of the stator core and the equivalent winding are set according to [12]. The material for the isolation layer is polyimide with the

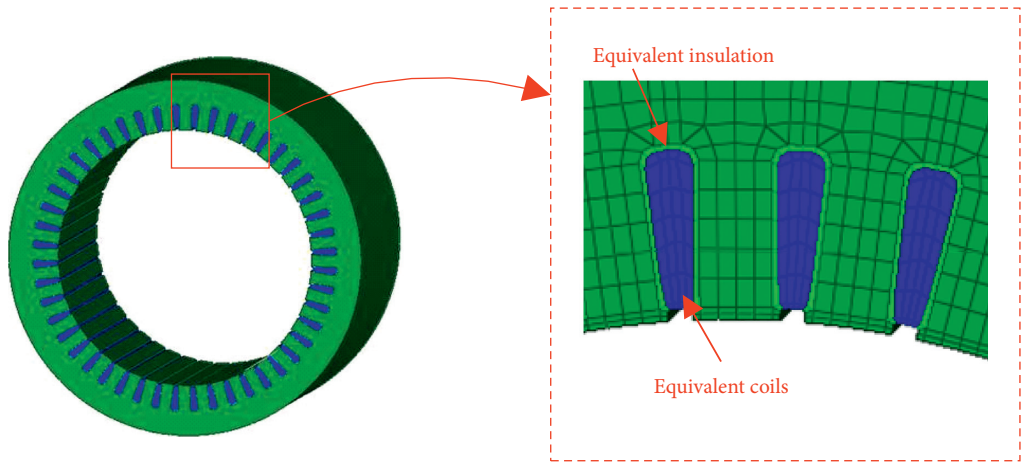


Figure 3: FE model of the stator core with winding.

default setting to be shown as follows: density $\rho = 1.2$ g/ml, elasticity modulus $E = 3$ GPa, and Poisson ratio $\mu = 0.35$.

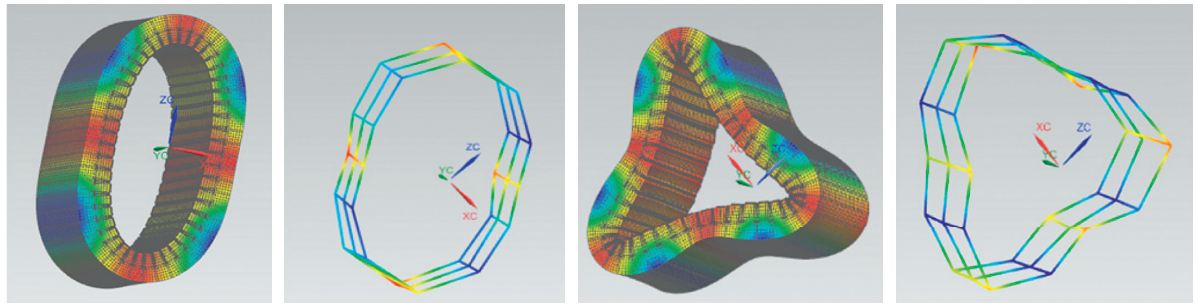
In the modal test, the frequency response functions are obtained under transient excitations. The stator is hung with an elastic slope, and a hammer is used to excite structural vibration. In total, 36 vibration-measuring points are distributed in the matrix along the stator surface, 3 circles with 12 test points equally spaced within one circle. Three axial acceleration sensors are used to acquire vibration signals. After the modal test, the results are fed to the correlation module of Virtual Lab to carry out correlation analysis and parameter optimization. In Figure 4, the comparison between simulated and tested modes is provided, showing the modal results with axial order $m = 0$ and circumferential order n being 0, 2, 3, and 4, respectively. The discrepancy stands for the relative error between the computational modal frequency and experimental modal frequency. MAC=1 means that the two modes are identical, while MAC=0 denotes that the mode shapes are orthogonal. As can be seen in Figure 4, after parameter calibration, the FE model can represent the natural vibration characteristics accurately. The relative errors for the 4 modes are less than 6.4%, with the correlation coefficient higher than 0.6. Apart from the stator, the material for other structural parts can be considered as isotropic, which will not be discussed in detail here.

4. Simulation of Excitation Forces

The NVH performance of the powertrain system depends on the running conditions. In most conditions, the vehicle velocity and motor torque are often transient under real-world driving. In general, the whole throttle acceleration and coasting deceleration are the two worst conditions in terms of powertrain whine noise for electric vehicles. Therefore, these two conditions are normally selected as the key conditions to be evaluated for powertrain NVH performance. Taking the whole throttle condition as an example, simulation of the order of vibration and noise has been carried out for the powertrain system. In order to obtain the order forces, the electromagnetic forces and bearing forces under constant-speed conditions are obtained at first, and then the order forces under the speed-varying condition can be calculated with interpolation.

4.1. Electromagnetic Forces under Constant-Speed Conditions. The powertrain system is equipped with a permanent magnet synchronous motor with 8 poles and 48 slots. FE analysis for the electromagnetic field in the software Maxwell has been performed to constant-speed conditions to obtain the electromagnetic forces. The simulation of the electromagnetic field is conducted from 1000 rpm to the maximum speed, with a step of 1000 rpm. According to the previous research work [15], satisfactory simulation accuracy can be achieved when applying interpolation to electromagnetic force calculation with this step.

Instead of using the electromagnetic forces at the nodes of each tooth for subsequent interpolation and simulation directly, the equivalent concentrated electromagnetic forces for each tooth are adopted. The node forces on the tooth are equalized to a concentrated axial force and a concentrated tangential force, with the distribution effect of the electromagnetic loading in circumferential direction neglected. This approach brings in little deterioration in simulation accuracy but significantly reduces the amount of data to be processed, hence improving simulation efficiency. The 48 slots of the motor are evenly distributed, and the electromagnetic force is sampled in circumference at 48 points. According to the sampling theorem, when carrying out FFT, only the forces for the first 24 spatial orders can be recognized. When the concentrated forces are adopted for NVH calculation, the contribution of forces with spatial orders higher than 24 ($n > 24$) is ignored. Generally, the electromagnetic vibration of the motor is heavily dependent on the circumferential structural modes with low orders. As the order increases, the vibration amplitude of the mode decreases with a speed of n^4 [16]. In addition, as the force order increases by a multiple of the number of poles, the amplitude of the electromagnetic force shows a decreasing trend. Hence, the application of concentrated electromagnetic force has little influence on simulation accuracy.

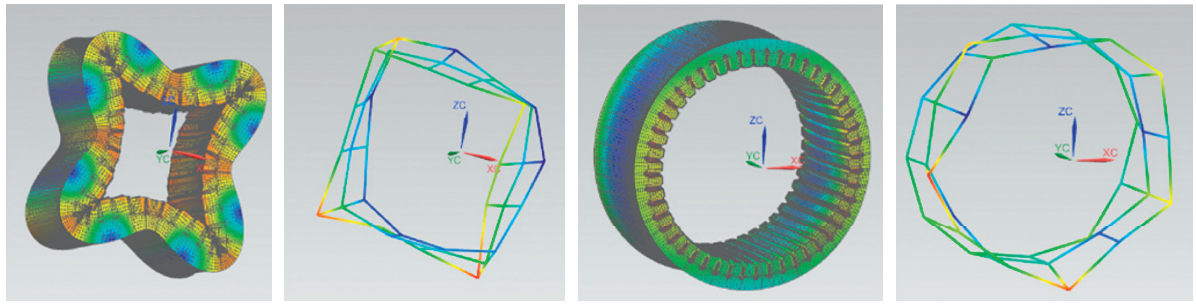


Simulation: 689.7 Hz
Error: 4.2%

Test: 662 Hz
MAC: 0.81

Simulation: 1837.1 Hz
Error: 2%

Test: 1801 Hz
MAC: 0.87



Simulation: 3329.5 Hz
Error: –6.4%

Test: 3557 Hz
MAC: 0.60

Simulation: 5491.8 Hz
Error: –3.1%

Test: 5667 Hz
MAC: 0.83

Figure 4: Modal results of the stator core with winding.

4.2. Bearing Forces under Constant-Speed Conditions. The reducer is a one-ratio two-stage gear transmission system, and there are two pairs of helical gears in the gearbox for speed slowdown and torque increasing. In addition, there is a pair of differential gears. Dynamic forces generate during gear-meshing process, which is transmitted to the axles first and then to the powertrain housing via bearings. In this paper, multi-body dynamic simulation is performed to obtain the exciting forces at the bearings under constant-speed conditions.

According to research work [17, 18], the flexibility of the housing has influence on dynamic meshing force characteristics. Hence, the powertrain housing is considered as a flexible body in the multi-body dynamic model. Before feeding the FE model which has been calibrated in Section 3 to the multi-body dynamic model, modal condensation is used to reduce the degrees of freedom of the housing model, hence improving efficiency. Provided with the highest frequency of the noise of interest $f_{\max}$, only the housing modes with natural frequency below $2f_{\max}$ are retained. In Section 4.3, the speed conditions selected for multibody dynamic simulation are introduced.

4.3. Order Forces under the Varying Speed Condition. In this section, the order forces under the acceleration condition can be obtained by applying interpolation. The notion "order" refers to how many times the frequency is referenced to the rotating frequency of the rotor. FFT is performed to obtain the frequency spectrums of the forces. In the end, force interpolation in the frequency domain between adjacent constant-speed conditions is performed . Cubic spline interpolation is used here, with the boundary condition at each endpoint being a not-a-knot boundary (the 3rd derivative at the endpoint equals that of the adjacent point).

The constant-speed conditions used for interpolation include speeds of two parts, with one part being the same as that used in electromagnetic force simulation and the other being additional speed conditions. The latter is included to take into account the influence of the natural vibration characteristics of the powertrain housing on the order bearing forces. Under each additional speed condition, the order frequency of the gear engagement matches well with one of the natural frequencies of the powertrain housing. Figure 5 illustrates how to select additional speed conditions. f_i represents the modal frequency of the powertrain housing. The order lines intersect with the resonance frequency lines, and the speeds at the intersection points are selected as the additional speed conditions. For example, speeds N_i^1 and N_i^2 are included in force interpolation to take into account the coupling effects of the housing mode f_i. If the number of the housing modes within the frequency range of interest is I and the number of force orders of interest is q, the number of additional speed conditions should be I^*q.

5. Vibration and Noise Simulation

5.1. Vibration Modeling and Loading. Before vibration simulation, the natural modes of the powertrain system should be computed by FE analysis. In this paper, the commercial software Nastran is used for the modal calculation, and then the modal results are imported into the software LMS Virtual Lab for force loading and vibration

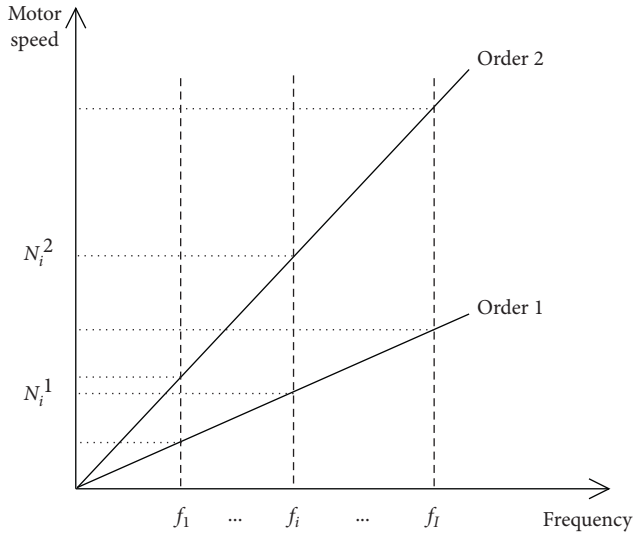


Figure 5: Schematic diagram of the additional steady-speed operation condition determination method.

simulation. Spring elements are used to model the suspension cushions, with the stiffness coefficients of each element set as the static stiffness coefficients of the suspension cushions.

A force treatment program "Force_gene.exe" is worked out to automatically generate order forces, match them with the FE geometry, and output a load file in the format of ".unv" which then can be imported into LMS Virtual Lab for vibration simulation. The program is much more efficient and error-less than loading manually , as the interpolation and loading of the order forces of all 48 stator teeth and 6 bearing holes can be completed by just running the program in seconds. It needs to be noted that the load on each stator tooth obtained by interpolation calculation is still a concentrated force. The code "Force_gene.exe" uniformly decomposed the concentrated force into dozens of point forces distributed on the tooth surface .

5.2. Acoustic Simulation. The boundary element method (BEM) [19] is a native method for simulation acoustic wave problems, especially for exterior acoustic problems. To overcome drawbacks of the conventional BEM, such as efficiency and large memory consumption, fast accelerated BEM has been proposed and applied for large-scale acoustic problems [20, 21]. In this work, acoustic transfer vector (ATV) from the vibration of the powertrain housing to sound pressure response is calculated by the acoustic FE simulation in the software LMS Virtual Lab. ATV can be regarded as the linear input-output transfer relation between the housing vibration and the response point of the sound field, which can be expressed by the following equation:

$$p(\omega) = \langle \mathrm{ATV}(\omega) \rangle \{V_n(\omega)\}. \quad (1)$$

$\{V_n(\omega)\}$ is the normal vibration velocity matrix of the powertrain housing, $p(\omega)$ denotes the sound pressure of the sound field response point, and $\langle \mathrm{ATV}(\omega) \rangle$ represents the ATV matrix. The ATV matrix depends on the housing geometry, acoustic impedance at the structure-air interface, acoustic field response position, acoustic signal frequency, and acoustic medium parameters, but it is not related to the surface vibration velocity of the structure. Therefore, ATV can be calculated without the presence of vibration velocity. For the electric powertrain system, NVH simulation analysis is usually required for different operating conditions, such as full-throttle acceleration, half throttle acceleration, and coasting deceleration. For acoustic simulation under multiple operating conditions, the ATV method is more efficient than the direct vibroacoustic FE method. Instead of calculating the sound pressure directly for multiple rounds, the former only needs one round of acoustic FE simulation as the ATV keeps invariable for different operating conditions.

6. Results and Discussion

6.1. Bearing Forces. The electromagnetic loads of the motor include two parts, namely, the electromagnetic forces on the teeth and the torque ripple on the rotor. In addition to the tooth order forces discussed in Section 4.1, the torque ripple on the motor rotor also contains the order components in multiples of the number of poles, such as order 8 and order 48. The harmonic torques can be also transmitted to the powertrain housing through the bearings and cause vibration and noise. If one needs to take into account the contribution of the torque ripple on bearing forces, the time-domain signal of the rotor torque obtained by electromagnetic simulation should be adopted in the multi-body dynamic model in Section 4.2. Figure 6 illustrates the amplitude curves of the order forces at one bearing of the input shaft of the reducer, including the orders caused by the electromagnetic torque ripple, i.e., order 8 and order 48, and those caused by meshing gears. The curve O_{ij} in the figure denotes the j-th order force caused by the i-th pair of meshing gears. Each curve in the figure denotes the amplitude of the vector sum of the order force, which has been converted into A-weighted level in decibel, with the level F_L calculated with equation (2). $F(n)$ is the force amplitude corresponding to rotor speed n, and the reference value of the force $F_0 = 1N$:

$$F_L = 20 \log \left[\frac{F(n)}{F_0} \right]. \quad (2)$$

Figure 6 indicates that the bearing forces caused by the motor torque ripple (orders 8 and order 48) are more than 30 dB (A) smaller than the gear-meshing order forces, which means that the sound pressure levels (SPL) caused by the former should be much smaller than the latter. It should be noted that the SPLs caused by the electromagnetic forces on the stator teeth could often reach or even exceed those caused by meshing gears, implying that the electromagnetic noise caused by the torque ripple should be a negligible value compared to that caused by the electromagnetic forces on the stator teeth. In the following NVH simulation, constant torques are used for inputs in multi-body dynamic analysis with torque ripples neglected since the overlooking of the

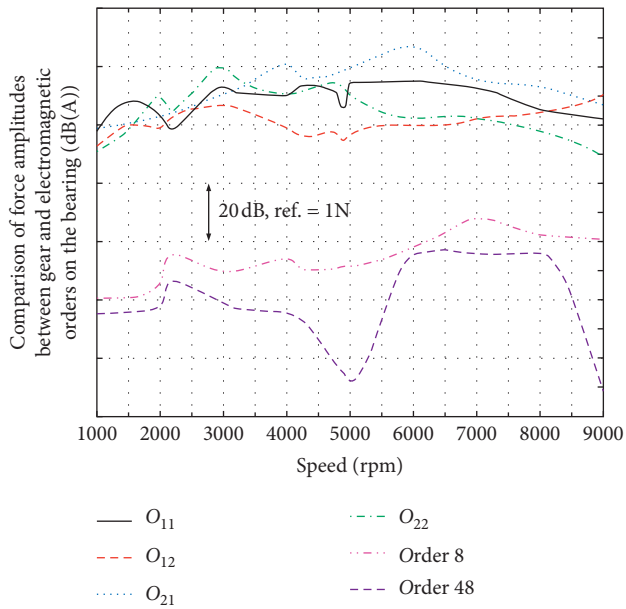


Figure 6: Comparison of force amplitudes between gear and electromagnetic orders on the bearing.

motor torque ripple does not result in significant loss in simulation accuracy of electromagnetic noise.

6.2. Experiment Verification and Analysis. The simulated vibration and noise results are compared with the NVH test data for verification. The test is conducted in a semi-anechoic laboratory, and the electric powertrain is installed on the test bench through the suspension system as shown in Figure 7. Four microphones are located 1 meter away from the powertrain housing in different directions, i.e., the front, back, right, and above, and a 3-axis acceleration sensor is positioned on the powertrain housing. The vibration and sound pressure signals are recorded under the full-throttle acceleration condition.

The vibration acceleration at the housing and the average SPLs of the 4 microphones are presented in Figure 8, with the vibration acceleration denoting the vector sum of the signals of the 3-axis acceleration sensor. As can be seen, the simulation curves and the experiment results show good consistency. The relative speed deviation of any local peak $|\delta_r|$ is below 8%, with $|\delta_r|$ calculated by formula (3), where N_{CAE} is the motor speed corresponding to the local peak value on the simulation curve and N_{test} is the motor speed corresponding to the local peak value on the experiment curve. The error of the average SPL is mostly below 10 dB (A), with the peak error of the 48th order around 7200 rpm being about 1 dB (A) and the peak error of the order O_{11} around 6800 rpm being about −8 dB (A):

$$\delta_r = \frac{N_{\text{CAE}} - N_{\text{test}}}{N_{\text{test}}}. \tag{3}$$

The maximum peaks on the 48-order vibration and noise curves appear near 7000 rpm, and the response

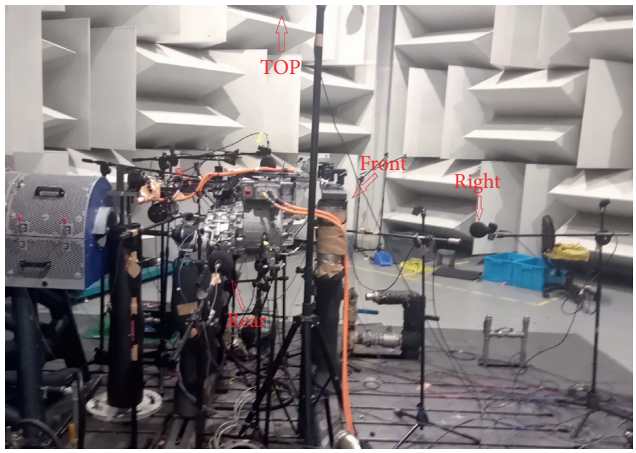


Figure 7: NVH test of the powertrain in a semianechoic chamber.

frequency is about 5600 Hz which is highly consistent with the resonant frequency of the "breathing" mode of the motor stator ($n = 0$, f 5600 Hz), as presented in Figure 4. This strong peak appears at around 5600 Hz due to the following two factors. On the one hand, the spatial zero-order "force pattern" ($n = 0$, f 5600 Hz) of the 48th order electromagnetic force matches well with the stator "breathing" mode in both shape and frequency, leading to strong resonance; on the other hand, for the "breathing" mode, the normal velocity of the stator is in the same phase, which means it has strong acoustic radiation efficiency. The results indicate that accurate simulation of the stator modes is very important for the calculation of motor-related noise, reinforcing the significance of the modal calibration in Section 3.

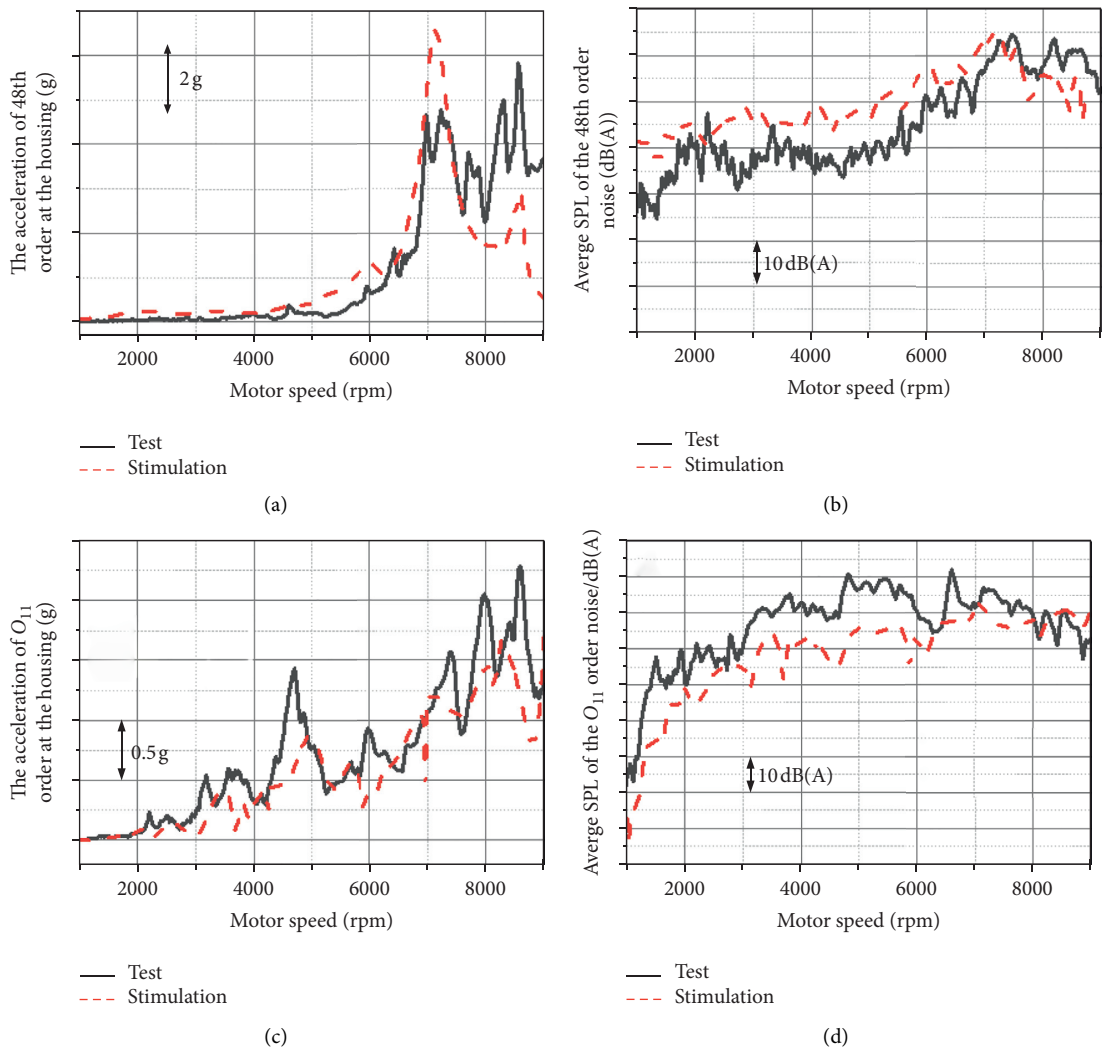


FIGURE 8: Comparison between simulation and test: (a) the acceleration of the 48th order at the housing, (b) average SPL of the 48th order noise, (c) the acceleration of the O_{11} order at the housing, and (d) average SPL of the O_{11} order.

7. Conclusions

This paper presents a method of NVH simulation analysis for the electric powertrain system under speed-varying conditions. Modal correlation analysis is performed to calibrate the natural modes of the motor stator and improve NVH simulation accuracy.

The calibrated simulated modes are in good agreement with the experimental modes. For the four modes of interest, the frequency errors are within 6.4%, and the MACs are not less than 0.6.

The computed bearing forces show that electromagnetic noise caused by the torque ripple would be a negligible value compared to that caused by the electromagnetic forces on the stator teeth. When multi-body dynamic simulation is used to calculate the bearing forces, ignoring the torque ripple of the motor rotor does not lead to significant loss in electromagnetic SPLs.

The vibration and sound results obtained by simulation and the test are in good agreement. The relative frequency deviation of local peaks between simulation and test curves is less than 8%. The peak error of the motor of 48-order SPL is about 1 dB (A) and that of the order O_{11} is about −8 dB (A).

The 48-order whine noise is strongly related to the breathing mode of the stator. When the circumferential 0-order component of the 48th-order electromagnetic force coincides with the stator breathing mode in space and frequency, the vibration and sound curves show strong local resonance peaks.

The influence of model parameters on simulation accuracy should be an interesting research topic in the future.

Data Availability

The data are not freely available due to the requirement of commercial confidentiality.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] P. Pellerey, V. Lanfranchi, and G. Friedrich, "Coupled numerical simulation between electromagnetic and structural models. Influence of the supply harmonics for synchronous machine vibrations," *IEEE Transactions on Magnetics*, vol. 48, no. 2, pp. 983–986, 2012.

[2] Q. Kang, P. Gu, C. Gong, and S. Zuo, *Test and Analysis of Electromagnetic Noise of an Electric Motor in a Pure Electric Car*, SAE Technical Paper, Grand Rapids, MI, USA, 2019.

[3] L. Humbert, P. Pellerey, and S. Cristaudo, "Electromagnetic and structural coupled simulation to investigate NVH behavior of an electrical automotive powertrain," *SAE International Journal of Alternative Powertrains*, vol. 1, no. 2, pp. 395–404, 2012.

[4] Y. Fang and T. Zhang, "Vibroacoustic characterization of a permanent magnet synchronous motor powertrain for electric vehicles," *IEEE Transactions on Energy Conversion*, vol. 33, no. 1, pp. 272–280, 2017.

[5] O. Harris, P. Langlois, and A. Gale, *Electric Vehicle Whine Noise—Gear Blank Tuning as an Optimization Option*, pp. 64–73, Gear Technology, Chicago, IL, USA, 2019.

[6] P. Yu, S.-Y. Chen, T. Zhang, and R. Guo, "Vibration response of an EV power train under mechanical-electromagnetic excitation," *Vibration and Shock*, vol. 35, no. 13, pp. 99–105, 2016.

[7] P. Yu, F.-F. Chen, T. Zhang, and R. Guo, "Vibration characteristics analysis of a central-driven electric vehicle powertrain," *Vibration and Shock*, vol. 34, no. 1, pp. 44–48, 2015.

[8] P. Vijayraghavan and R. Krishnan, "Noise in electric machines: a review," *IEEE Transactions on Industry Applications*, vol. 35, no. 5, pp. 1007–1013, 1999.

[9] J. F. Gieras, C. Wang, and J. C. Lai, *Noise of Polyphase Electric Motors*, CRC Press, Boca Raton, FL, USA, 2018.

[10] N. Chandrasekhar, C. Tang, N. Limsuvan et al., *Current Harmonics, Torque Ripple and Whine Noise of Electric Machine in Electrified Vehicle Applications*, SAE Technical Paper, Detroit, MI, USA, 2017.

[11] T. C. Lim and R. Singh, *A Review of Gear Housing Dynamics and Acoustics Literature*, NTRS Report NAS 1.26:183110, Ohio State University, Columbus, OH, USA, 1988.

[12] W. Deng, S. Zuo, H. Sun, S. Wu, and G. Zhang, "Modal analysis of a claw-pole alternator considering orthotropy of the stator core and windings," *Vibration and Shock*, vol. 36, no. 12, pp. 43–49, 2017.

[13] S. Zuo, Y. Zhang, J. Yan, G. Zhang, F. Lin, and S. Wu, "Optimization of vibration and noise in permanent magnet synchronous motor considering stator anisotropy," *Journal of Xi'an Jiaotong University*, vol. 51, no. 5, pp. 60–68, 2017.

[14] P. Millithaler, É. Sadoulet-Reboul, M. Ouisse, J.-B. Dupont, and N. Bouhaddi, "Structural dynamics of electric machine stators: modelling guidelines and identification of three-dimensional equivalent material properties for multi-layered orthotropic laminates," *Journal of Sound and Vibration*, vol. 348, pp. 185–205, 2015.

[15] Q. Deng, "A high efficiency NVH simulation methodology for vehicle driving motor based on electromagnetic force approximation," *Journal of Applied Acoustics*, vol. 38, no. 6, pp. 932–938, 2019.

[16] Y. Chen, Z. Zhu, and S. Ying, *Analysis and Control of Electric Motor Noise*, Zhejiang University Press, Hangzhou, China, 1987.

[17] G. Zheng, X. Huang, and D. Guo, "Effects of flexible gearbox body on dynamic meshing performance of its gear pair," *Journal of Vibration and Shock*, vol. 36, no. 13, pp. 140–145, 2017.

[18] P. K. Singh, *Study of Effect of Variation in Micro-Geometry of Gear Pair on Noise Level at Transmission*, SAE Technical Paper, Pune, India, 2015.

[19] L. G. Copley, "Fundamental results concerning integral representations in acoustic radiation," *The Journal of the Acoustical Society of America*, vol. 44, no. 1, pp. 28–32, 1968.

[20] H. Wu, Y. Liu, and W. Jiang, "A fast multipole boundary element method for 3D multi-domain acoustic scattering problems based on the Burton-Miller formulation," *Engineering Analysis with Boundary Elements*, vol. 36, no. 5, pp. 779–788, 2012.

[21] H. Wu, Y. Liu, and W. Jiang, "A low-frequency fast multipole boundary element method based on analytical integration of the hypersingular integral for 3D acoustic problems," *Engineering Analysis with Boundary Elements*, vol. 37, no. 2, pp. 309–318, 2013.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9406479, 10 pages
https://doi.org/10.1155/2020/9406479

Research Article

A Novel Median-Point Mode Decomposition Algorithm for Motor Rolling Bearing Fault Recognition

Ganzhou Yao, Bishuang Fan, Wen Wang, and Haihang Ma

School of Electrical and Information Engineering, Changsha University of Science & Technology, Changsha, China

Correspondence should be addressed to Bishuang Fan; fbs@csust.edu.cn

Received 31 July 2020; Accepted 13 September 2020; Published 16 November 2020

Academic Editor: Yong Chen

Precise fault recognition of motor rolling bearing fault is playing a significant role in any machinery and equipment. However, conventional decomposition methods fail to completely reveal the fault signal information of motor rolling bearing due to mixed modes problem. To solve the problem, the median-point mode decomposition (MMD) method is presented. The MMD method uses sort-based inversion to sort out each variation of the same time interval for better and specific mode decomposition, with the assistance of the advanced envelope curve formed by the median points between adjacent extreme points. It certainly alleviates the mixed mode during the iteration of intrinsic mode functions (IMFs). Therefore, comparison results are simulated in the proposed MMD method with conventional methods. Experiment of motor rolling bearing fault is operated for fault recognition in order to demonstrate the MMD algorithm.

1. Introduction

Rolling bearings are common components in rotating machines, which have been significant in the industry. The motor signal is a nonlinear, nonstationary weak signal with strong randomness. In the acquisition process, it will be affected by external environmental actions or noise interference such as power frequency, leading to mixed modes in the IMF components. Therefore, the preprocessing of this type of signal is an important research problem. Meanwhile, fault signal of the motor cannot be intuitively observed due to its characteristic complexity, so it needs to be decomposed or extracted in time domain and frequency domain and fault characteristic values from multiple angles should be obtained. Feature extraction is the core content of fault recognition. The accuracy of the signal process and that of feature extraction will directly affect the reliability of fault recognition. Thus, HHT is an adaptive time-frequency analysis method to be used in the feature extraction of fault recognition.

Conventional signal processing techniques can only detect stationary and linear signals [1]. Wavelet transform was studied for nonstationary signals and time-sfrequency analysis [2], but the wavelet base function limits the result of it, which may lead to a priori assumption on the characteristics of the investigated vibration signal [3]. As a self-adaptive signal processing method, empirical mode decomposition (EMD) is analyzed to decompose the complicated signal into a set of complete and intrinsic mode functions (IMFs) [4, 5].

However, mixed mode problem is one of the major drawbacks of EMD, caused by the screening process in the EMD algorithm and the discontinuity of the eigenmode function of a certain time scale and several time scales [6]. Mixed mode problem leads to the decomposed IMFs becoming distorted because the signals are mixed with discontinuous high-frequency weak noise interference and it confuses the time-frequency distribution, making each IMF lack physical meaning.

A simple mixed mode example would be like two identical signals, one having low-order random noise and the other not; the results of EMD decomposition can be quite different [7–10]. Mixed modes in bearing faults cause the fatal breakdown of machines and inestimable economic losses [11–14]. In order to overcome the above problems, ensemble empirical mode decomposition (EEMD) is studied

as a new solution for mixed mode problem, which is through adding finite white noise to the investigated signal. However, the Gaussian white noise may make it difficult to determine an ensemble mean as the different iterations can generate different number of IMFs [15–18]. Furthermore, the EEMD method is hard to be self-adaptive as it requires an amplitude of noise and ensemble number as parameters. Therefore, it is significant to detect the existence and severity of a bearing fault with an efficiently fast, accurate method.

In this paper, a novel median-point algorithm with time interval sort-based inversion is developed. EMD and EEMD algorithms with some of their drawbacks are reviewed. The rest of the paper is organized as follows: In Section 2, the principle of the proposed median-point mode decomposition is presented. Then, detail process simulations of MMD are shown in Section 3, followed by the flowchart of the MMD method. Finally, simulations of EMD and EEMD based on the same original mode as MMD and simulated fault recognition are all given to demonstrate that the proposed method based on MMD obtains a more precise mode decomposition result. The proposed MMD method can be applied in practice, particularly in fault recognition of rolling element bearings since its occurrence.

2. Principle of the Proposed Median-Point Mode Decomposition (MMD)

Median-point mode decomposition (MMD) can be treated as a screening process, which is a self-adaptive method and can decompose any complex signal into a list of intrinsic mode functions (IMFs), which must meet two conditions as follows in Table 1.

All the local extrema are identified as $x(t)$. In EMD, the first step is to construct the upper envelope and lower envelope in the signal by interpolating the local maxima and minima, respectively, using cubic spline [11]. However, in MMD, we apply sort-based inversion to detect out all periods in the same frequency and then, respectively, employ only the median point between adjacent extreme points of one specific part, to gain the median-point-fit-curve $m(t)$ for further managements.

Huge difference among EMD, EEMD, and MMD is that the sort-based inversion algorithm is adopted in MMD to sort the obtained time intervals from small to large. Set a default maximum value rate of time intervals earlier. Then, when the rate of change exceeds the set value, the system defaults to take the time interval value before the change as the maximum time interval value $T_{\max}$ of this required mode.

The median-point-fit-curve is formed by cubic spline function, under two different conditions, listed in Table 2.

Thus, the difference between the local extrema of $x(t)$ and median-point-fit-curve m(t) is marked as equation (1), which should meet the condition in Table 1:

$$h(t) = x(t) - m(t). \tag{1}$$

Repeat the above steps until $h(t)$ is an IMF, and then, set $c_i(t) = h(t)$. Then, compute the residue $r_i(t) = x(t) - c_i(t)$ and set $x(t) = r_i(t)$ and repeat the above steps to extract the next IMF until $r_i(t)$ is monotonic or constant.

The result of MMD algorithm can be expressed as

$$x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t), \tag{2}$$

where $x(t)$ is decomposed into a series of IMFs $c_j(t)$ and a residue $r(t)$. For better presentation of the principle of MMD, we have listed the steps of MMD, as shown in Table 3.

3. Detail Process of MMD

The original signal composed of signals with different amplitude and frequency ratios is crucial to the EMD\EEMD mode mixing problems. As the principle of MMD is presented completely in Section 2, an example is presented as follows, where $x(t)$ is composed of x_1, x_2, x_3, x_4, and x_5:

$$x_1(t) = 0.01t, \tag{3}$$

$$x_2(t) = 0.1\sin(2\pi t), \tag{4}$$

$$x_3(t) = 0.12\sin(6\pi t), \tag{5}$$

$$x_4(t) = 0.15\sin(16\pi t_2), \tag{6}$$

$$x_5(t) = 0.35\sin(76\pi t_1), \tag{7}$$

where $0 \le t \le 2$, $0.3 \le t_2 \le 0.6$, and $1.3 \le t_1 \le 1.6$, shown in Figure 1.

The original signal $x(t)$ consists of constituent signals with different degrees of frequency separation, which is shown in Figure 1. Mixed mode exists in nonlinear and nonstationary signals. In order to verify the sensitivity of MMD to signal changes, the proposed method adds new interference processing in the time $t_1 = 0.3$ s and $t_2 = 1.3$ s and ends at the time of 0.6 s and 1.6 s, respectively. Each component in $x(t)$ contains only a simple vibration mode (single instantaneous frequency), and the signals of these components can completely represent the real physical information in the original signal.

For comparison, the simulation signal $x(t)$ is analyzed using the EMD and EEMD method and the decomposition results are displayed in Figures 2 and 3.

Notice that when EMD is operated on the original signal $x(t)$, the result is as shown in Figure 2. Mixed mode problem makes the scale of the first-order IMF1 (c_1) different, and the scale of IMF2 (c_2) is also affected by c_2, while c_3 and c_4 contain the same scale signal. It can be judged that there are obvious mixed modes existing, leading to the mode component becoming seriously distorted, as compared with the original signal. It is indistinct that the problem of mixed modes appears at IMF1-4 below, showing that the EMD method fails to provide the reasonable decomposition.

The components y_1, y_2, y_3, y_4, and y_5 in original signal $x(t)$ are defined in (equations (3)–(7)). From top to the bottom of Figure 2, each subfigure represents IMFs with ascending order and is produced by the EMD method.

Table 1: Conditions of IMF.

Condition 1	The number of signal extreme points is equal to zero point or the difference in them is within 1.
Condition 2	At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Table 2: Conditions of the maximum time interval value T_{max}.

Condition 1	When the interval of adjacent extreme points is larger than T_{max}, the value of median point would be the magnitude and amount of time of $x(t)$ between the current extreme points.
Condition 2	When the interval is less than T_{max}, the MMD assigns the value of median point from the current adjacent extreme points and the value of median point corresponding to the original signal, together to the value of median point.

Table 3: The MMD algorithm.

Step 1	Identify all the local extrema of $x(t)$.
Step 2	Obtain the local maxima and minima of $x(t)$.
Step 3	Gain all the time intervals between adjacent extreme points.
Step 4	Apply sort-based inversion algorithm for time intervals from high frequency to low.
Step 5	Determine the rate of change of the time interval and set a maximum of time interval T_{max}.
Step 6	Gain different values of median point in different periods of time intervals based on two conditions of T_{max}.
Step 7	In each sorted period of time intervals, through cubic spline function, form the median-point-fit-curve with all the gained median points $m(t)$.
Step 8	Set $h(t) = x(t) - m(t)$.
Step 9	Repeat the above steps until $h(t)$ is an IMF, check in Table 1, and then set $c_i(t) = h(t)$.
Step 10	Compute the residue $r_i(t) = x(t) - c_i(t)$.
Step 11	Set $x(t) = r_i(t)$ and repeat the above steps to extract the next IMF until $r_i(t)$ is monotonic or constant.

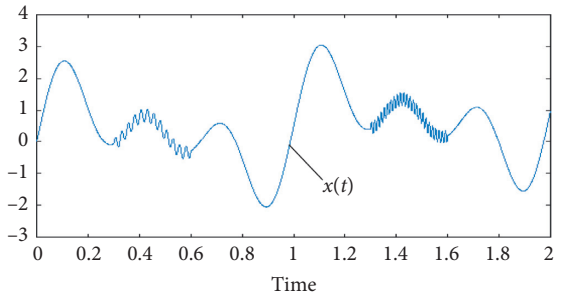


Figure 1: Synthetic signal waveform $x(t)$.

Despite the previous example [1] showing the EMD's accurate decomposition of a synthetic signal, the result above indicates that the mixed modes problem containing mixed components of the input signal cannot be decomposed successfully. Therefore, the same original signal $x(t)$ is taken as the input signal for the EEMD method for better comparison, shown as follows.

In Figure 3, it can be observed that when mixed modes occur, the signal components of different scales coexist in the same order of IMF. In other words, signal components with different frequencies coexist in the same order of IMF. From top to the bottom of Figure 3, each subfigure represents IMFs with ascending order produced by the EEMD method. As EEMD performing the signal $x(t)$, mixed modes can be reduced to a certain extent, but it cannot be eliminated fundamentally, and the decomposition result cannot reveal the signal characteristics and provide accurate information.

Note that MMD has multiresolution analysis and the advantages of signal analysis such as local adaptability, shown in Figure 4, where IMF1 is decomposed without the influence of mixed mode problem.

The process of the method for decomposing signals into each IMF is shown in Figure 4, demonstrating the advantage of self-adaptiveness and high efficiency in MMD.

The red curve in Figure 4 is the median-point-fit-curve $m(t)$, and the blue curve is the original signal $x(t)$; the difference between $x(t)$ and $m(t)$ can be obtained as an IMF if conditions meet equally (Table 1), denoted as $h(t)$. Even with a complex original signal in Figure 1, it can be noticed that imf1 $h(t) = x(t) - m(t)$ without obvious mixed mode.

Hence, the IMFs $h(t)$ equals the difference between the original signal $x(t)$ and the median-point-fit-curve $m(t)$. The MMD algorithm is operated in all five different composition processes in five different time intervals sorted by the

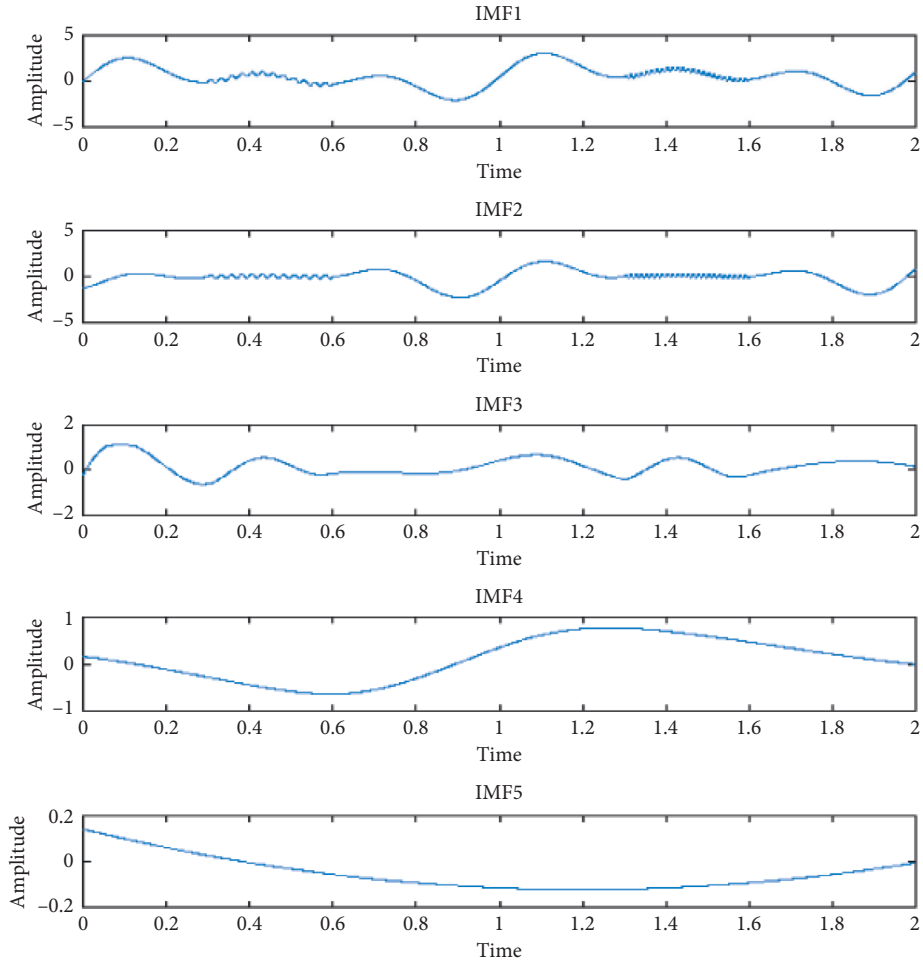


Figure 2: EMD result of signal $x(t)$.

ranking algorithm, in order to obtain IMF1-5, shown in Figure 5.

The result in Figure 5 demonstrates that the MMD method can effectively decompose the added interferences and normal signal into the correct constituent signals in various cases, alleviating mixed modes problem and being self-adaptive at the same time.

It can be seen from the results of IMFs in Figure 5 that MMD algorithm can decompose a series of IMFs from high to low frequencies, without the influence of mixed mode problem.

As the problem of mixed mode occurs, an IMF can cease to have physical meaning by itself, suggesting falsely that there may be different physical processes represented in a mode. In MMD, when acquiring IMF components, because too many iterations would damage the integrity of the signal and its physical meaning, the number of iterations needs to be limited. Therefore, the criterion to end iterations used in this method is already written in Tables 1 and 2 and Step 11 of Table 3.

Additionally, observing the differences of EMD and EEMD shown in Figures 2 and 3, the decomposition result of the EEMD method is better than that of the EMD method. However, EEMD takes three more steps to iterate out the final IMF component. Thus, the result of MMD using the same original signal $x(t)$ given in Figure 5 represents better mode decomposition.

Applying MMD to decompose $x(t)$ resulted in a series of IMFs, where the *imf*1-5 denote all the IMFs, showing a successful decomposition of four smoothly sinusoidal signals and single residual, accordingly. As can be seen, MMD can solve the problem of mixed modes well with the mode component very similar to the original signal. Comparing Figures 3–5, the IMFs decomposed by MMD is obviously more accurate than the decomposition results of EMD and EEMD. The frequency of each IMF is sequentially reduced, and the waveform transformation is more regular. It shows that MMD can avoid mixed mode because it could separate high-frequency and low-frequency components clearly and obtain the meaningful signal sufficiently. It can also prove that MMD maintains the adaptability in signal decomposition.

In the interim, the implementation flowchart of this proposed MMD method is shown in Figure 6. The $x(t)$ represents original signal in Figure 1. $c(t)$ stands for each of IMFs $h(t)$, and $r(t)$ denotes residue, which equals to $x(t) - c(t)$. The median-point-fit-curve m(t) is formed by cubic spline function. At first, identify all the local extrema of original signal $x(t)$ to obtain the local maxima and minima of

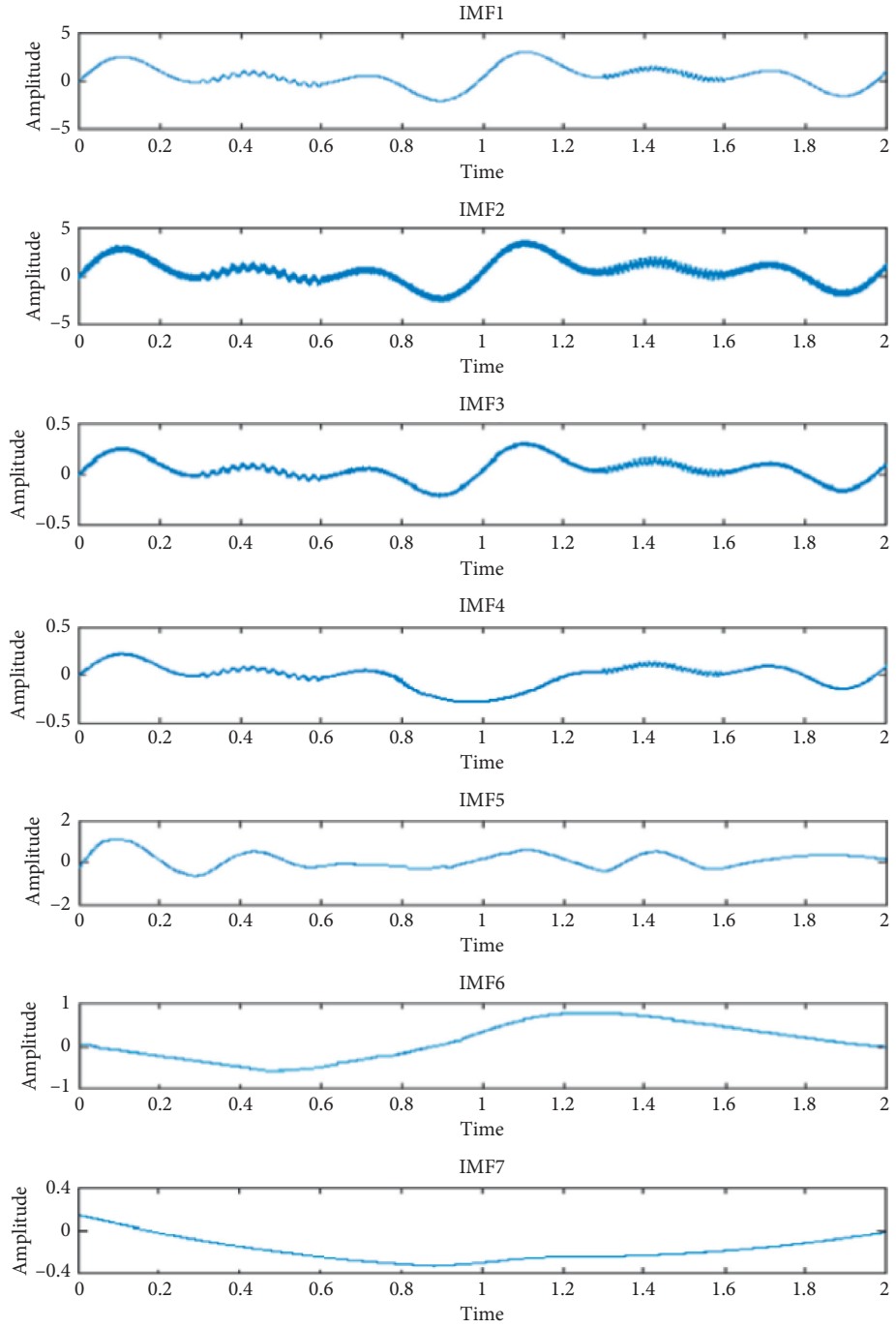


Figure 3: EEMD result of signal $x(t)$.

$x(t)$. Then, the time interval in all the adjacent extreme points is arranged in ascending order with sort-based inversion, selecting out the different frequency periods for MMD to operate, respectively. Check two conditions about the pre-set maximum of time interval Tmax of Table 2, then the cubic spline function is used to form the median-point-fit curve in each sorted time interval, and the median-point-fit curve obtained is processed in the next step according to the EEMD and EMD methods. Finally, the MMD algorithm achieves self-adaptive mode decomposition with the alleviation in mixed modes.

4. Motor Fault Recognition Experiments

The characteristic complexity in motor fault signal makes it hard to be detected. Generally, engineers and researchers adopt different diagnostic methods for different bearing faults of motors, but each one needs the separation from the decomposition and extraction of modes.

When a bearing fault occurs in an asynchronous motor, its vibration frequency will change significantly, and for different types of bearing faults, the characteristic frequency of the fault produced is also different.

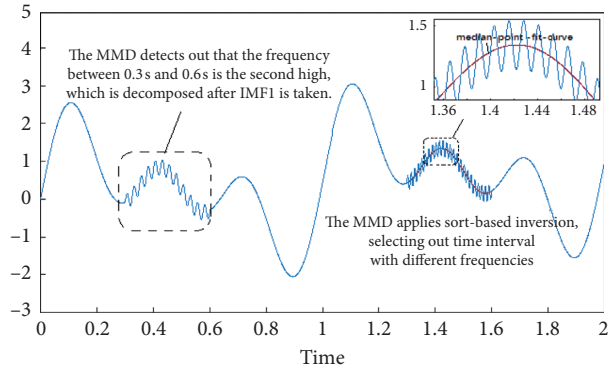


Figure 4: Synthetic signal waveform $x(t)$ for obtaining IMF1.

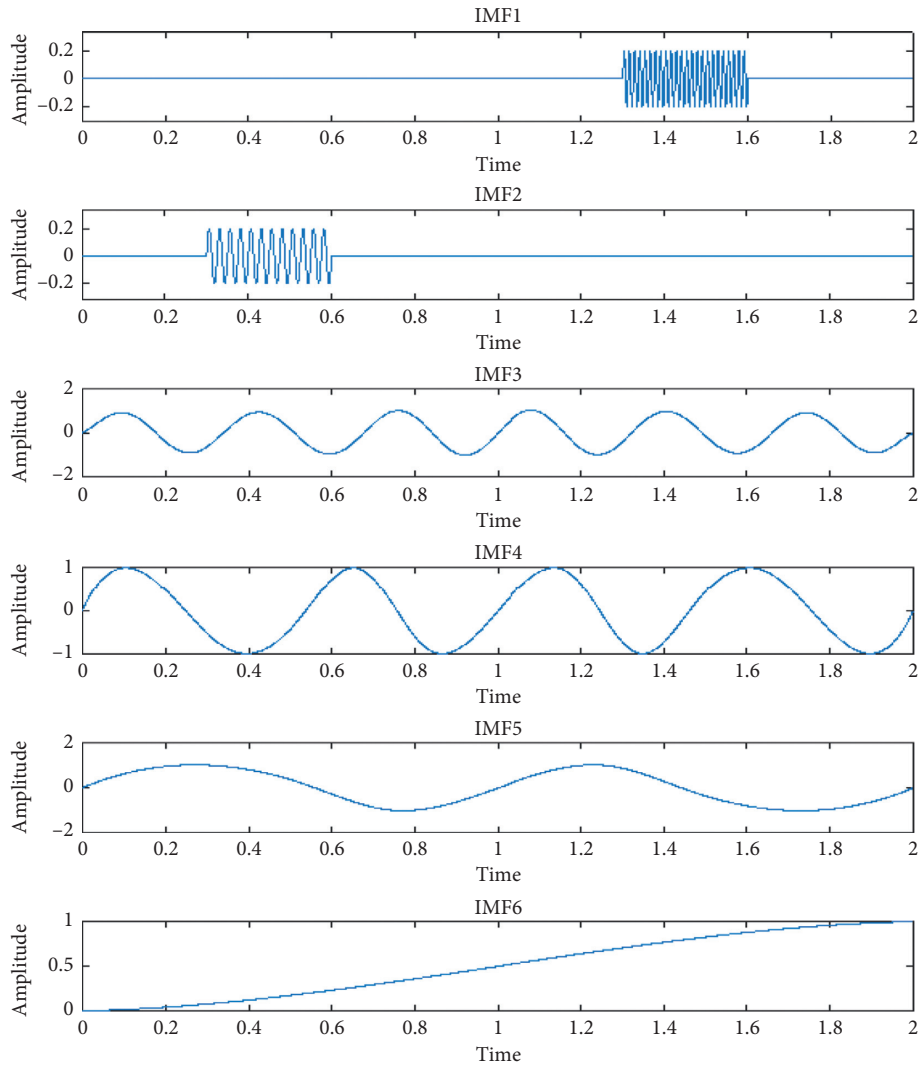


Figure 5: MMD result of signal $x(t)$.

Therefore, the type of bearing failure can be identified by the vibration characteristic frequency. The following is the vibration characteristic frequency formula of various bearing faults. The expression of outer ring fault fOD, inner ring fault fID, rolling element fault fBD, and cage fault fCD, are shown as follows [12]:

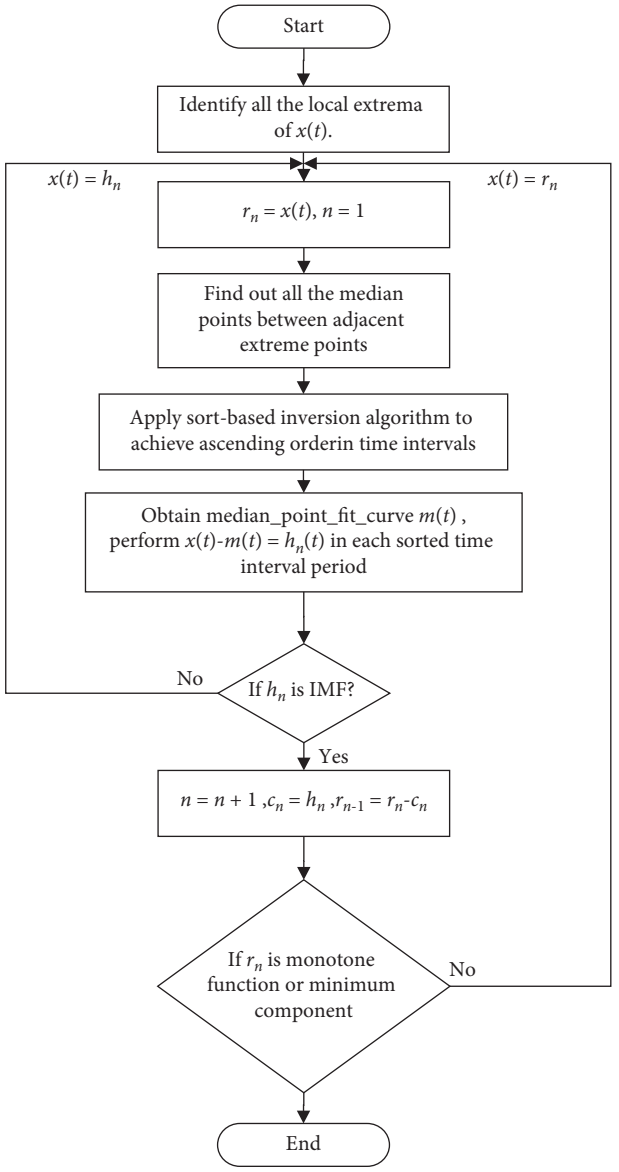


Figure 6: Implementation flowchart of MMD.

$$f_{\mathrm{OD}} = \frac{n}{2} f_{rm}\left(1 - \frac{d_b}{d_p}\cos\Phi\right), \tag{8}$$

$$f_{\mathrm{ID}} = \frac{n}{2} f_{rm}\left(1 + \frac{d_b}{d_p}\cos\Phi\right), \tag{9}$$

$$f_{\mathrm{BD}} = \frac{d_p}{2d_b} f_{rm}\left[1 - \left(\frac{d_b}{d_p}\cos\Phi\right)^2\right], \tag{10}$$

$$f_{\mathrm{CD}} = \frac{1}{2} f_{rm}\left(1 - \frac{d_b}{d_p}\cos\Phi\right), \tag{11}$$

where f_{rm} is the rotation frequency of motor, d_b and d_p are the diameter of the bearing rolling elements and the diameter of the bearing cage, respectively, n is the number of the bearing rolling elements, and Φ is the contact angle of rolling element.

As we can see above, the rolling element of motor rolling bearing fault is simulated and the time-domain waveform of the fault vibration signal is shown in Figure 7, where the vertical axis represents the vibration signal of the motor. For better observation, an enlarged view of Figure 7 during the time of zero to two seconds is presented in Figure 8. At the same time, the four IMF components (IMF1~IMF4) and one residual term (Res) obtained by adaptive MMD decomposition of the fault vibration signal are shown in Figure 9. Note that from the corresponding kurtosis value of each IMF component, we can conclude that since the kurtosis value of the IMF component of the 4th layer is the largest, the IMF4 component contains a lot of obvious fault characteristic information.

Therefore, the characteristics of the vibration signal as the rolling bearing outer ring in motor fault are verified, which demonstrates the effectiveness of the MMD method for the fault recognition.

Note that the MMD algorithm is able to alleviate the mixed modes problem in fault signal, where each IMF shows a certain periodicity. In this proposed method, the algorithm based on MMD and sort-based inversion is used to separate and alleviate the mixed modes. MMD decomposition of each quasi-margin term is re-decomposed to realize the self-adaptive function, making sure every IMF meets the conditions in Table 1. The result of EMD and MM obtained are both shown in Figure 9, illustrating through comparison with the conventional method that the algorithm successfully separated mixed modes problems in motor fault.

In Figure 9, the signal of IMF1 is completely extracted in the MMD method, while the EMD method still has mixed mode problem. Note that the resonance occurs with specific resonance frequency, and we manage to analyze with Hilbert–Huang spectrum for further needs of fault recognition.

In order to respond to the relationship between time-frequency-amplitude more intuitively, the three-dimensional Hilbert spectrum based on the information from the above IMFs is drawn in Figure 10. In Figure 10, there are fluctuations in the low-frequency part, but basically no energy distribution on the high-frequency part, which can be seen as linearly distributed and stable.

For better observation in the low-frequency part, comparative IMFs marginal spectrums of EMD and MMD in motor rolling bearing fault are given in Figure 11. The decomposition result of MMD has 5 IMFs. The IMFs contain enough physical meaning which are called effective intrinsic functions (EIMF). False intrinsic mode function (FIMF) components denote no physical meaning in IMFS. As can be seen in Figure 11, the result from EMD of Figure 11(a) conducts more numbers of FIMFs than MMD, which means the MMD method has better performance, particularly in fault recognition of rolling element bearings.

It can be seen from Figure 11(b) that the largest amplitude is around 0.35 with the frequency of near 38 Hz. According to theoretical calculation in equation (9), the inner ring is faulty with the calculated frequency of 37.6 Hz. Thus, the frequency near 38 Hz occupies the main

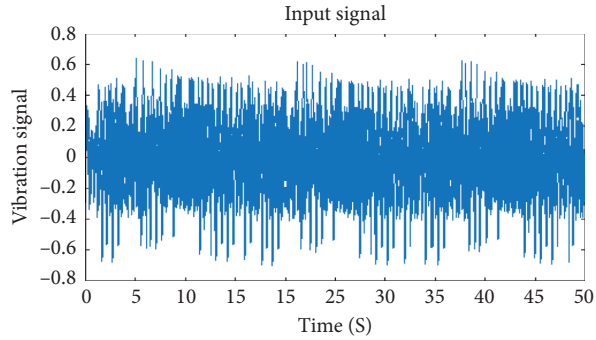


Figure 7: Synthetic signal waveform of motor rolling bearing simulated fault.

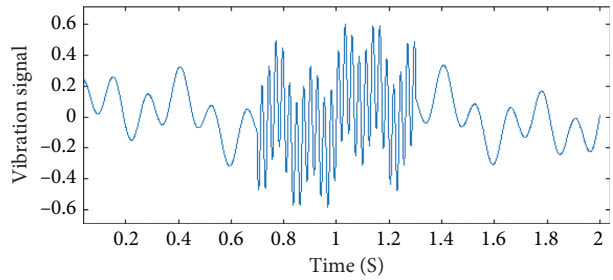


Figure 8: Synthetic signal waveform of motor rolling bearing simulated fault from the time of 0 s to 2 s.

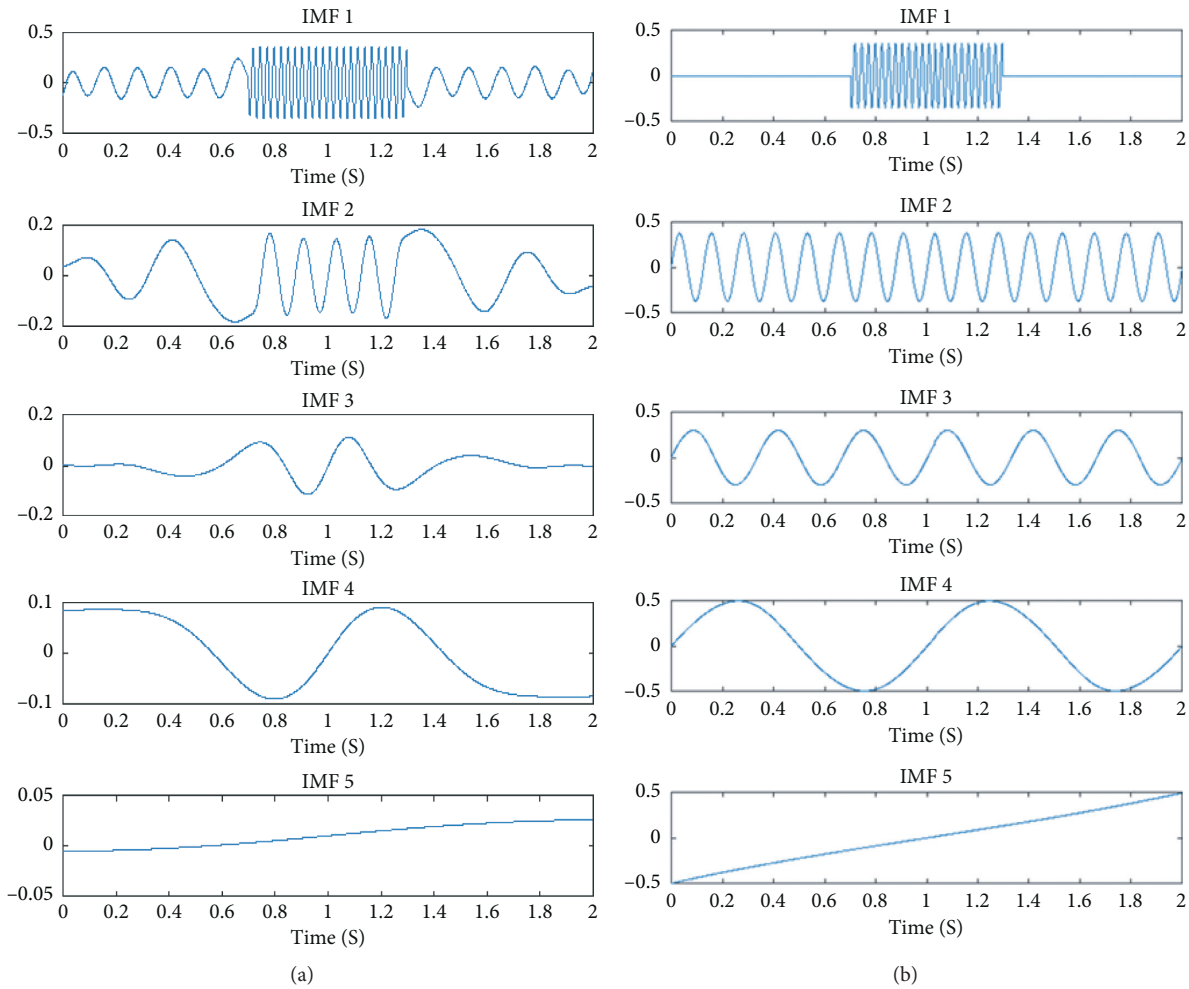


Figure 9: Result of signal of motor rolling bearing simulated fault. (a) EMD method; (b) MMD method.

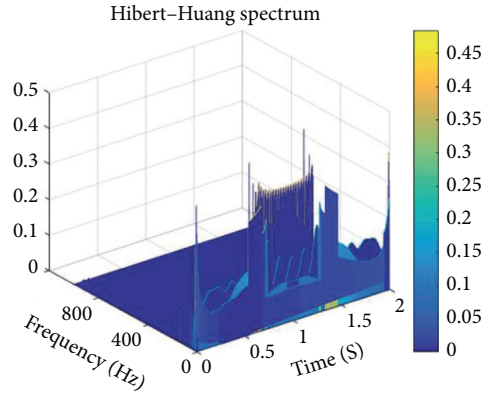


Figure 10: Hilbert–Huang spectrum waveform of motor rolling bearing simulated fault.

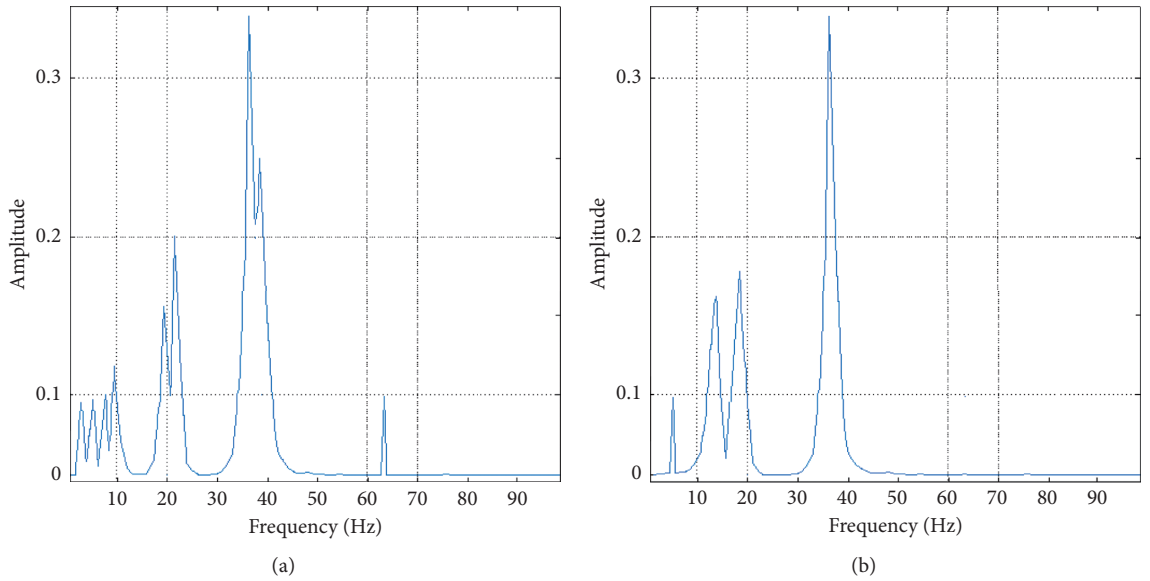


Figure 11: Each IMF marginal spectrum of MMD in motor rolling bearing simulated fault. (a) EMD method; (b) MMD method.

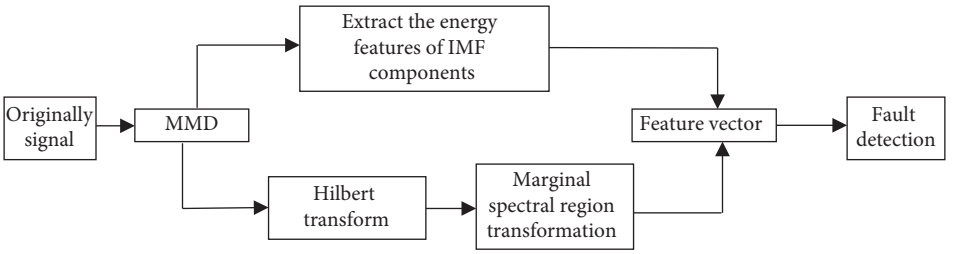


Figure 12: Process of MM when applied to fault recognition.

components, representing ability gathering, which proves obvious fault information and can be treated as a motor rolling bearing fault. That is, the MMD method can effectively extract the signal feature effectively and avoid the mixed modes problem.

Figure 12 shows the whole process of MMD applied in practice for extracting the motor fault; thus, the detected feature vector verifies the effectiveness of the proposed algorithm.

5. Conclusions

A fault recognition method for motor rolling bearing fault is put forward in this paper, which is based on a novel median-

point mode decomposition (MMD) with sort-based inversion algorithm. The MMD method is not only suitable for analyzing complex multicomponent signals but also chosen to precondition the vibration signal of the roller bearing to produce a set of IMF components. For the fact that the vibration signal is nonlinear and unstable, the MMD method keeps the algorithm self-adaptive for sorting out each variation of the extreme points interval with better and specific mode decomposition. Comparison simulations and experiments are operated to highlight the advantages of MMD in dealing with mixed mode problem in nonlinear signals.

In summary, MMD is a better choice when the signal needs time-frequency analysis, especially when the signal is nonlinear and nonstationary. The proposed method MMD keeps the advantages of EMD and EEMD and avoid mixed mode, which makes it capable of capturing the features of the signal in motor rolling bearing fault accurately. [13–18]

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Ganzhou Yao and Bishuang Fan contributed equally to this work.

Acknowledgments

This work was supported by the National Science and Technology project called "Research on Coordinated Control methods of Single-Phase-to-Ground Fault Flexible Arc Suppression and Protection for Distribution Networks" (No. 51877011).

References

[1] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[2] R. Gabriel and P. Flandrin, "One or two frequencies? The empirical mode decomposition answers," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 85–95, 2008.

[3] K. D. Seger, M. H. Al-Badrawi, J. L. Miksis-Olds, N. J. Kirsch, and A. P. Lyons, "An empirical mode decomposition-based recognition and classification approach for marine mammal vocal signals," *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. 3181–3190, 2018.

[4] C. Amo, L. de Santiago, R. Barea, A. LópezDorado, and L. Boquete, "Analysis of gamma-band activity from human EEG using empirical mode decomposition," *Sensors*, vol. 17, no. 5, p. 989, 2017.

[5] Y. Li, J. Liu, and Y. Wang, "Railway wheel flat recognition based on improved empirical mode decomposition," *Shock and Vibration*, vol. 2016, Article ID 4879283, 14 pages, 2016.

[6] J. Zheng, J. Cheng, and Y Yang, "Partly ensemble empirical mode decomposition: An improved noise-assisted method for eliminating mode mixing," *Signal Processing*, vol. 96, pp. 362–374, 2014.

[7] G. Li, Z. Yang, and H. Yang, "Noise reduction method of underwater acoustic signals based on uniform phase empirical mode decomposition, amplitude-aware permutation entropy, and pearson correlation coefficient," *Entropy*, vol. 20, no. 12, p. 918, 2018.

[8] J.-C. Nunes and E. Delechelle, "Empirical mode decomposition: Applications on signal and image processing," *Advances in Adaptive Data Analysis*, vol. 1, no. No. 1, pp. 125–175, 2009.

[9] C. Wang, H. Li, and D. Zhao, "A preconditioning framework for the empirical mode decomposition method," *Circuits System Signal Process*, vol. 37, no. 12, pp. 5417–5440, 2018.

[10] R. T. Rato, M. D. Ortigueira, and A. G. Batista, "On the HHT, its problems, and some solutions," *Mechanical Systems and Signal Processing*, vol. 22, no. 6, 2008.

[11] R. Ho and K. Hung, "A comparative investigation of mode mixing in EEG decomposition using EMD, EEMD and M-EMD," in *IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 203–210, Penang, Malaysia, April 2020.

[12] X. Hu, S. Peng, and W.-L. Hwang, "EMD revisited: A new understanding of the envelope and resolving the mode-mixing problem in AM-FM signals," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1075–1086, 2012.

[13] T.-L. Kung and C.-N. Hung, "Estimating the subsystem reliability of bubblesort networks," *Theoretical Computer Science*, vol. 670, pp. 45–55, 2017.

[14] G. Xua, Z. Yangb, and S. Wang, "Study on mode mixing problem of EMD," in *Proceedings of the Joint International Information Technology, Mechanical and Electronic Engineering Conference*, Chongqing China, May 2016.

[15] Y. R. Du, L. H. Chen, and H. Jin, "A new view of mode mixing phenomenon," *Applied Mechanics and Materials*, vol. 532, pp. 134–137, 2014.

[16] Y. Kopsinis and S. McLaughlin, "Investigation and performance enhancement of the empirical mode decomposition method based on a heuristic search optimization approach," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, 2008.

[17] Y. O. Adu-Gyamfi, N. O. Attoh-Okine, and A. Y. Ayenu-Prah, "Critical Analysis of different hilbert-huang algorithms for pavement profile evaluation," *Journal of Computing in Civil Engineering*, vol. 24, no. 6, 2010.

[18] B. Xu, Y. Sheng, P. Li, Q. Cheng, and J. Wu, "Causes and classification of EMD mode mixing," *Vibroengineering Procedia*, vol. 22, pp. 158–164, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5648058, 10 pages
https://doi.org/10.1155/2020/5648058

Research Article

Stability Coordinated Control of Distributed Drive Electric Vehicle Based on Condition Switching

Zhao Jingbo,[1] **Chen Jie**,[2] **and Liu Chengye**[2]

[1]*Changzhou Institute of Technology, Changzhou 213032, China*
[2]*Jiangsu University of Technology, Changzhou 213001, China*

Correspondence should be addressed to Zhao Jingbo; 66822871@qq.com

Received 25 July 2020; Revised 29 September 2020; Accepted 18 October 2020; Published 31 October 2020

Academic Editor: Yong Chen

The distributed drive electric vehicle is a complex hybrid system including discrete events and continuous events. In order to coordinate the longitudinal and lateral motion of the distributed drive electric vehicle, a hierarchical control method was proposed. In the upper layer, the body attitude tracking controller based on sliding mode control algorithm was established to accurately analyze the driving expectation and to track the longitudinal speed, the lateral speed, and the yaw rate of the vehicle. In the lower layer, the switching controller based on the hybrid theory was established to improve the driving stability under various working conditions. The switching controller can switch between control strategies according to the working conditions. The joint simulation was carried out under various working conditions using Simulink and CarSim software. The results showed that the controller can coordinate the longitudinal and lateral motion of the vehicle well in linear acceleration and sinusoidal acceleration conditions and can strictly track the driving expectation and maintain the desired body posture. And another, the controller can be switched according to the working conditions and control strategies accurately and smoothly and can ensure stable driving in the constant speed single lane change condition. The controller can reveal the continuous behavior characteristics of the vehicle and reflect the characteristics of discrete events by coordinating the longitudinal and lateral motion of the vehicle. It improves the stability and control performance of the distributed drive electric vehicle under various working conditions.

1. Introduction

Since the 21st century, with the increasing environmental and energy problems, distributed drive electric vehicle as a new type of new energy vehicle has gradually entered the field of vision of researchers. The chassis layout of the distributed drive electric vehicle is novel. Four driving motors are installed in the wheel rim, respectively, and the transmission structure of the vehicle is cancelled, and the drive motor is controlled independently [1]. Therefore, it has the characteristics of fast response speed and independent torque control and brings huge control potential for the vehicle. As a kind of vehicle chassis control technology, the longitudinal and transverse motion control technology plays an important role in improving the driving stability of vehicles. However, the research on the longitudinal and transverse motion control technology is mainly focused on the traditional internal combustion engine vehicles, and the control system is not perfect. The optimization algorithm of longitudinal and horizontal tire force is single, which cannot comprehensively consider the various working conditions faced by the vehicle in the process of driving. It is easy to cause the unreasonable distribution of the longitudinal and lateral forces of the vehicle and affect the driving attitude and handling stability of the vehicle. Some scholars use the average distribution of vehicle tire force to improve the stability of the vehicle, which does not fully consider the impact of vehicle steering and vehicle load transfer on tire force, so the vehicle stability cannot be fully improved. Therefore, it is of great significance to study the longitudinal and lateral motion of the vehicle under multiple working conditions to control the vehicle's driving attitude and improve the lateral driving stability and economy.

In order to solve the above problems, this paper proposes a distributed drive electric vehicle stability controller based on multi driving conditions. The controller adopts a top-down hierarchical control architecture. The upper controller tracks the desired body posture based on the sliding mode control algorithm and calculates the longitudinal force, lateral force, and yaw moment required by the vehicle. The lower controller establishes the vehicle condition switching controller by analyzing the continuous state characteristics and discrete state characteristics of the vehicle system based on the hybrid theory and designs the control strategy for each working condition which scientifically and reasonably distributes the longitudinal and transverse forces of the vehicle, so as to ensure that the vehicle can keep stable running in all driving conditions and comprehensively improve the driving stability of the vehicle. Finally, a simulation platform is built based on Simulink and CarSim to verify the effectiveness of the stability controller for distributed drive electric vehicles under multiple working conditions.

2. Stability Coordination Controller

Based on a hierarchical control framework, a stability coordinated controller for multi driving conditions of a distributed drive electric vehicle is established. The specific structure is shown in Figure 1. The stability coordination controller mainly includes the following parts: the upper vehicle reference model and body attitude tracking controller, the lower working condition switching controller, and the actuator controller. The controller receives the steering wheel angle signal and accelerator pedal signal from the driver and calculates the expected driving state of the vehicle through the reference model. The body attitude tracking controller tracks the expected driving state of the vehicle. The condition switching controller divides the driving condition of the vehicle into a straight driving condition and steering condition and optimizes the distribution of the force for the different driving conditions of the vehicle. At last, the actuator controller controls the vehicle drive/steering motor. Finally, the purpose of improving the driving stability of vehicles is achieved.

3. Upper Controller

3.1. Reference Model. The reference model is used to receive the driver's operation information (including steering wheel angle and accelerator pedal opening) and calculate the expected running state information of the vehicle (including the expected longitudinal speed, the expected lateral speed, and the expected yaw rate of the vehicle). At the same time, the reference model transmits the expected information to the vehicle body attitude tracking controller to provide the tracking target for the body motion controller.

In order to avoid coupling between the longitudinal system and transverse system, the longitudinal system and transverse system are designed separately in this paper.

Yaw rate is the key data to represent the driving state of the vehicle. Therefore, it is necessary to obtain the relationship between steering wheel input and yaw rate of the vehicle. The paper establishes a linear two degrees-of-freedom vehicle model as the reference model of the vehicle steering system to represent the relationship. The output results of the reference model are all the desired data.

The transfer matrix of vehicle linear two degree-of-freedom steering model is as follows.

$$\begin{bmatrix} \dot{V}_y \\ \dot{\varphi} \end{bmatrix} = \begin{bmatrix} \frac{k_f + k_r}{mV_x} & \frac{l_f k_f - l_r k_r}{mV_x} - V_x \\ \frac{l_f k_f - l_r k_r}{I_z V_x} & \frac{l_f^2 k_f + l_r^2 k_r}{I_z V_x} \end{bmatrix} \begin{bmatrix} V_y \\ \varphi \end{bmatrix} + \begin{bmatrix} -\frac{k_f}{m} & -\frac{k_r}{m} \\ -\frac{l_f k_f}{I_z} & \frac{l_r k_r}{I_z} \end{bmatrix} \begin{bmatrix} \delta_f \\ \delta_r \end{bmatrix}. \tag{1}$$

Through mathematical derivation, the desired yaw rate of the vehicle can be expressed as

$$\dot{\varphi} = \frac{V_x \delta_f}{\left(1 + K_g V_x^2\right)\left(l_f + l_r\right)}. \tag{2}$$

Among them,

$$K_g = \frac{m}{\left(l_f + l_r\right)^2}\left(\frac{l_f}{k_r} - \frac{l_r}{k_f}\right). \tag{3}$$

After considering the yaw rate constraint, the desired yaw rate can be expressed as

$$\dot{\varphi}_d = \min\left\{|\varphi|, \left|\mu \cdot g / V_x\right|\right\}, \tag{4}$$

where μ is the ground friction coefficient and g is 9.8 m/s^2.

The reference model of the longitudinal system is mainly to obtain the expected longitudinal speed of the vehicle, which can be determined by the longitudinal acceleration in the time domain.

$$V_{xd} = V_{x0} + \int_0^t a_{xd}(\tau)\mathrm{d}\tau, \tag{5}$$

where k_f and k_r are the side deflection stiffness of front and rear wheels, respectively; δ_f and δ_r are the front and rear wheel angles; V_x is the longitudinal speed of the vehicle; V_y is the lateral velocity of the vehicle; $\dot{\varphi}$ is the yaw rate of the vehicle; m is the mass of the vehicle; l_f is the distance from vehicle centroid to front axle; l_r is the distance from vehicle centroid to rear axle; I_z is the moment of inertia; and V_{x0} is the initial speed of the vehicle.

3.2. Body Attitude Tracking Controller. The function of the body motion controller is to calculate the expected longitudinal total torque, expected total lateral moment, and expected yaw moment of the vehicle through the advanced control algorithm according to the information of the reference model. However, the vehicle system has complex nonlinear characteristics. The paper selects the

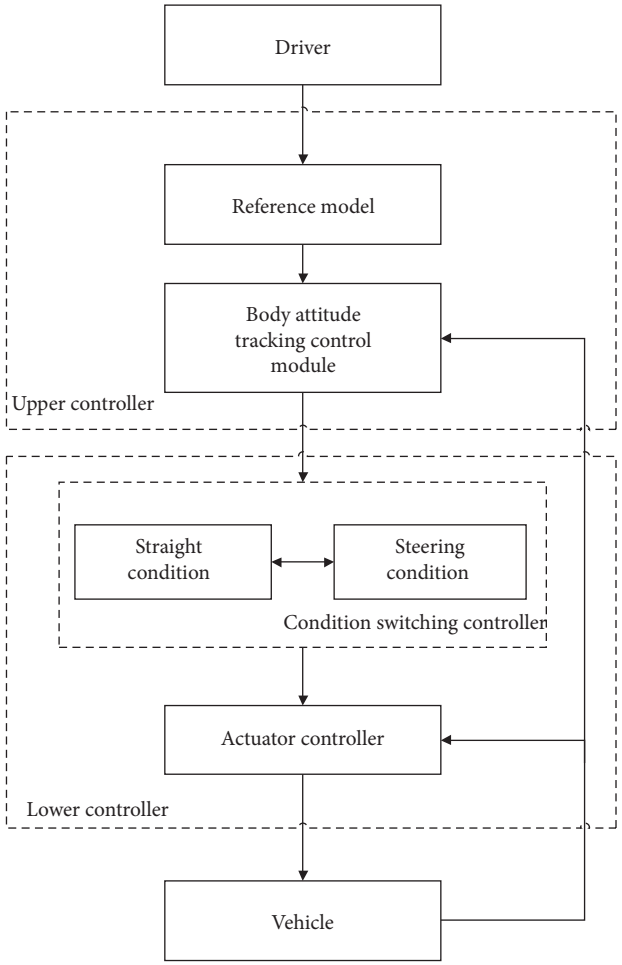


Figure 1: Framework of stability coordination controller.

sliding mode variable structure control algorithm to track the longitudinal speed, lateral speed, and yaw angle of the vehicle by comparing a variety of advanced control algorithms.

The sliding mode control algorithm can make the tracking error converge to zero on the designed sliding mode surface and can track the expectation of the reference model output well.

The sliding surface is designed as shown in the following equations:

$$S_1 = V_x - V_{xd}, \tag{6}$$

$$S_2 = V_y - V_{yd}, \tag{7}$$

$$S_3 = \dot{\varphi} - \dot{\varphi}_d. \tag{8}$$

The design sliding mode variable structure control rate can be expressed as

$$\dot{S}_k = u_{kn}, \quad \forall k \in \{1, 2, 3\}. \tag{9}$$

In order to suppress the chattering of the system, the linear saturation function is used as the sliding surface S_1, S_2, and S_3.

$$u_{kn} = -\eta_{kn}\mathrm{sat}\left(\frac{S_k}{\phi_k}\right), \quad \forall k \in \{1, 2, 3\},$$

$$\mathrm{sat}\left(\frac{S_k}{\phi_k}\right) = \begin{cases} \dfrac{S_k}{\phi_k}, & \text{if } |S_k| < \phi_k, \\ \mathrm{sgn}\left(\dfrac{S_k}{\phi_k}\right), & \text{if } |S_k| \geq \phi_k. \end{cases} \tag{10}$$

The Lyapunov function is constructed to determine the stability of the system, as shown in the following equation:

$$V_k = \frac{S_k^2}{2}, \quad \forall k \in \{1, 2, 3\}, \tag{11}$$

$$\dot{V}_k = S_k \cdot \dot{S}_k = -S_k \cdot \mathrm{sat}\left(\frac{S_k}{\phi_k}\right) \leq 0 \quad \forall k \in \{1, 2, 3\}. \tag{12}$$

It can be seen from equation (12) that the controller satisfies the stability condition and is stable.

Finally, the total longitudinal force, total lateral force, and yaw moment required for vehicle attitude tracking are obtained.

$$F_{xd} = m\left(\dot{V}_x - \dot{V}_y\dot{\varphi}\right) = m\left(\dot{V}_{xd} - \eta_{1n}\mathrm{sat}\left(\frac{S_1}{\phi_1}\right) - \dot{V}_y\dot{\varphi}\right),$$

$$F_{yd} = m\left(\dot{V}_y + V_x\dot{\varphi}\right) = m\left(\dot{V}_{yd} - \eta_{2n}\mathrm{sat}\left(\frac{S_2}{\phi_2}\right) + \dot{V}_x\dot{\varphi}\right),$$

$$M_{zd} = I_z\ddot{\varphi} = I_z\left(\ddot{\varphi}_d - \eta_{3n}\mathrm{sat}\left(\frac{S_3}{\phi_3}\right)\right). \tag{13}$$

Here, F_{xd} and F_{yd} represent the expected total longitudinal force and expected total lateral force. M_{zd} is the desired yaw moment.

4. Lower Controller

4.1. Condition Switching Controller. Based on hybrid control theory, the main function of the controller is to determine and switch real-time driving conditions according to vehicle information and switch the corresponding control strategy to optimize the tire force distribution. The distributed drive electric vehicle is a hybrid system, which can switch the driving mode of the vehicle in real time according to the change of the discrete signal of the vehicle [2]. The driving condition of the vehicle is divided into two driving conditions: straight driving condition and steering condition, and appropriate control strategies are developed, respectively, as shown in Figure 2.

The hybrid system is modeled by automata, as shown in the following equation:

$$H = (Q, X, V, Y, l\mathrm{int}, f, \ln v, E, \Psi), \tag{14}$$

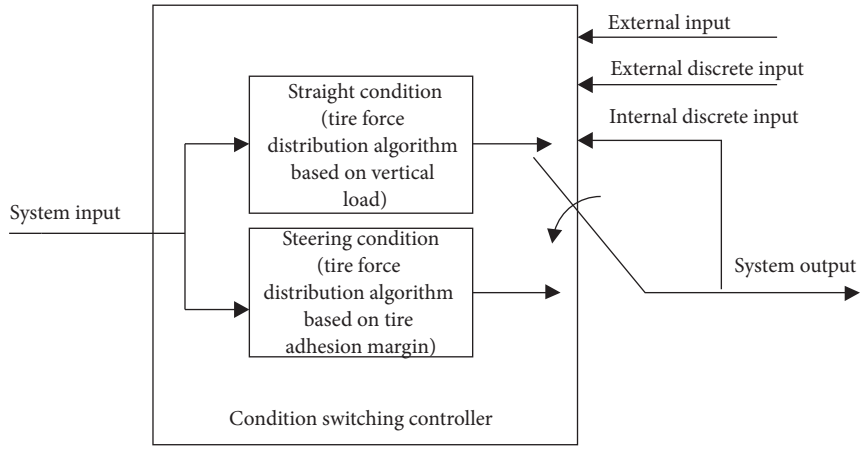


Figure 2: Condition switching controller.

where Q is the driving condition of the vehicle: {straight condition, steering condition}; X is the continuous state variable of the system: $\{F_{xfl}, F_{xfr}, F_{xrl}, F_{xrr}, F_{yfl}, F_{yfr}, F_{yrl}$, and $F_{yrr}\}$; and V is the continuous input variable $\{a_x, a_y, F_{xd}, F_{yd}$, and $M_{zd}\}$ and discrete input variable $\{S_1$ and $S_2\}$.S_1 and S_2 are the control strategies of two driving conditions; Y is the continuous output variable: $\{F_{xfl}, F_{xfr}, F_{xrl}, F_{xrr}, F_{yfl}, F_{yfr}, F_{yrl}$, and $F_{yrr}\}$; $l$int is the initial state of the system; $\ln v$ is the set of invariant state quantity of the system E is the set of discrete switching events: $\{E_1$ and $E_2\}$; and ψ specifies an allowable input field for each state. Here, F_x, F_y, and F_z represent the longitudinal force, lateral force, and vertical load of the vehicle. At the same time, fl, $frrl$, and rr represent the left front, right front, left rear, and right rear wheels of the vehicle, respectively. a_x and a_y represent the longitudinal and lateral acceleration of the vehicle.

When the vehicle is in the straight driving condition, the vehicle adopts the tire force distribution method based on the vertical load of the tire; when the vehicle is in the steering condition, the vehicle adopts the tire force distribution method based on the minimum tire adhesion margin. The monitoring data are driver steering wheel angle δ, steering wheel angle velocity δ_ω, and vehicle yaw angle acceleration $\dot{\varphi}$. The monitoring data switching thresholds are set, respectively. When the monitoring data are lower than the system set thresholds, it is determined that the vehicle is in straight running condition; otherwise, it is determined that the vehicle is in steering condition.

4.2. Optimized Distribution of Tire Force. The distributed drive electric vehicle has four drive motors and four steering motors. The degree of freedom required to control is far less than the number of controllable actuators. The system is highly redundant and overdrive system. Therefore, the control strategy is designed by control distribution theory. The optimal distribution of tire force based on the control distribution theory can effectively improve the dynamic response of the vehicle.

Under the condition of straight driving, the longitudinal force has a great influence on the driving state of the vehicle. And then, the vertical load of the tire will move between the front and rear axles. The driving force distribution method based on tire load can better meet the requirements of vehicle power and safety in the straight driving condition.

The vertical load of tire is shown in the following equations:

$$F_{zfl} = m\left(\frac{gl_r}{2(l_r+l_f)} - \frac{ha_x}{2(l_r+l_f)} - \frac{hl_r a_y}{t_f(l_r+l_f)}\right), \quad (15)$$

$$F_{zfr} = m\left(\frac{gl_r}{2(l_r+l_f)} - \frac{ha_x}{2(l_r+l_f)} + \frac{hl_r a_y}{t_f(l_r+l_f)}\right), \quad (16)$$

$$F_{zrl} = m\left(\frac{gl_r}{2(l_r+l_f)} + \frac{ha_x}{2(l_r+l_f)} - \frac{hl_f a_y}{t_r(l_r+l_f)}\right), \quad (17)$$

$$F_{zrr} = m\left(\frac{gl_r}{2(l_r+l_f)} + \frac{ha_x}{2(l_r+l_f)} + \frac{hl_f a_y}{t_r(l_r+l_f)}\right). \quad (18)$$

The total vertical load of the wheel can be expressed by the following equation:

$$F_{zt} = F_{zfl} + F_{zfr} + F_{zrl} + F_{zrr}. \quad (19)$$

Finally, the driving force of each wheel can be shown in the following equation:

$$F_{xij} = \frac{F_{zij}}{F_{zt}} F_{zd}, \quad i \in \{f, r\}, j \in \{l, r\}, \quad (20)$$

where t_f and t_r are the front and rear track width and h is the height from the vehicle center of mass to the ground. At this driving condition, the tire lateral force is evenly distributed under the straight driving condition.

When the vehicle is in the steering condition, the longitudinal force, lateral force, and yaw moment generated by the vehicle to maintain the body attitude will have an impact on the driving state of the vehicle. At this condition, the tire force distribution should focus on improving the stability of the vehicle. And the tire force distribution method based on

the tire load coefficient is adopted. The smaller the load factor of the vehicle tire, the greater the potential of the tire, the higher the stability of the vehicle.

$$\min \quad J = \sum_{\substack{i=f,r \\ j=l,r}} \frac{F_{xij}^2 + F_{yij}^2}{\mu_{ij}^2 F_{zij}^2}, \tag{21}$$

where μ is the ground adhesion coefficient and F_z is the vertical force of each tire.

At steering condition, the distributed longitudinal force, lateral force, and yaw moment shall also meet the kinematic equation of the vehicle.

$$\begin{aligned} F_{xd} &= F_{xfl} + F_{xfr} + F_{xrl} + F_{xrr}, \\ F_{yd} &= F_{yfl} + F_{yfr} + F_{yrl} + F_{yrr}, \\ M_{zd} &= l_f\left(F_{yfl} + F_{yfr}\right) - l_r\left(F_{yrl} + F_{yrr}\right) + \frac{t_f}{2}\left(-F_{xfl} + F_{xfr}\right) \\ &\quad + \frac{t_r}{2}\left(-F_{xrl} + F_{xrr}\right). \end{aligned} \tag{22}$$

In addition, the longitudinal and lateral forces of the vehicle during driving should meet the constraints of the vertical forces of the tire. That is to say, the limit condition of the friction circle should be met.

$$F_{xij}^2 + F_{yij}^2 \le \mu_{ij}^2 F_{zij}^2. \tag{23}$$

At the same time, the driving force and lateral force of the vehicle also need to meet the maximum torque requirements of the motor.

$$F_{xij} \le \frac{T_{\max}}{r}, \quad i \in \{f, r\}, j \in \{l, r\}, \tag{24}$$

where $T_{\max}$ is the maximum output torque of the motor and r is the effective radius of the tire.

In this paper, the interior point method of quadratic programming (SQP) is used to solve the problem with inequality constraints. Finally, the *S*-function in Simulink is used to write the objective function. And the optimization problem is solved iteratively to get the optimal distribution of tire force.

4.3. Actuator Controller. The main function of the actuator controller is to convert the received optimal driving force and lateral force into the driving torque and steering angle of the actuator motor. Accurate control of vehicle actuators is the key to improve vehicle driving stability.

It can be seen from tire dynamics that the longitudinal force of tire can be realized by directly controlling the torque of the driving motor. According to the longitudinal force model of single wheel and the principle of moment balance, the driving moment of the driving motor can be calculated by using the following equation:

$$T_{wij} = R_{wij} F_{wij} + J_{wij} \omega_{wij} + T_{bij}, \quad i \in \{f, r\}, j \in \{l, r\}. \tag{25}$$

The tire lateral force cannot be directly transformed into the wheel angle of the steering motor. It needs to be solved indirectly by the inverse model of tire cornering. The tire model shown in equation (26) is used to realize the model [3]. The tire model can better represent the linear relationship between tire lateral force and tire cornering angle:

$$F_y = -CG_x \frac{\mu}{k} \tan^{-1}\left(\frac{k\alpha}{\mu}\right). \tag{26}$$

G_x and k are factors defined as

$$\begin{aligned} G_x &= \sqrt{1 - \left(\frac{F_x}{\mu F_z}\right)^2}, \\ k &= C\frac{\pi}{2}\frac{1}{F_z}. \end{aligned} \tag{27}$$

According to the tire model, the angle between the tire running direction and the coordinate axis can be expressed by the following equation:

$$\alpha = \frac{\mu}{k} \tan\left(\frac{-F_y k}{CG_x \mu}\right). \tag{28}$$

The angle between the driving direction of the tire and the coordinate axis can be known from the side slip of the tire:

$$\sigma_{ij} = \delta_{ij} + \alpha_{ij}. \tag{29}$$

At the same time, it can be seen from the vehicle dynamics model:

$$\begin{aligned} \sigma_{lf,rl} &= \tan^{-1}\left(\frac{\left(V_y - l_r\varphi\right)}{\left(V_x \mp t_f \varphi/2\right)}\right), \\ \sigma_{lr,rr} &= \tan^{-1}\left(\frac{\left(V_y - l_r\varphi\right)}{\left(V_x \mp t_r \varphi/2\right)}\right). \end{aligned} \tag{30}$$

Finally, it can be seen that the required wheel angle of the vehicle can be expressed by the following equation:

$$\delta_{ij} = \sigma_{ij} - \alpha_{ij}, i \in \{f, r\}, j \in \{l, r\}. \tag{31}$$

5. Simulation Verification

In order to prove the effectiveness of the stability coordinated controller designed, this paper establishes a joint simulation model of Simulink and CarSim and carries out simulation verification under the conditions of linear acceleration, sinusoidal acceleration condition, and uniform single lane change condition. The input of the coordination controller is the acceleration determined by the steering wheel angle and accelerator pedal opening of the vehicle. The simulation vehicle includes four driving motors and four

Table 1: Main parameters of the vehicle.

Parameters	Numerical value
Distance from the center of the mass to front axle, l_f (m)	1.232
Distance from the center of the mass to rear axle, l_r (m)	1.468
Wheel base, L (m)	2.7
Vehicle mass, m (kg)	1723
Front axle track, t_f (m)	1.416
Rear axle track, t_r (m)	1.375
Height from the center of mass to ground, h_g (m)	0.54
Effective radius of tire, R (m)	0.28

steering motors, and the specific parameters are shown in Table 1.

5.1. Linear Acceleration Condition. The linear acceleration condition is mainly to verify the effectiveness of the vehicle coordination controller for driver driving expectation tracking.

In the condition of linear acceleration, the steering wheel angle input is always 0. The initial speed of vehicle is 36 km/h, and the acceleration curve can be shown in Figure 3. The ground is flat. The coefficient of adhesion is 0.8. The acceleration is given to the vehicle in the 2 seconds when the vehicle is running, and the acceleration is increased to 1.5 m/s^2 in one second and keep it for 5 seconds. At last, the vehicle acceleration is reduced to 0 in the sixth second.

The specific simulation results are shown in Figures 4 and 5.

Figure 4 shows the longitudinal speed chart of the vehicle. When the vehicle starts to accelerate in 1 s and stops to accelerate in 6 s, the vehicle speed accelerates from 36 km/h to 57.6 km/h. The vehicle speed curve is smooth, which tracks the expected speed of the vehicle well and meets the acceleration expectation of the driver. Figure 5 shows the actual acceleration curve of the vehicle. In the simulation process, the actual acceleration curve can better track the expected acceleration curve, and the vehicle response is fast and accurate, which fully meets the acceleration expectation of the driver. It can be seen from the simulation results of the linear acceleration condition that the actual driving speed of the vehicle can track the target speed quickly and accurately. At the same time, it can meet the acceleration needs of the driver in the linear condition, and the coordination controller can better track the driving expectation of the vehicle. Therefore, the coordinated controller is effective.

5.2. Sinusoidal Acceleration Condition. The sinusoidal acceleration condition is to verify the effectiveness of the vehicle stability coordination controller to improve the lateral stability under the steering condition. The steering wheel input and accelerator pedal input are shown in Figures 6 and 3. The initial speed of the vehicle was set as 36 km/h. The ground is flat, and the coefficient of adhesion is 0.8.

The simulation results are shown in Figures 7–9.

Figure 7 shows the comparison between the actual yaw rate and the desired yaw rate of the vehicle. When the steering wheel angle and longitudinal speed increase continuously, the yaw rate of the vehicle will produce errors when tracking the desired yaw rate of the vehicle, but the errors are small where maximum error is less than 0.3 deg/s. It cannot have a big impact on the stability of the vehicle. In this condition, the controller can complete the tracking of the yaw rate as a whole and track the longitudinal direction of the vehicle better speed. It has a better impact to improve the stability of the vehicle. In the first four seconds of the simulation, the vehicle can track the longitudinal acceleration very well. However, the longitudinal acceleration of the vehicle will fluctuate due to the excessive change rate of the steering angle of the vehicle at four seconds and the sixth. Therefore, it can quickly track the expected acceleration of the vehicle after the fluctuation. To sum-up, the vehicle coordination controller can better complete the driver's steering intention, track the vehicle's yaw acceleration and acceleration, and improve the vehicle's yaw stability and safety to a certain extent.

5.3. Uniform Single Lane Change Condition. The uniform single lane change condition is mainly to verify the effectiveness of the hybrid controller in the process of vehicle driving. In this case, it is necessary to design a driver model to convert the model path input into the vehicle steering wheel input. The specific driver model is not described here. The steering wheel angle input is shown in Figure 10. The vehicle speed is set as 36 km/h. The ground is flat without slope, and the ground adhesion coefficient is 0.8.

As shown in Figure 11, the speed of the simulation vehicle is maintained at 35.999 km/h when driving straight, and the speed fluctuates slightly when turning. The maximum error is 0.003 km/h in this condition. The impact on the overall speed can be ignored basically. The vertical and horizontal coordination controller can better track the longitudinal speed of the vehicle.

It can be seen from Figure 12 that the vehicle cannot fully track the desired yaw rate in which the maximum error is less than 0.25 deg/s when turning. However, the yaw rate curve of the vehicle is smooth and has no fluctuation, which can better maintain the yaw stability of the vehicle under this working condition in this simulation process. Figure 13 is a schematic diagram of vehicle working condition switching. 1 represents straight driving condition and 2 represents steering condition. It can be seen from Figure 13 that the hybrid controller can switch vehicle conditions smoothly according to the change of

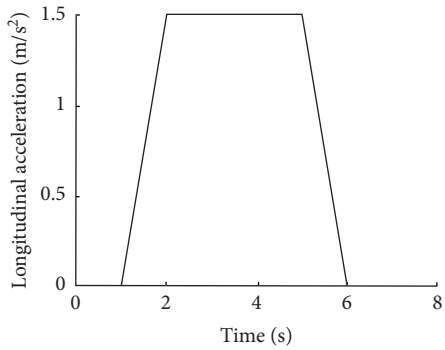


Figure 3: Target curve of longitudinal acceleration.

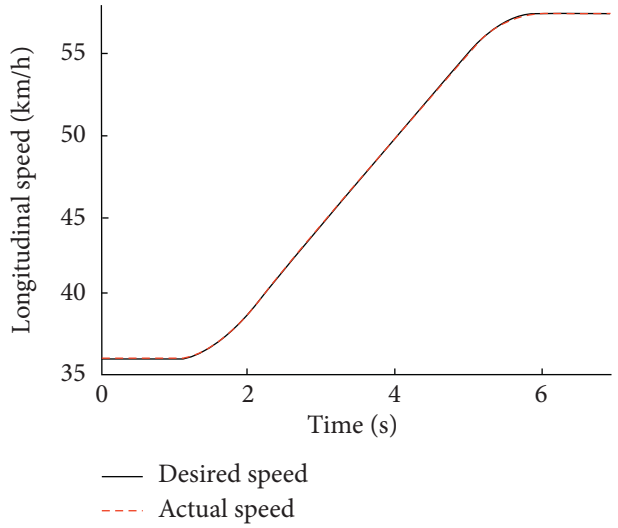


Figure 4: Longitudinal speed.

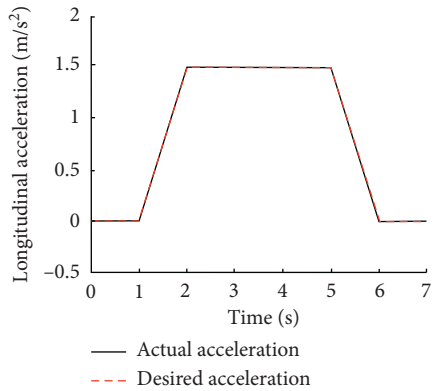


Figure 5: Actual longitudinal acceleration curve.

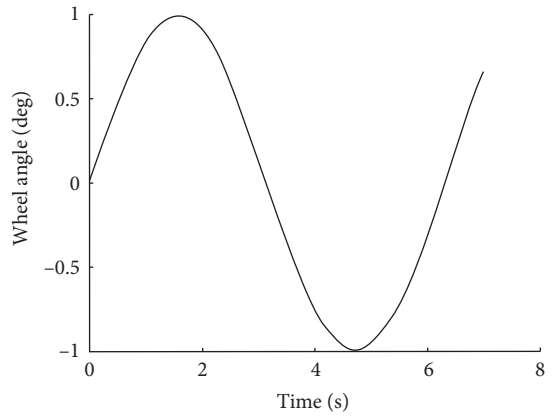


Figure 6: Front-wheel angle step input.

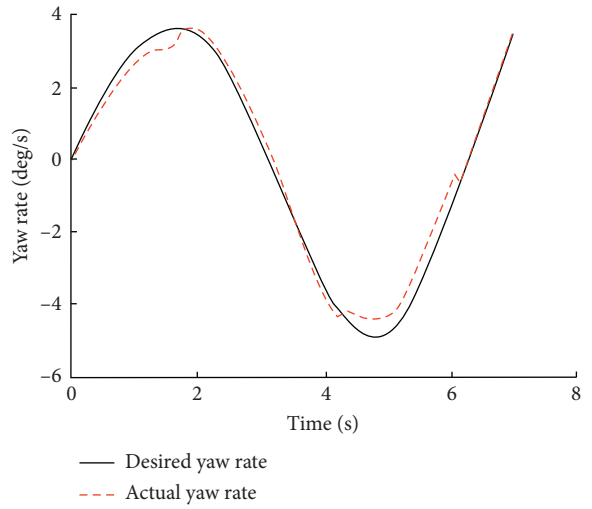


Figure 7: Comparison of yaw rate.

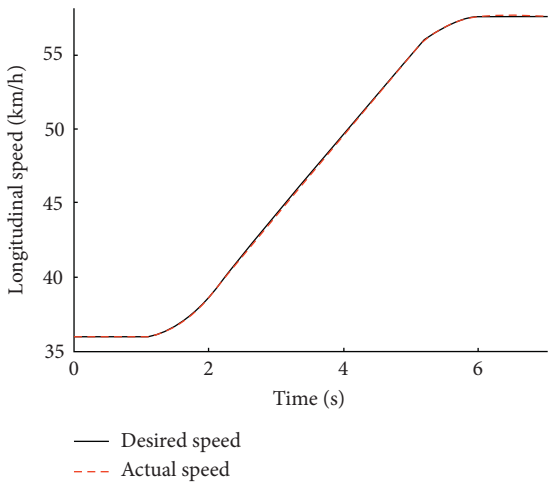


Figure 8: Comparison of longitudinal velocity diagram.

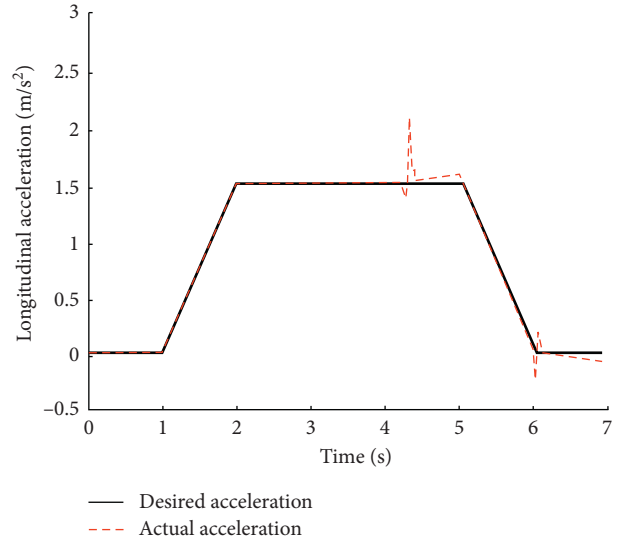


Figure 9: Comparison of longitudinal acceleration.

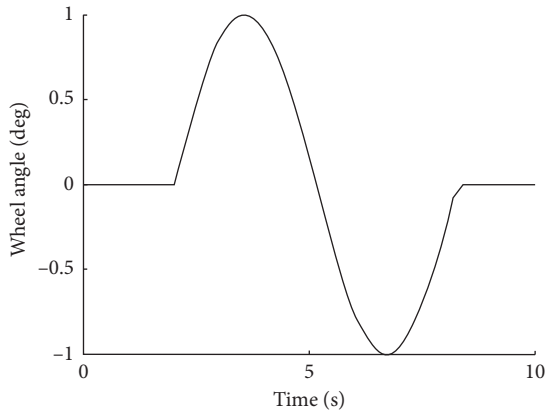


Figure 10: Steering wheel angle of single moving line.

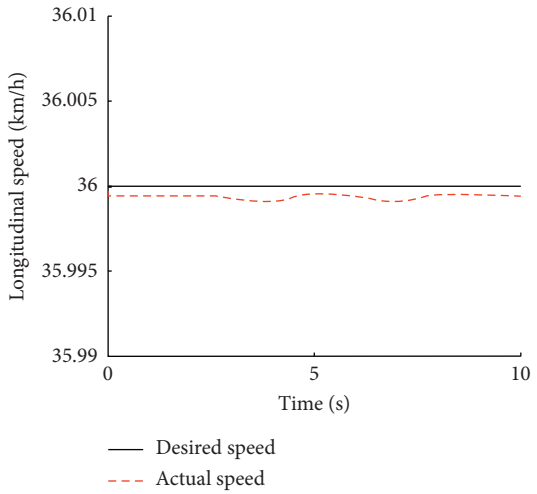


Figure 11: Comparison curve of longitudinal speed.

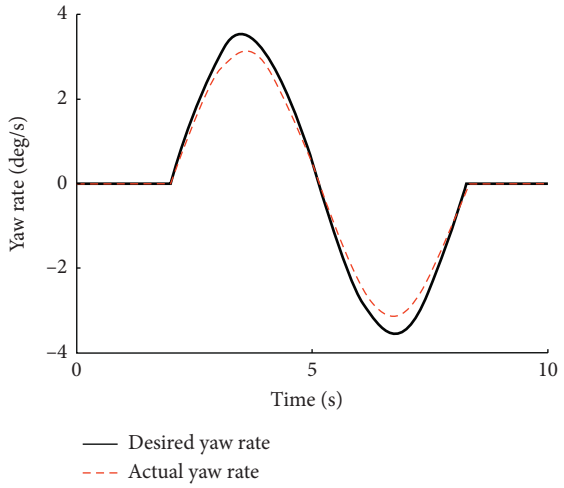


Figure 12: Comparison curve of yaw rate.

vehicle monitoring data and select the corresponding control strategy to optimize the distribution of vehicle tire force.

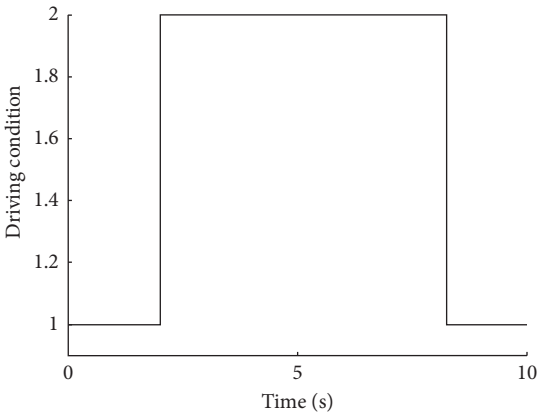


Figure 13: Working condition switching diagram.

6. Conclusion

In order to improve the driving stability of the vehicle under multiple working conditions, the longitudinal and transverse stability controller of the distributed drive electric vehicle is established in this paper. The body attitude tracking controller is established based on the sliding mode variable structure control idea, and the vehicle condition switching controller is established based on the hybrid control theory. At last, the distributed vehicle is simulated and verified under the multi conditions. The simulation results show that the vehicle stability coordination controller can meet the driver's driving expectations and improve vehicle stability and safety. The vehicle condition switching controller can switch vehicle working conditions and control strategies in real time according to vehicle monitoring information and optimize tire force distribution to realize yaw stability control of the vehicle under different working conditions and meet the driving requirements of drivers under different working conditions. Therefore, the vehicle stability controller is effective.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Funding from the National Natural Science Foundation of China (Grant no. 61503163) and the Key University Science Research Project of Jiangsu Province (Grant no. 18KJA580004) is gratefully acknowledged.

References

[1] Z. Yu and X. Lu, "Review on vehicle dynamics control of distributed drive electric vehicle," *Journal of Mechanical Engineering*, vol. 49, no. 8, pp. 105–114, 2013.

[2] L. Hai-mei, N. Zhang, B. Shao-yi, F. Jun-ping, and J.-b. Zhao, "Dynamics and switching control of hybrid power steering system of distributed drive electric vehicle," *Science and Technology and Engineering*, vol. 16, no. 17, pp. 283–291, 2016.
[3] S.-I. Sakai, H. Sado, and Y. Hori, "Dynamic driving/braking force distribution in electric vehicles with independently driven four wheels," *Electrical Engineering in Japan*, vol. 138, no. 1, pp. 79–89, 2002.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3213052, 10 pages
https://doi.org/10.1155/2020/3213052

Research Article

Rotor Temperature Safety Prediction Method of PMSM for Electric Vehicle on Real-Time Energy Equivalence

Anjian Zhou,[1,2] **Changhong Du**,[2] **Zhiyuan Peng**,[2] **Qianlei Peng**,[2] **and Datong Qin**[1]

[1]*State Key Laboratory of Mechanical Transmission & School of Automotive Engineering, Chongqing University, Chongqing 400044, China*
[2]*Chongqing Changan New Energy Automobile Technology Co., Ltd., Chongqing 401133, China*

Correspondence should be addressed to Anjian Zhou; 20183201027g@cqu.edu.cn

Received 19 June 2020; Accepted 21 August 2020; Published 14 October 2020

Guest Editor: Esam Hafez Abdelhameed

The load capacity of the permanent magnet synchronous motor is limited by the rotor temperature, and the excessive temperature of the rotor will bring potential thermal safety problems of the system. Therefore, the accurate prediction of the rotor temperature of the permanent magnet synchronous motor for the electric vehicle is crucial to improve the motor performance and system operation safety. This paper studied the heating mechanism and the energy flow path of the motor and built the heat energy conversion model of the stator and rotor. The real-time algorithm to predict the rotor temperature was constructed based on the dissipative energy conservation of the stator of the motor rotor temperature. And the prediction method of the initial rotor temperature is fitted using the experimental results when the system is powered on. Finally, the test platform was set up to validate the rotor temperature accuracy. The results show that the motor rotor temperature estimation error under the dynamic operating condition is within ±5. The research provides a solution to improve the performance and thermal safety of the permanent magnet synchronous motor for electric vehicles.

1. Introduction

The permanent magnet synchronous motor (PMSM) is widely used in pure electric vehicles due to high-power density, high efficiency, and high torque. Thermal safety and peak performance are the difficulties of PMSM development. The copper and the iron losses are the main sources for the temperature rise of rotor magnetic steel, and the temperature of rotor magnetic steel directly determines the duration of the peak power of the motor. Therefore, the research of the rotor temperature prediction can not only ensure the thermal safety of the motor but also improve the peak performance of the motor. Meanwhile, the coercive force of magnetic steel is closely related to temperature, which decreases with the rise of temperature. When the temperature of the rotor magnetic steel exceeds the limit value, the irreversible demagnetization will happen. In general, the irreversible demagnetization should be avoided under the operating condition of the motor [1–3]. In fact, the torque capacity of PMSM is usually lower than its actual torque capacity to avoid overheating failure of the motor without the high-precision rotor temperature prediction [4, 5].

It is difficult to obtain the temperature of rotor magnetic steel by direct measurement when the motor is running [6]. The rotor temperature measurement methods include sliding ring and wireless temperature sensor. But these two methods have high cost and low engineering feasibility, so they cannot be applied in batch. Compared with the direct temperature measurement with an integrated sensor to the rotor, a mature rotor temperature algorithm has advantages in development cost and fast response of thermal protection [7, 8]. But the real-time rotor temperature prediction technology faces some challenges, such as thermal model complexity, algorithm safety, and temperature prediction accuracy [9]. In the current research, the rotor temperature prediction methods mainly include three directions. The first method predicts the rotor temperature with the empirical formula by the indirect variables [10]. The second method is to subdivide the motor

into elements, establish the thermal resistance between elements, and form the thermal network model [11]. The third method is to measure the counterelectromotive force of the motor and calculate the residual flux density of the motor [12]. The actual rotor temperature is obtained by querying the corresponding relationship between the residual flux density and the rotor temperature [13, 14].

Nevertheless, there are some shortcomings in these research methods. Firstly, the temperature change of the rotor of the motor under natural cooling condition is not taken into account after the whole vehicle system is powered down. As a result, the initial temperature of the rotor cannot be assigned to calculate when the system is powered up again. Secondly, it only predicts the rotor temperature under normal temperature conditions without considering the influence of ambient temperature on the rotor temperature characteristics, resulting in poor adaptability and limited accuracy of the algorithm [15]. In addition, when measuring the rotor temperature with the counterelectromotive force method, the motor current should be unloaded. It is not practical to predict the real working condition of the vehicle. This paper will comprehensively consider the thermal nodes that affect the rotor temperature of the motor, and obtain the law of the rotor temperature characteristics of the motor through the test method. The rotor temperature algorithm is built under different environmental temperatures and load conditions, so as to improve the performance and operation safety of the motor system [16, 17].

2. Main Problem

To predict the temperature of the rotor accurately, the mechanism of heat generation and conduction for the motor should be researched. Considering the complexity of the thermal characteristics on the actual motor work condition, the energy transfer paths inside the motor system were simplified as shown in Figure 1.

The temperature rise of the rotor is affected by the copper loss P_{cu}, the iron loss P_{iron}, the mechanical loss P_{mech}, and the coolant dissipation P_w [18–20]. The loss exchanges with the environment in the form of heat to attain the thermodynamic equilibrium. Meanwhile, the stator generates loss or heat when the three-phase current reacts on the stator. As a source, the stator would heat on the rotor with a power of P_r, dissipate to the air with a power of P_{s-air}, and dissipate to the coolant with a power of P_w. Also, the rotor would dissipate to the air with a power of P_{r-air}.

As a main heat source of the stator, the copper loss is caused by three-phase current passing through the stator winding cross section. To eliminate the irregularity of stator current in the winding, the current in the stator winding section is simplified and equivalent to uniform distribution. The copper loss is estimated by the following formula [21, 22]:

$$P_{Cu} = nI_{\text{phase}}^2 R_{\text{phase}},$$
$$R_{\text{phase}} = R_{20}\frac{(235 + T_{\text{en}})}{(235 + 20)}, \tag{1}$$

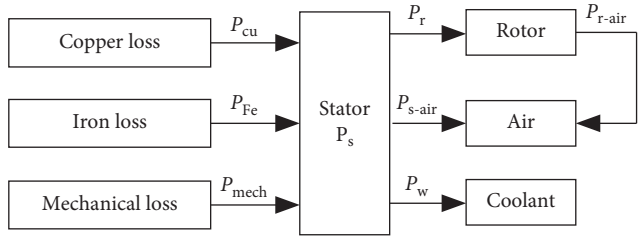


Figure 1: The energy transfer paths of the motor system.

where n is the phase number of the motor, I_{phase} is the phase current, R_{phase} is the phase resistance, R_{20} is the resistance of the stator winding at an ambient temperature of 20, and T_{en} is the ambient temperature.

The iron loss includes the hysteresis loss and the eddy current loss. The hysteresis loss is caused by the change of alternating magnetic field caused by the alternating current in the stator winding. The eddy current loss is caused by the induced current as the magnetic field changes in the core. The iron loss is calculated by the following formula [23, 24]:

$$P_{\text{iron}} = k_h f B_m^2 + k_e f^2 B_m^2, \tag{2}$$

where k_h is the hysteresis loss coefficient, k_e is the eddy current loss coefficient, f is the armature field alternating frequency, and B_m is the amplitude of flux density of the stator core.

The mechanical loss consists of the bearing friction loss and the windage loss. The mechanical loss is calculated by the following formula [25, 26]:

$$P_{\text{mech}} = k_c C_f \pi \rho_{\text{air}} \omega_m^3 r^4 l, \tag{3}$$

where k_c is the coefficient of surface roughness, C_f is the friction coefficient, π is the constant parameter of Pi, ρ_{air} is the density of air, ω_m is the angular velocity of the rotor, l is the length of the rotor, and r is the radius of the rotor.

As the motor works, most of the heat is taken away by the coolant and the rest is carried away by air. The heat dissipated by the coolant is estimated by the following formula [27, 28]:

$$P_w = \rho_w C_w A_w v \frac{(T_{\text{in}} - T_{\text{out}})}{(t_2 - t_1)}, \tag{4}$$

where ρ_w is the density of the coolant, C_w is the specific heat of the coolant, A_w is the section area of the cooling pipe, v is the flow velocity of the coolant, T_{in} and T_{out} are the temperatures of the coolant at the inlet and the outlet, and t_1 and t_2 are the beginning and the ending time.

The heat carried away by air is evaluated by the following formula [29].

Its Newton function is given by

$$P_{s-\text{air}} = \delta A_s \frac{(T_s - T_{\text{en}})}{(t_2 - t_1)},$$
$$P_{r-\text{air}} = \alpha A_r \frac{(T_r - T_{\text{en}})}{(t_2 - t_1)}, \tag{5}$$

where δ and α are the coefficients of convection heat transfer for the stator and the rotor [30], A_s is the area of convection heat transfer between the stator surface and the air, A_r is the area of convection heat transfer between the rotor surface and the air, and T_s and T_r are the temperatures at the surface of the stator and the rotor:

$$\begin{cases} \delta = 9.73 + 14V_s^{0.62}, \\ \alpha = 9.73 + 14V_r^{0.62}, \end{cases} \tag{6}$$

where V_s and V_r are the air velocity of the cooling surface for the stator and the rotor, respectively.

The estimation accuracy of rotor temperature is greatly influenced by factors of motor operation condition and environment temperature. In order to obtain high accuracy rotor temperature, equivalent and accurate modeling solutions will be used to build a rotor temperature model based on the running state and stop state, respectively.

It is clear that the heating power of the stator mainly consists of three parts including copper loss, iron loss, and mechanical loss according to the energy flow analysis during motor running state from Figure 1. The conservation of energy can be expressed as follows.

From this, the decision function corresponding to the segmentation hyperplane equation can be solved, which is given by

$$P_s = P_{Cu} + P_{\text{iron}} + P_{\text{mech}}. \tag{7}$$

The absorbing energy of the stator will change its temperature during the period of time, so it is concluded as follows:

$$P_s = C_s M_s \frac{(T_{s2} - T_{s1})}{(t_2 - t_1)}, \tag{8}$$

where C_s is the specific heat of the stator, M_s is the mass of the stator, T_{s1} and T_{s2} are the stator temperatures at interval time points of sample period, respectively, and t_1 and t_2 are interval time points of sample periods, respectively.

Meanwhile, the stator will bring heat energy to cooling water, atmosphere, and rotor as a heating energy resource. Therefore, the absorbing heat power of the rotor can be concluded based on the conservation of energy as follows:

$$P_s = P_r + P_w + P_{s-\text{air}} + P_{r-\text{air}}. \tag{9}$$

The absorbing energy of the rotor will change its temperature during the period of time, so it is concluded as follows:

$$P_r = C_r M_r \frac{(T_{r2} - T_{r1})}{(t_2 - t_1)}, \tag{10}$$

where C_r is the specific heat of the rotor, M_r is the mass of the rotor, and T_{r1} and T_{r2} are the rotor temperatures at interval time points of the sample period, respectively.

3. Method

Due to the rotor temperature variation during motor running state, updating of the rotor temperature can be attained by putting the previous rotor temperature into formula (10) and adopting a real-time iterative algorithm per operation period. Therefore, combine formulas (8) and (9) to build an estimation model of rotor temperature as follows:

$$T_{r-\text{act}} = T_{r2} = T_{r1} + \Delta T_r,$$

$$\Delta T_r = \frac{C_s M_s (T_{s2} - T_{s1}) - (P_w + P_{s-\text{air}} + P_{r-\text{air}})(t_2 - t_1)}{C_r M_r}, \tag{11}$$

where $T_{r-\text{act}}$ is the real-time rotor temperature from the estimation model.

When the motor comes into a stop state, the heat energy of the rotor brings to the atmosphere and its temperature goes down to the environment temperature along stop time. Therefore, the rotor temperature model under the motor stop state can be attained by obtaining the relationship between rotor temperature and stop time.

When the vehicle is powered on, initial rotor temperature can be attained by adopting the previous rotor temperature $T_{r-\text{pre}}$, the stop time t_{stop}, and the environment temperature T_{en} as the following steps:

(1) Obtain the value of $T_{r-\text{pre}}$ and t_{stop} which are saved in electrically erasable programmable read-only memory (EEPROM) of the motor controller last time.

(2) Receive the environment temperature T_{en} from the air-conditioning controller, and initialization time point t_0 can be captured by looking up the table of rotor temperature and the stop time from Figure 2 at different environment temperatures T_1 and T_2.

(3) According to t_{stop} and t_0 from step (1) and step (2), initial rotor temperature $T_{r-\text{init}(T1)}$ and $T_{r-\text{init}(T2)}$ of different environment temperatures T_1 and T_2 can be obtained by looking up the table of rotor temperature and stop time from Figure 2 at the time point of $t_0 + t_{\text{stop}}$.

(4) Based on output results from step (3), rotor initial temperature at different environment temperatures T_{en} can be matched as follows:

$$T_{r-\text{init}(T)} = \xi T_{r-\text{init}(T1)} + (1 - \xi) T_{r-\text{init}(T2)}, \tag{12}$$

where ξ is the matching coefficient depending on environment temperature as shown in Table 1.

A real-time control algorithm is constructed to estimate the rotor temperature at different environment temperatures and operation states based on the rotor temperature model. The algorithm process and software frame are introduced in Figure 3.

Firstly, it is necessary to make sure whether the system is powered on or not, and then judge motor operation state by actual motor torque and speed. When the motor comes into stop state (Flg = 0), initial rotor temperature is attained by the estimation model in the stop state. Next step, when motor torque and speed are checked by controller, real-time

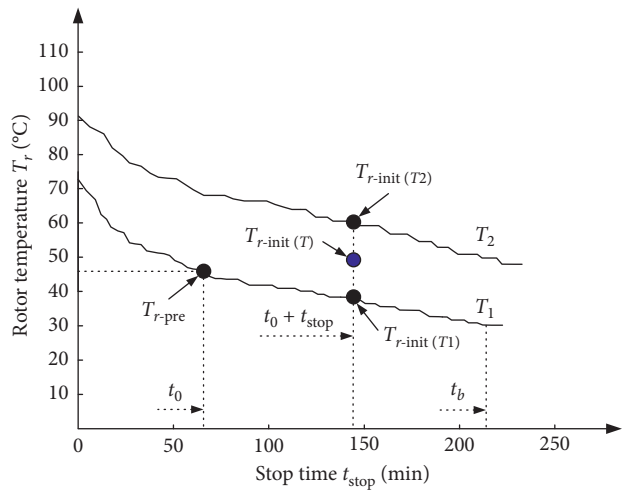


Figure 2: The relationship between rotor temperature and stop time.

Table 1: The matching coefficient for the rotor initial temperature.

T_{en}	0°C–10°C	10°C–20°C	20°C–30°C	30°C–40°C	40°C–50°C	50°C–60°C	60°C–70°C
ξ	0.1	0.3	0.3	0.5	0.5	0.1	0.1

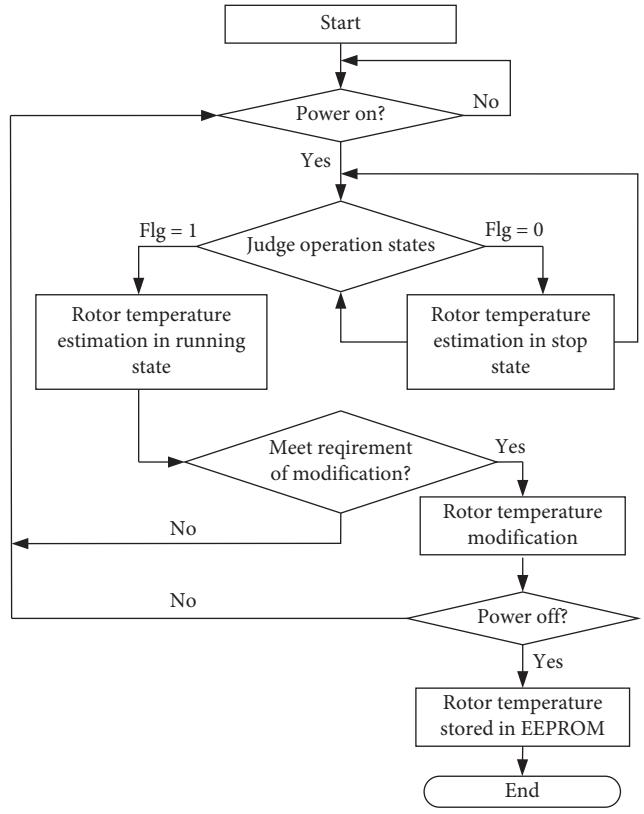


Figure 3: The algorithm process of rotor temperature.

rotor temperature is calculated by the estimation model in running state. Rotor temperature needs to be modified during the motor running state if it meets the requirement of the modification strategy. Finally, when it comes into power off for the system, real-time rotor temperature is stored into the memorizer of EEPROM and the previous rotor temperature can be used again in next power on for the system.

There is an accumulative error to adopt a real-time iterative algorithm to estimate rotor temperature. Therefore, it

is necessary to build a modification strategy to improve the estimation accuracy of rotor temperature as follows:

$$\begin{aligned} n_{\text{cal}-1} \le \left|n_{\text{mot}}\right| \le n_{\text{cal}-2}, \\ \left|T_{\text{mot}-\text{trq}}\right| \le T_{\text{cal}}, \\ \Delta\psi_{\text{mot}} \le \Delta\psi_{\text{cal}}, \end{aligned} \tag{13}$$

where n_{mot} is the actual motor speed, $n_{\text{cal}-1}$ and $n_{\text{cal}-2}$ are low- and high-level limitation of motor speed, respectively, $T_{\text{mot}-\text{trq}}$ is the actual motor torque, T_{cal} is the motor torque limitation, $\Delta\psi_{\text{mot}}$ is the changing rate of motor actual flux linkage, and $\Delta\psi_{\text{cal}}$ is the changing rate limitation of motor flux linkage.

It is necessary to limit the changing rate to eliminate the prominent variation between estimation value and modification value and avoid power break-off or torque cut-down in a short time. Therefore, the updated value of rotor temperature in a running period should be modified based on the following equation:

$$T_{\text{var}} = T_{r-\text{act}} + k\left(T_{\text{tab}} - T_{r-\text{act}}\right), \tag{14}$$

where T_{tab} is the rotor temperature from looking up the flux linkage table for motor and k is the changing rate for rotor temperature modification.

The numerical model of the rotor temperature algorithm is built by relevant experiments and embedded in the software of the motor control system to satisfy the practicability of the rotor temperature algorithm.

According to the algorithm established above, the prediction of the rotor temperature will produce accumulated errors with the extension of calculation time. In order to ensure the accuracy of rotor temperature prediction, equation (14) is used to correct the rotor temperature.

Firstly, the counterelectromotive force of the motor corresponding to each rotor temperature was obtained through experiments. And the motor flux was calculated by using the following formula:

$$\psi_{\text{mot}} = \frac{E_{\varphi}}{\omega_{\text{mot}}} = \frac{7.8E_{\text{mot}}}{\eta_{\text{mot}} p_{\text{mot}}}, \tag{15}$$

where E_{φ} is the maximum phase electromotive force, ω_{mot} is the electrical angular frequency, E_{mot} is the phase electromotive force, and p_{mot} is the pole pairs of the motor.

As a result, the numerical model of the motor flux and rotor temperature was built, as shown in Figure 4.

The total cooling dissipation of the motor includes three parts: cooling water dissipated power and stator and rotor dissipated power to air. The relationship between the total cooling dissipation of the system and the stator temperature change rate is established as shown in Figure 5. The calibration and optimization of the numerical model were carried out by experiments under different load conditions to improve the prediction accuracy of the rotor temperature model.

In order to obtain the initial rotor temperature when the motor system is powered on, the numerical model is established corresponding to the rotor temperature and downtime under the state of natural cooling of the motor system. The

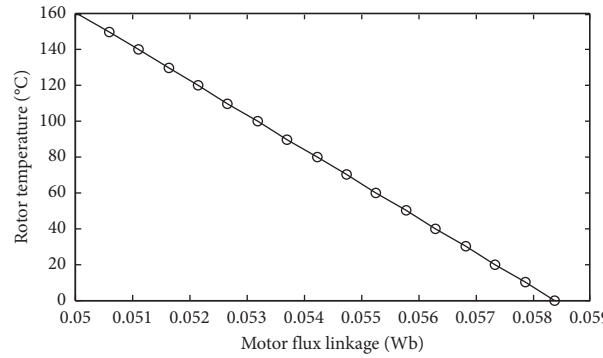


Figure 4: The relationship between the rotor temperature and the motor flux linkage.

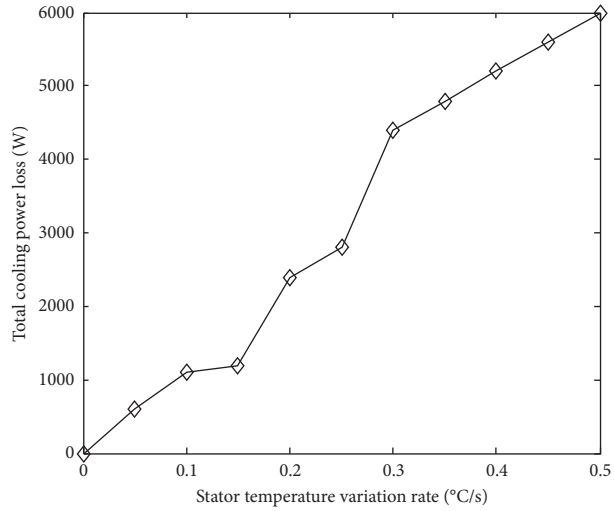


Figure 5: The numerical model of the total cooling dissipation.

motor was run at high power until the rotor temperature rose to the equilibrium point at each ambient temperature under the conditions of 10°C, 30°C, 50°C, and 70°C, respectively. And the expression of rotor temperature and shutdown time was fitted by the polynomial, as shown in Figures 6–9.

4. Results

To validate the real-time control algorithm of the rotor temperature prediction model established in this paper, an AVL test system is used to build a motor rotor temperature accuracy experimental platform, as shown in Figure 10. A slip ring is used to draw out the thermal resistance for the rotor temperature, as shown in Figure 11. The experimental platform consists of power dynamometer, battery simulator, temperature box, cooling system, motor system, electrical parameter tester, adjustable low-voltage power supply, and related sensors.

To ensure the prediction accuracy of the algorithm in the different environment temperatures under the condition of changing load, the motor was running under different loads in the ambient temperature 30°C and 70°C, respectively. As shown in Figures 12–14, the maximum error between the prediction value and the experimental results under fixed load conditions is within ±6°C.

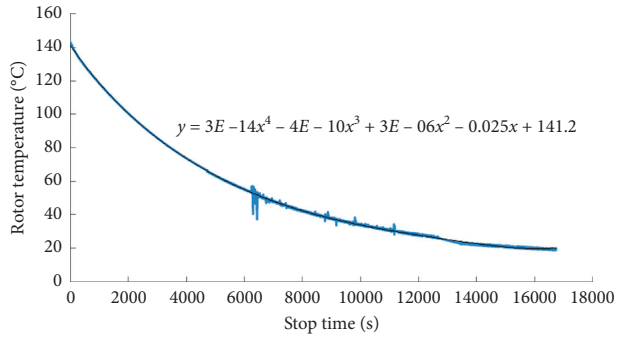


Figure 6: The rotor temperature curve at environment temperature 10°C.

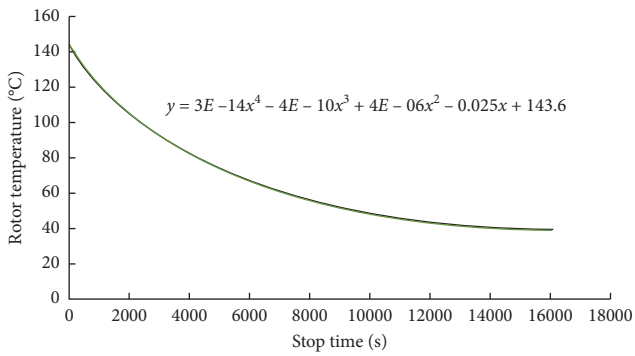


Figure 7: The rotor temperature curve at environment temperature 30°C.

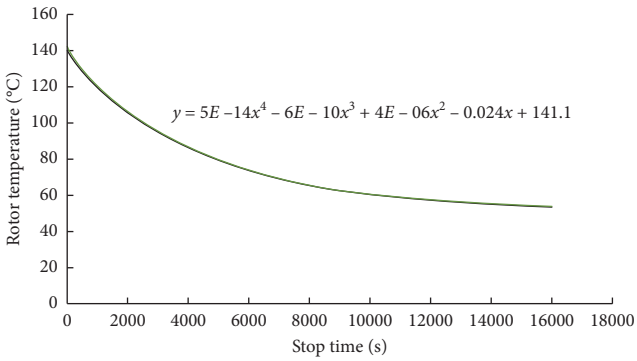


Figure 8: The rotor temperature curve at environment temperature 50°C.

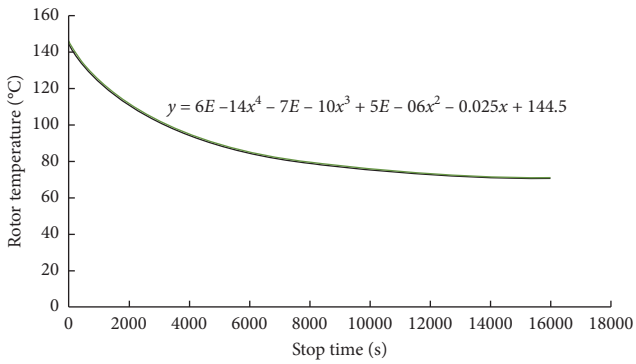


Figure 9: The rotor temperature curve at environment temperature 70°C.

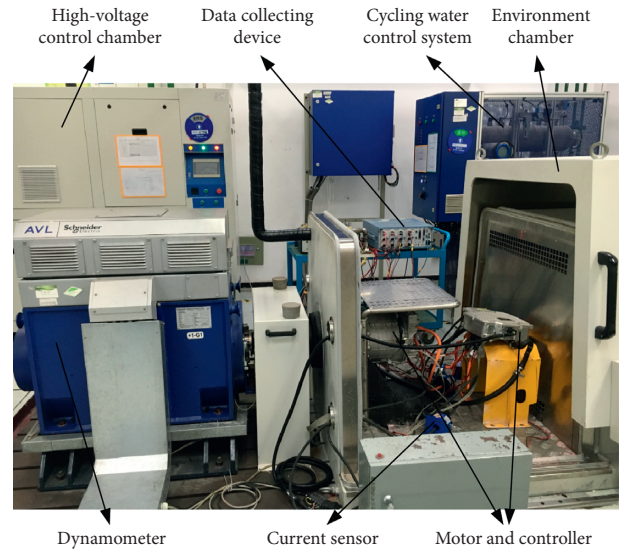


Figure 10: The motor rotor temperature accuracy experimental platform.

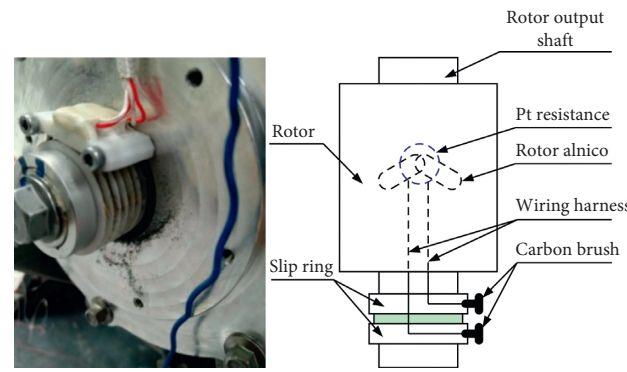


Figure 11: The schematic of the rotor temperature test using the slip ring.

To validate the proposed algorithm accuracy rotor temperature under the changing load condition, the motor was running with the vehicle real variable load in the environment temperatures of 30°C and 70°C, respectively. The comparison indicates that the maximum dynamic error is within ±5°C between the predicted and the measured values, as shown in Figures 15 and 16.

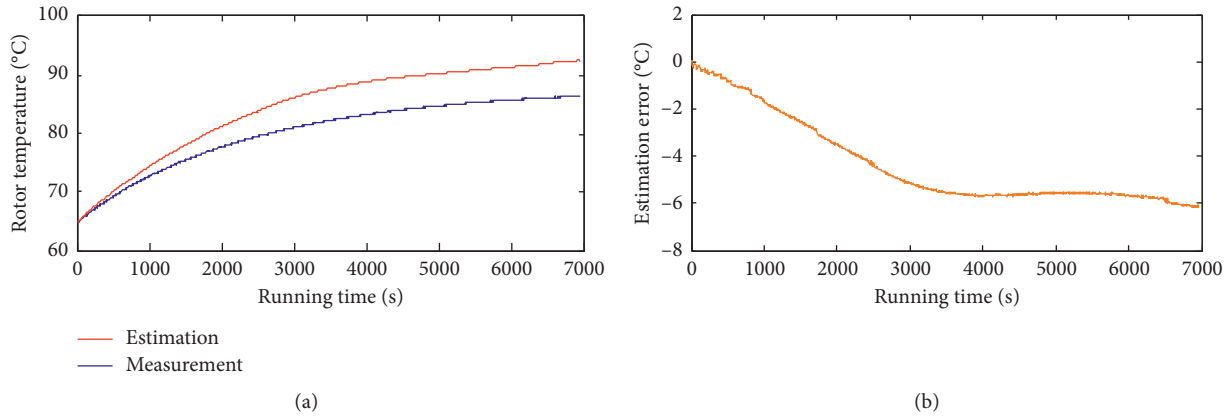


Figure 12: The relationship between the rotor temperature and the motor flux linkage.

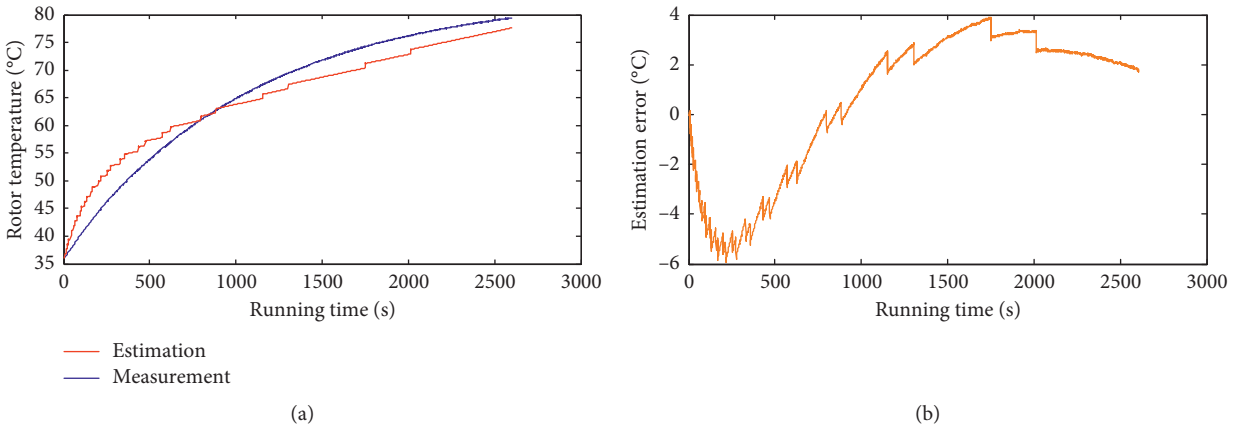


Figure 13: Test results at constant load condition (5000 rpm/68 Nm and environment temperature 30°C).

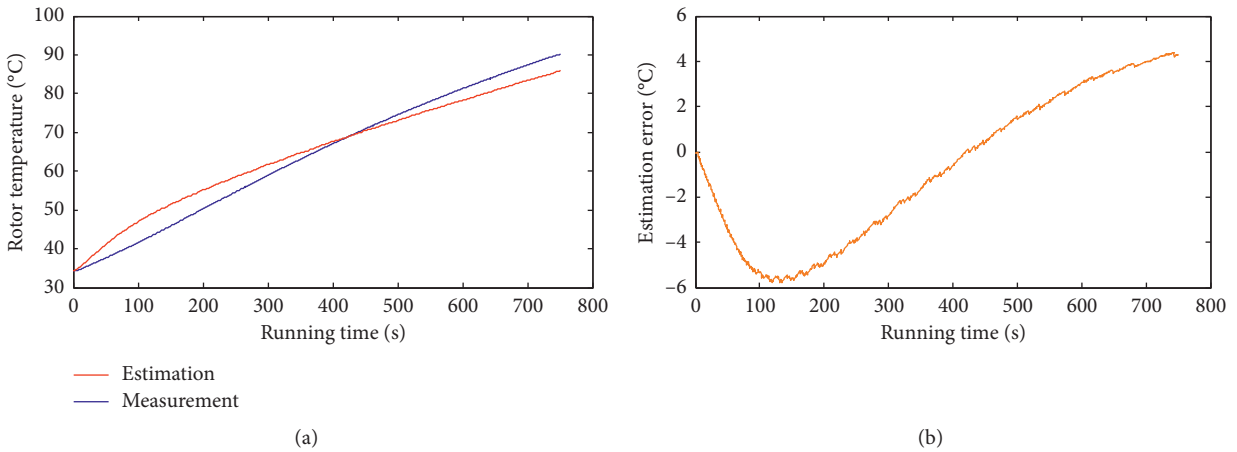


Figure 14: Test results at constant load condition (4000 rpm/193 Nm and environment temperature 30°C).

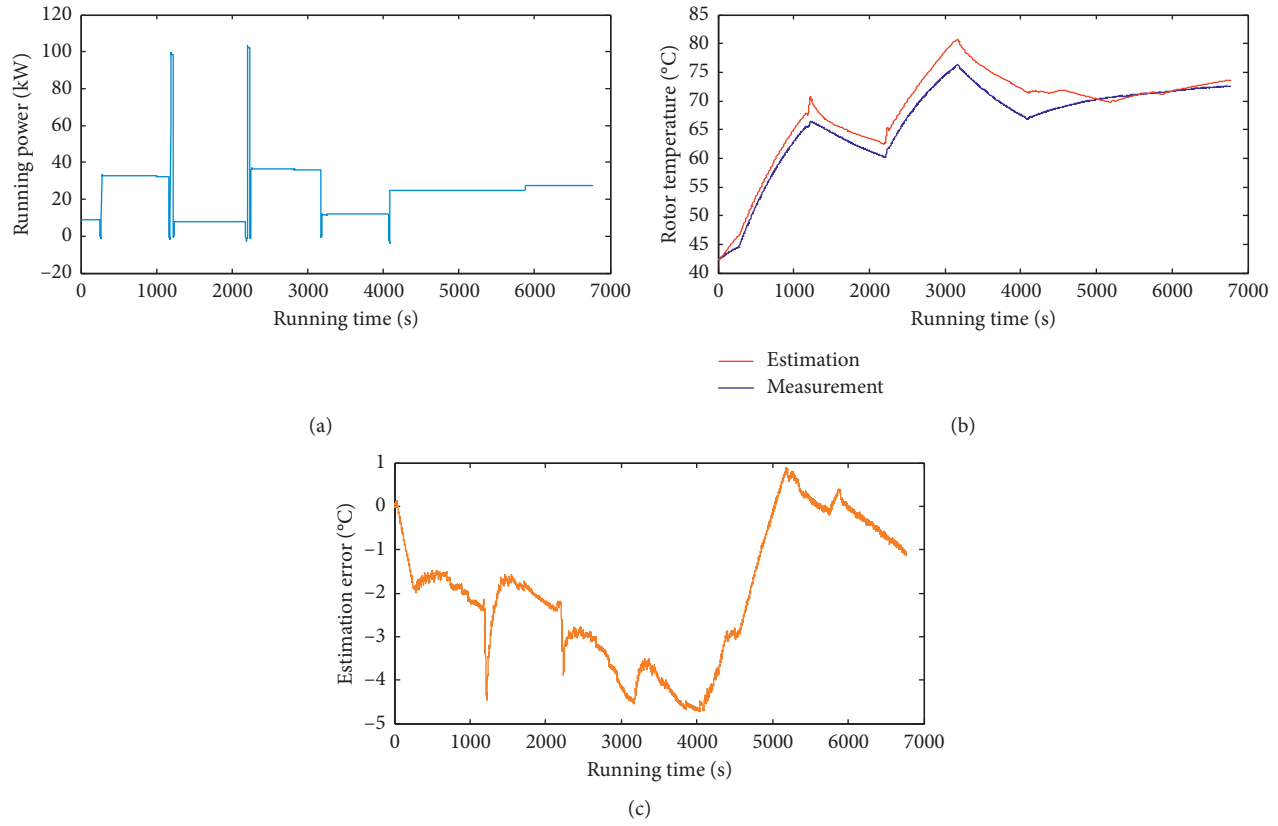


Figure 15: Test results at variational load condition (environment temperature 30°C).

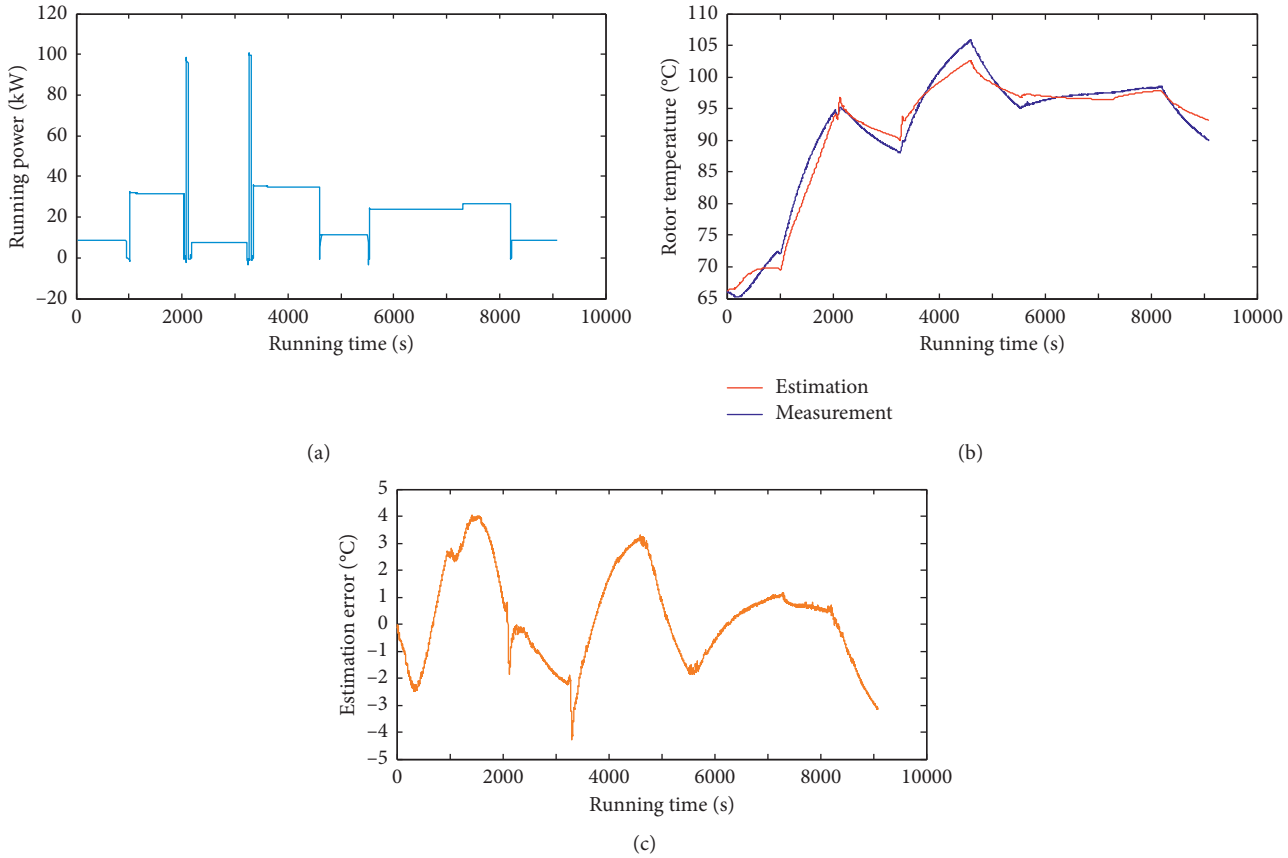


Figure 16: Test results at variational load condition (environment temperature 70°C).

5. Conclusion

A method to estimate the rotor temperature of the permanent magnet synchronous motor in this paper has been proposed. The method is characterized with the equivalent thermal model of rotor temperature estimation by analyzing the principle of heat generation and heat transferring path inside the motor system during operation based on the conservation of energy for the stator heat consumption and establishing a numerical model of rotor temperature estimation by experiment. Different constant load power is adopted to motor real operation states at different environment temperatures and the numerical model of total cooling power is optimized by comparing rotor temperature errors between real test and model calculation to improve the estimation accuracy of the rotor temperature model. Then, the model estimation accuracy of rotor temperature is validated at different environment temperatures and variational load power, and the test result shows that dynamic estimation accuracy between measurement and estimation is within ±5°C. According to the high accuracy estimation of rotor temperature in this research, duration operation time of motor peak power can be significantly expanded because the protection threshold of rotor temperature is increased to improve the peak performance of motor in electric vehicle.

Data Availability

All data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Key R&D Plan Program (2018YFB0106101).

References

[1] I. Boldea, L. N. Tutelea, L. Parsa, and D. Dorrell, "Automotive electric propulsion systems with reduced or no permanent magnets: an overview," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 10, pp. 5696–5711, 2014.

[2] H. Liang, Y. Chen, S. Liang, and C. Wang, "Fault detection of stator inter-turn short-circuit in PMSM on stator current and vibration signal," *Applied Sciences*, vol. 8, no. 9, p. 1677, 2018.

[3] D. Huang, W. Li, Y. Wang, and Z. Cao, "Influence of magnetic slot wedge on rotor losses and temperature field of PMSM," *Electric Machines and Control.Magnetics*, vol. 20, pp. 60–66, 2016.

[4] Y. Chen, S. Liang, W. Li, H. Liang, and C. Wang, "Faults and diagnosis methods of permanent magnet synchronous motors: a review," *Applied Sciences*, vol. 9, no. 10, p. 2116, 2019.

[5] W. Li, Y. Chen, X. Li, and S. Liang, "Matching quality detection system of synchronizer ring and cone," *Applied Sciences*, vol. 9, no. 17, p. 3622, 2019.

[6] J. Dong, Y. Huang, L. Jin et al., "Thermal optimization of high-speed permanent motor," *IEEE Transactions on Magnetics*, vol. 50, Article ID 7018504, 2014.

[7] K.-S. Kim, B.-H. Lee, and H.-J. Kim, "Thermal analysis of outer rotor type IPMSM using thermal equivalent circuit," in *Proceedings of the 15th International Conference on Electrical Machines and Systems*, pp. 1–4, Sapporo, Japan, October 2012.

[8] Y. Liu, J. Li, Z. Chen, D. Qin, and Y. Zhang, "Research on a multi-objective hierarchical prediction energy management strategy for range extended fuel cell vehicles," *Journal of Power Sources*, vol. 429, pp. 55–66, 2019.

[9] N. Rostami, M. R. Feyzi, J. Pyrhonen, A. Parviainen, and M. Niemela, "Lumped-parameter thermal model for axial flux permanent magnet machines," *IEEE Transactions on Magnetics*, vol. 49, no. 3, pp. 1178–1184, 2013.

[10] J. D. McFarl and T. M. Jahns, "Investigation of the rotor demagnetization characteristics of interior PM synchronous machines during fault conditions," *IEEE Transactions on Industry Applications*, vol. 50, pp. 2768–2775, 2013.

[11] T. Reichert, T. Nussbaumer, and J. W. Kolar, "Split ratio optimization for high-torque PM motors considering global and local thermal limitations," *IEEE Transactions on Energy Conversion*, vol. 28, no. 3, pp. 493–501, 2013.

[12] K.-C. Kim and D.-S. Ryu, "Torque characteristic with respect to the load angle of a permanent magnet motor," *IEEE Transactions on Magnetics*, vol. 48, no. 11, pp. 4200–4203, 2012.

[13] X. Jannot, J.-C. Vannier, C. Marchand, M. Gabsi, J. Saint-Michel, and D. Sadarnac, "Multiphysic modeling of a high-speed interior permanent-magnet synchronous machine for a multiobjective optimal design," *IEEE Transactions on Energy Conversion*, vol. 26, no. 2, pp. 457–467, 2011.

[14] L. Guangjin, J. Ojeda, E. Hoang et al., "Thermal-electromagnetic analysis for driving cycles of embedded flux-switching permanent-magnet motors vehicular technology," *IEEE Transactions on Magnetics*, vol. 61, pp. 140–151, 2012.

[15] D. Joo, J.-H. Cho, K. Woo, B.-T. Kim, and D.-K. Kim, "Electromagnetic field and thermal linked analysis of interior permanent-magnet synchronous motor for agricultural electric vehicle," *IEEE Transactions on Magnetics*, vol. 47, no. 10, pp. 4242–4245, 2011.

[16] D. H. Lim and S. C. Kim, "Thermal performance of oil spray cooling system for in-wheel motor in electric vehicles," *Applied Thermal Engineering*, vol. 63, no. 2, pp. 577–587, 2014.

[17] T. Deng, Z. Su, J. Li et al., "Advanced angle field weakening control strategy of permanent magnet synchronous motor," *System Control Engineering*, vol. 68, pp. 3425–3435, 2019.

[18] Y. Xie, Z. Wang, X. Shan, and Y. Li, "The calculations and analysis of 3D transient magnetic-thermal-solid coupling for squirrel-cage induction motors based on multi fields," *Proceedings of the CSEE.Magnetics*, vol. 36, pp. 3076–3084, 2016.

[19] X. Sun, Y. Shen, S. Wang, G. Lei, Z. Yang, and S. Han, "Core losses analysis of a novel 16/10 segmented rotor switched reluctance BSG motor for HEVs using nonlinear lumped parameter equivalent circuit model," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 2, pp. 747–757, 2018.

[20] X. Sun, Z. Shi, G. Lei, Y. Guo, and J. Zhu, "Analysis and design optimization of a permanent magnet synchronous motor for a campus patrol electric vehicle," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10535–10544, 2019.

[21] N. Yang, M. Zhang, B. Guo, and C. Xiao, "Analysis of temperature field in ironless stator axial field permanent motor," *Computer Simulation Magnetics*, vol. 32, pp. 259–263, 2015.

[22] L. Chen, W. Zhao, and J. Ji, "Thermal analysis and calculation of fault-tolerant permanent magnet machine by using equivalent thermal network method," *Magnetics*, vol. 43, pp. 45–50, 2016.

[23] C. Li, F.-C. Huang, and Y.-Q. Wang, "An applicable real-time thermal model for temperature prediction of permanent magnet synchronous motor," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 231, no. 1, pp. 43–51, 2017.

[24] H. Toda, K. Senda, S. Morimoto, and T. Hiratani, "Influence of various non-oriented electrical steels on motor efficiency and iron loss in switched reluctance motor," *IEEE Transactions on Magnetics*, vol. 49, no. 7, pp. 3850–3853, 2013.

[25] A. Boglietti, A. Cavagnino, and M. Lazzari, "Fast method for the iron loss prediction in inverter-fed induction motors," *IEEE Transactions on Industry Applications*, vol. 46, no. 2, pp. 806–811, 2010.

[26] J. Nerg, M. Rilla, and J. Pyrhonen, "Thermal analysis of radial-flux electrical machines with a high power density," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 10, pp. 3543–3554, 2008.

[27] C. Li, *Study on the Cooling System for Mini Vehicle Induction Motor.*Harbin Institute of Technology, Harbin, China, 2013.

[28] L. Liu, *The Study of Thermal Characteristics in Various Conditions Cooling System of Permanent Magnet Synchronous Motor in Pure Electric Vehicle*, Hefei University of Technology, Hefei, China, 2015.

[29] H. Liu, L. Yang, and F. Sun, "Study of numerical method determining inner temperature rise of asynchronous motor based on thermographic measurement," *Magnetics*, vol. 27, pp. 496–500, 2006.

[30] W. Li, S. Li, Y. Xie, and S. Ding, "Stator-rotor coupled thermal field numerical calculation of induction motors and correlated factors sensitivity analysis," *Proceedings of the CSEE. Magnetics*, vol. 24, pp. 85–91, 2007.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9718345, 9 pages
https://doi.org/10.1155/2020/9718345

Research Article

Fault Detection of the Wind Turbine Variable Pitch System Based on Large Margin Distribution Machine Optimized by the State Transition Algorithm

Mingzhu Tang,[1] **Jiahao Hu**,[1] **Zijie Kuang**,[1] **Huawei Wu**,[2] **Qi Zhao**,[1] **and Shuhao Peng**[1]

[1]*School of Energy and Power Engineering, Changsha University of Science & Technology, Changsha 410114, China*
[2]*Hubei Key Laboratory of Power System Design and Test for Electrical Vehicle, Hubei University of Arts and Science, Xiangyang 441053, China*

Correspondence should be addressed to Huawei Wu; whw_xy@hbuas.edu.cn

Received 31 July 2020; Revised 31 August 2020; Accepted 8 September 2020; Published 13 October 2020

Academic Editor: Yong Chen

Aiming at solving the problem that the parameters of a fault detection model are difficult to be optimized, the paper proposes the fault detection of the wind turbine variable pitch system based on large margin distribution machine (LDM) which is optimized by the state transition algorithm (STA). By setting the three parameters of the LDM model as a three-dimensional vector which was searched by STA, by using the accuracy of fault detection model as the fitness function of STA, and by adopting the four state transformation operators of STA to carry out global search in the form of point, line, surface, and sphere in the search space, the global optimal parameters of LDM fault detection model are obtained and used to train the model. Compared with the grid search (GS) method, particle swarm optimization (PSO) algorithm, and genetic algorithm (GA), the proposed model method has lower false positive rate (FPR) and false negative rate (FNR) in the fault detection of wind turbine variable pitch system in a real wind farm.

1. Introduction

With the rapid development of the wind power industry, the installed capacity and quantity of wind turbines are growing continuously. The World Wind Energy Association predicts that by the end of 2020, the global installed capacity will reach 1.9×10^6 mw [1]. However, the availability of wind turbines is not ideal due to the increasing failure rate and maintenance cost of wind turbines along with the development of wind farms. The wind turbine variable pitch system is one of the important parts of the wind turbine, which has a complex internal mechanical structure and operated in a harsh environment that will lead to its failure rate significantly higher than other wind turbine subsystems. Since the safe and stable operation of the variable pitch system directly affects the operation efficiency of wind turbines, fault detection of the variable pitch system is of great significance for stable and efficient generation of wind turbines [2, 3].

The fault detection method is generally divided into the model-based method and the data-driven method [4]. The model-based fault detection method needs to establish an accurate mathematical model for the diagnosis object through mathematical and physical knowledge and detect faults by observing the change of the residual value [5]; the residual value of an equipment under normal state should be zero or close to zero, and it is not zero when the equipment is disturbed or malfunctioned. This method can be divided into the parameter estimation method [6], state estimation method [7], and equivalent space method [8]. The model-based fault detection method can quickly get a more accurate mathematical model and detect faults accurately for the system with simple structure. However, for the fault detection of large-scale wind turbines, the modeling process is easy to be affected by various parameters which will influence the robust performance and the accuracy of the fault detection and even makes it difficult to locate the wind turbine internal fault causes.

The data-driven fault detection method extracts useful information through various data processing and analysis methods based on the collected data, compares the collected historical data with the real-time data of the system, and analyzes their potential relationship so as to carry out fault detection. Being capable of detecting the fault of the equipment through data analysis, this method does not need to establish an accurate mathematical model. It does not depend on the complexity and uncertainty of the system, so a good detection performance is obtained of it. As the method meets the needs of the industrial big data era, it is widely used in the industrial field. Artificial neural networks (ANNs), SVM, LDM, and other models are usually used for fault detection of equipment in the data-driven fault detection method.

The ANN is a classic data-driven model based on mimicing the biological nervous system. It can automatically analyze and infer the input information to detect faults by simulating the physiological structure and thinking mode of the human brain. The application of ANNs in wind turbine fault detection has a good detection performance. Concerning the problem of sensor fault of the wind turbine, Qiu et al. proposed a damage prediction method for the offshore wind turbine tower structure based on ANNs, which can improve the accuracy of fault prediction [9]. In the case of gearbox fault, Chen et al. proposed a fault diagnosis method based on wavelet analysis and neural networks to diagnose the wind turbine gearbox and predict the early fault signs and obtained good results [10]. The ANN has the ability of self-study which is similar to the human brain. It has good robustness to the interference and noise of the system. However, due to the nature of the black box, this method is difficult to make a good explanation for specific faults. Moreover, it has high requirements for data in actual use and requires high running cost.

SVM is another classic data-driven model based on global optimization. It has good performance and can solve the problems of multiclassification recognition and regression prediction [11, 12], which has been widely recognized in the field of wind turbine fault research. Hang et al. proposed a wind turbine fault diagnosis method based on a multiclass fuzzy SVM classifier to improve the accuracy of fault diagnosis [13]. Rotating parts in wind turbines are one of the key objects in fault diagnosis of wind turbines. However, in fact, the vibration signals collected from the rotating parts are generally non-Gaussian and nonstationary, and the fault samples are very limited. Liu et al. proposed a wind turbine fault diagnosis method based on a diagonal spectrum and clustering binary tree SVM, which achieved good results [14]. Although having a good performance on simple binary classification problems, SVM is ineffective in dealing with large-scale data problems and sensitive to model parameters and data integrity.

The distribution machine supported by large margin theory can find the distribution model according to the sample distribution characteristics while considering the sample mean value and sample variance. Compared with the former two models, LDM has higher fault detection performance. In the fault detection of wind turbines, Tang et al. proposed a cost-sensitive large margin distribution machine (CLDM) to solve the problems of class imbalance data and misclassification unequal cost of large wind turbine data sets, which has effectively improved the fault detection performance [15].

The data-driven fault detection model has good practicability in actual wind turbine fault detection and fault diagnosis. However, most of these models depend on the selection of parameters, so it is necessary to use the parameter optimization algorithm to quickly and accurately find the global optimal model parameters. The GS, PSO, and GA are most commonly used to optimize the parameters of the fault detection model in wind turbine fault detection. Aguilar et al. proposed a multiobjective particle swarm optimization (MOPSO) algorithm for the electrical fault of variable-speed wind turbines, which improved the stability of wind turbines [16]. Concerning the problem that the traditional threshold setting is difficult to identify the abnormal operation of wind turbines, Zhang et al. put forward a new backpropagation neural network (BPNN) anomaly identification model combined with GA, which provides good performance effect for abnormal identification of wind turbines [17]. Yan et al. optimized the parameters of SVM by the GS method in wind turbine fault detection to improve the diagnostic accuracy [18]. PSO, GA, and GS can achieve approximate global optimal solution for parameter optimization of a simple model, but it is easy to fall into local optimum when used in fault detection of large and complex wind turbines.

The STA is a parameter optimization algorithm with four state transition operators, facing the complex fault detection problem; the global optimal value can be quickly and accurately found by the four state transformation operators alternately, which is suitable for detecting the complex fault of the wind turbine variable pitch system. Because of its strong performance and practicability, the STA has solved many problems in the industry and other fields [19, 20].

It is of great significance to choose a fault detection model with proper performance. However, in the fault detection model based on machine learning, parameter optimization is an important process, and how to select appropriate parameters to enable the detection model to meet the fault detection standard of the wind turbine variable pitch system is the key and difficult problem of all machine learning models. Therefore, an improved LDM model based on the STA is studied with an aim to effectively finding out the optimal model parameters, making it meet the fault characteristics of the variable pitch system, and improving the accuracy of fault detection.

2. Large Margin Distribution Machine

If the traditional machine learning algorithm based on margin theory for optimization is adopted, attention should be paid to find the minimum margin between samples, such as SVM, which can be adopted to find the hyperplane that maximizes the minimum margin between two kinds of samples in the optimization process [21]. However, the method only focuses on the support vectors that only account for a small proportion in a large number of samples, while the rest of the sample information is not considered in the learning process. The above method will lead to the loss of some samples of useful information as well as reduction of

the learning ability of the algorithm for samples; in addition, the learning effect remains to be improved.

The LDM proposed by Zhang and Zhou is used to find the separation hyperplane according to the distribution characteristics of samples under the premise of considering the margin distribution of the whole sample [22]. Compared with the support vector machine which only optimizes the minimum margin, it has stronger generalization performance. Figure 1 shows the different results of the final classification hyperplane due to different margin considerations in the classification process.

In Figure 1, the triangle icon refers to the first type of sample, the square icon refers to the second type of sample, the elliptical dotted line shows the potential distribution of the two types of samples, and the red triangle and red square indicate the distribution mean of the two types of samples. If the classification hyperplane is searched based on the minimum margin between the two types of samples as $h_{\min}$ in the figure, it can be found to intersect with the potential distribution range of the right sample, and there is the possibility of misclassification; if the overall distribution of two types of samples has been considered in the classification hyperplane with the classification plane as h_{dist} in the figure, the conclusion can be drawn that it has better classification performance and stronger robustness for two types of samples [23].

For LDM, we should set $X = [\phi(x_1), \ldots, \phi(x_m)]$ and set the algorithm $y = [y_1, \ldots, y_m]$ as the m-dimensional column vector, where $\phi(\cdot)$ is the feature mapping through a positive definite function $\kappa(\cdot, \cdot)$, Y is an m-order square matrix, and the diagonal is $y_1, \ldots, y_m$; therefore, the mean value of the margin can be defined as follows:

$$\gamma_m = \frac{1}{m}\sum_{i=1}^{m} y_i \omega^{\mathrm{T}} \phi(x_i) = \frac{1}{m}(Xy)^{\mathrm{T}}\omega. \tag{1}$$

The margin variance is as follows:

$$\gamma_v = \frac{1}{m}\sum_{i=1}^{m}\left(y_i \omega^{\mathrm{T}}\phi(x_i) - \gamma_m\right)^2 = \omega^{\mathrm{T}} X \frac{mI - yy^{\mathrm{T}}}{m^2} X^{\mathrm{T}}\omega. \tag{2}$$

It is important to make a linear combination of margin mean and margin variance into an optimization problem, introduce L2-norm as the regularization term, and select hinge loss for the loss function; therefore, the formalization of LDM is as follows:

$$\begin{aligned} \min_{\omega,\xi_i} \quad & \frac{1}{2}\left\|\omega^2\right\| + \lambda_1\gamma_v - \lambda_2\gamma_m + \frac{C}{m}\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i\omega^{\mathrm{T}}\phi(x_i) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ \forall i \in [m], \end{aligned} \tag{3}$$

where parameters λ_1 and λ_2 are trade-off parameters and are used to adjust the weight of the margin mean and margin variance in the objective function, while C is a loss function parameter. Although the theory of large margin distribution has achieved good results in theory and practice, the classification surface may show unbalanced tendency in the face of the number of unbalanced margins and samples with noise, and the robustness to noise is not strong. Therefore, the model needs further development and improvement.

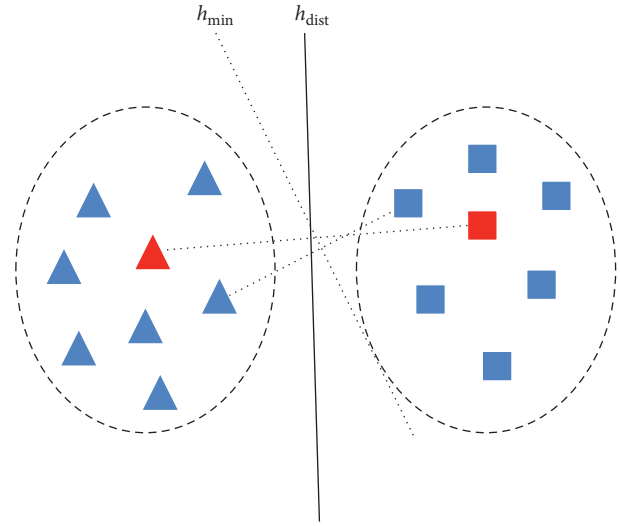


Figure 1: Minimum margin hyperplane and margin distribution hyperplane.

3. The State Transition Algorithm

Being a global optimization method proposed by Zhou et al. [24] the STA is an individual-based intelligent stochastic global optimization method. It uses different state transformation operators to operate independently through the given current solution, thus generating the candidate solution set and finding out the solution better than the current candidate solution in the candidate solution set, which serves as a new solution of the update iteration. The process should be repeated till the certain termination condition is met.

In brief, the STA is based on the state space of modern control theory, which treats the solving process of the optimization problem as the process of state transition and treats the generation and update of the solution as the formation and update of the state.

3.1. The State Transformation Operator. The state-space expression in modern control theory is used as the unified framework of the candidate set, and the state transformation operators are designed for the framework. The unified framework of candidate solutions for the STA is as follows:

$$\begin{cases} x_{k+1} = A_k x_k + B_k u_k, \\ y_{k+1} = f(x_{k+1}), \end{cases} \tag{4}$$

where $x_k = [x_1, x_2, \ldots, x_n]^T$ is the current state and represents a candidate solution in the optimization problem, A_k

and B_k are the state transition matrices, which are random matrices and equivalent to state transformation operators, u_k is a function of the historical state and current state, which is equivalent to a control variable, and $f(\cdot)$ is the objective function, that is, the fitness function.

The four state transition operators in the STA correspond to four search functions, and each state transformation operator can form a regular geometric neighborhood with unique shape and adjustable size. State transformation operators mainly include rotation transformation operator, translation transformation operator, expansion transformation operator, and axesion transformation operator.

(1) The rotation transformation operator:

$$x_{k+1} = x_k + \alpha \frac{1}{n\|x_k\|_2} R_r x_k, \tag{5}$$

where $\alpha > 0$ is the rotation factor; $R_r \in R^{n\times n}$ is a random matrix with its element values evenly distributed between $[-1, 1]$; $\|\cdot\|_2$ is the vector L2-norm, and the function of the rotation transformation operator is to search in the hypersphere with α as the radius.

(2) The translation transformation operator:

$$x_{k+1} = x_k + \beta R_t \frac{x_k - x_{k-1}}{\|x_k - x_{k-1}\|_2}, \tag{6}$$

where $\beta > 0$ is the translation factor; the value range of $R_t \in R$ is $[0, 1]$, meeting the uniform distribution. As a heuristic search operator, the translation transformation operator can search with β as the maximum length from point x_{k-1} to point x_k along the line.

(3) The expansion transformation operator:

$$x_{k+1} = x_k + \gamma R_e x_k, \tag{7}$$

where $\gamma > 0$ is the expansion factor and $R_e \in R^{n\times n}$ is a diagonal matrix, with its element value of nonzero, complying with the Gaussian distribution. As the global search operator, the expansion transformation operator can expand each element in x_k to the whole range of $(-\infty, +\infty)$, thus realizing the search of the whole space.

(4) The axesion transformation operator:

$$x_{k+1} = x_k + \delta R_a x_k, \tag{8}$$

where $\delta > 0$ is the axesion factor and $R_a \in R^{n\times n}$ is a sparse random diagonal matrix, with its element value of nonzero, complying with the Gaussian distribution. Being a heuristic search operator with relatively strong single-dimensional search ability, the axesion transformation operator can search along the axesion axis direction.

4. Large Margin Distribution Machine Optimized by the State Transition Algorithm

Fitness function is a main factor affecting the convergence speed and finding the optimal solution of the parameter optimization algorithm, and it is an evaluation criterion to select and update the optimal solution in the process of parameter optimization. In the STA, the mean accuracy of the LDM optimization model which was verified by 10-fold cross-validation is used as the fitness function to judge the selection and update of the current parameter state; if the accuracy is higher than that of the current optimal state, the new parameter will be used as a better solution to update the current state, and if the accuracy is lower than that of the current optimal state, the parameter will be abandoned for the next iteration. The fitness function is as follows:

$$\text{fitness} = \frac{\sum_{i=0}^{k_{cv}} \text{accuracy}\left(\text{LDM}\left(\lambda_1, \lambda_2, C\right)\right)}{k_{cv}}, \tag{9}$$

where $k_{cv} = 10$ is the number of cross-validations, λ_1, λ_2, and C are the three parameters in LDM, which are the margin variance parameter, margin mean parameter, and loss function parameter, respectively. The meaning and value range of three parameters are shown in Table 1.

LDM parameters are adjusted by the STA, and the three parameters in LDM are taken as a three-dimensional vector form, a state in the STA. The new candidate solution set is generated by alternately using the four transformation operators of rotation, expansion, axesion, and translation.

The use of the fitness function of the improved LDM and the selection and updated pseudocode of the current optimal state solution are given in Algorithm 1.

However, $\text{Best}_0(\lambda_{10}, \lambda_{20}, C)$ refers to the initial state, and the three parameters of LDM are assigned from Step 6 to Step 8; the training set is adopted to train the adjusted LDM algorithm to establish the learning model in Step 9; the testing set is used to predict the model in Step 10; the classification accuracy of the predicted results is used as the evaluation criterion of fitness function in Step 11; the rotation transformation, expansion transformation, axesion transformation, and the function of selection and update are realized from Steps 12 to 14, and the discriminant rules for selection and update follow the fitness function Fitness based on predicted accuracy of LDM. If the specified termination criterion is met, the output solution $\text{Best}(\lambda_1, \lambda_2, C)$ will be the global optimal parameters to improve the LDM. Figure 2 shows the specific process of the STA selecting the optimal parameters of LDM by the fitness function.

5. Experimental Results and Analysis

The experimental data used the wind turbine variable pitch system fault data of one year's SCADA data set collected by a wind farm in East China, including variable pitch main power supply fault, variable pitch blade server drive temperature over-limit fault, and variable pitch system emergency stop fault. The number of fault samples and the number of fault features of the three fault data are shown in Table 2.

According to the different fault detection of the wind turbine variable pitch system, the sample set in normal operation should be classified as normal, and the sample set in failure should be classified as a fault. It is important to

Table 1: Meaning and value range of LDM parameters.

Parameter	Meaning	Value range
λ_1	The trade-off parameter of margin variance, which is adopted to adjust the weight of margin variance	$[2^{-1}, 2^{10}]$
λ_2	The trade-off parameter of margin mean, which is adopted to adjust the weight of margin mean	$[2^{-1}, 2^{10}]$
C	The loss function parameter, which is adopted to adjust the weight of the loss function in the objective function	$[2^{0}, 2^{20}]$

(1) **Best**⟵**Best**$_0$ (λ_{10}, λ_{20}, **C**)
(2) repeat
(3) if $\alpha < \alpha_{\mathbf{min}}$ then
(4) α⟵$\alpha_{\mathbf{max}}$
(5) end if
(6) λ_1⟵**Best** (1)
(7) λ_2⟵**Best** (2)
(8) **C**⟵**Best** (3)
(9) **L DM**⟵ (λ_1, λ_2, **C**, **training set**)
(10) **accuracy** (**L DM**)⟵**testing set**
(11) **Fitness**⟵**accuracy** (**L DM**)
(12) **Best**⟵**rotation transformation** (Fitness, **Best**, **SE**, β, α)
(13) **Best**⟵**expansion transformation** (Fitness, **Best**, **SE**, β, γ)
(14) **Best**⟵**axesion transformation** (Fitness, **Best**, **SE**, β, δ)
(15) α⟵α/$\mathbf{f_c}$
(16) Until the specified termination criterion is met
(17) Output **Best**

Algorithm 1: Optimal parameters of the improved LDM.

divide the whole sample set into two parts with each part containing normal data and fault data, which are used as a training set and testing set, respectively. The training set is mainly used to train the fault detection model, and the testing set is used to predict the model. The parameters of the STA are set as $\alpha_{\max} = 1$, $\alpha_{\min} = 1e-4$, $\beta = 1$, $\gamma = 1$, $\delta = 1$, SE = 30, and $f_c = 2$.

In order to verify that the STA can be adopted to improve the parameter adjustment of LDM and the improved LDM is effective for fault detection of the wind turbine variable pitch system, measures should be taken to introduce GS, PSO, and GA into the model parameter optimization method for comparison. The evaluation indexes were four indexes produced by the confusion matrix, including accuracy, F1-score, FPR, and FNR:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \\ F1 - \text{score} &= \frac{2}{(1/(\text{TP}/(\text{TP} + \text{FP}))) + (1/(\text{TP}/(\text{TP} + \text{FN})))}, \\ \text{FPR} &= \frac{\text{FP}}{\text{TN} + \text{FP}}, \\ \text{FNR} &= \frac{\text{FN}}{\text{TP} + \text{FN}}, \end{aligned} \tag{10}$$

where TP means the actual sample is positive and the prediction is positive; FP means the actual sample is negative and the prediction is positive; TN means the actual sample is negative and the prediction is negative; FN means the actual sample is positive and the prediction is negative.

In terms of the fault of wind turbine variable pitch main power supply, the boxplot of accuracy is shown in Figure 3. The comparison results of F1-score, FPR, and FNR are shown in Table 3.

The results indicated that the detection accuracy and F1-score of the improved LDM based on the STA for the variable pitch main power supply fault were higher than the values in terms of the methods of parameter adjustment through PSO, GA, and GS. FNR and FPR were the lowest among the four parameter adjustment methods.

For wind turbine variable pitch blade server drive temperature over-limit fault situation, Figure 4 shows the accuracy boxplot. The comparison results of F1-score, FPR, and FNR are shown in Table 4.

The results indicated that the detection accuracy and F1-score of the improved LDM based on the STA for wind turbine variable pitch blade server drive temperature over-limit fault were the highest while FNR and FPR were lower than the other three parameter adjustment methods.

For the wind turbine variable pitch system emergency stop fault situation, Figure 5 shows the accuracy boxplot. The comparison results of F1-score, FPR, and FNR are shown in Table 5.

The results indicated that the detection accuracy and F1-score of the improved LDM based on the STA for wind turbine variable pitch system emergency stop fault were

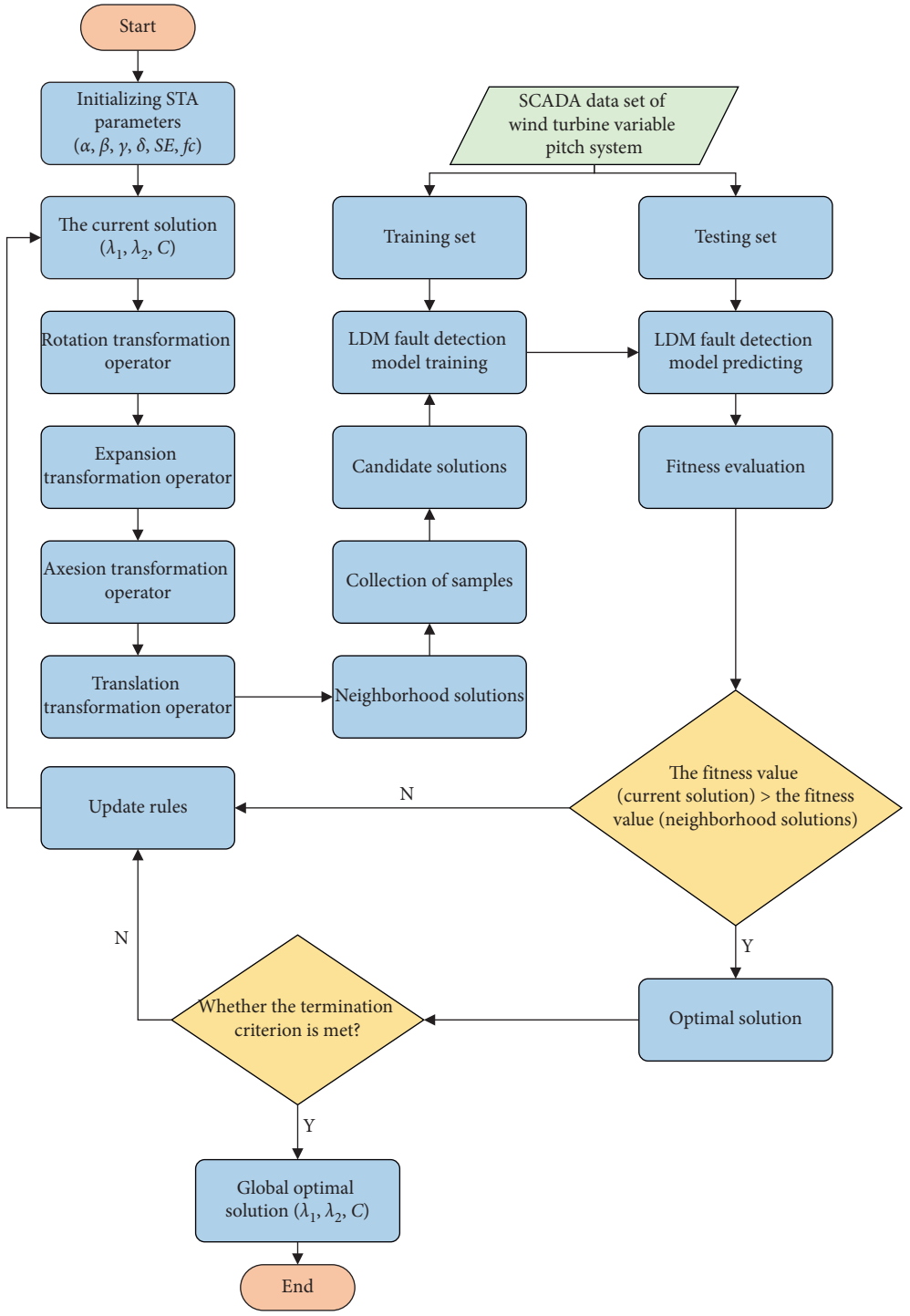


Figure 2: The specific process of the STA selecting optimal parameters of LDM by fitness function.

Table 2: The number of fault samples and the number of fault features of the three variable pitch system fault data.

Fault type	Number of fault samples	Number of fault features
Variable pitch main power supply fault	2902	212
Variable pitch blade server drive temperature over-limit fault	4864	212
Variable pitch system emergency stop fault	5893	212

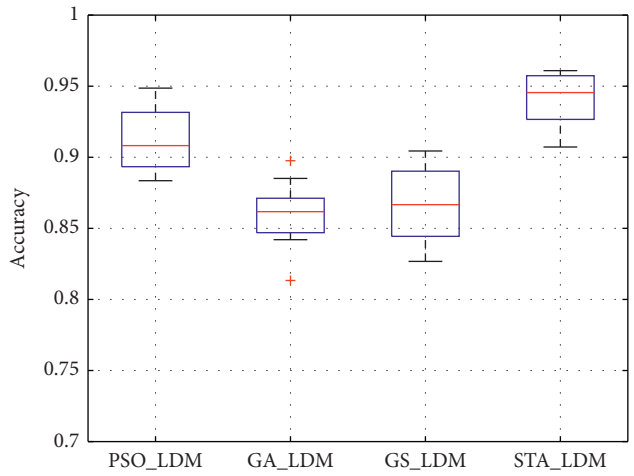


Figure 3: Boxplot of variable pitch main power supply fault detection accuracy.

Table 3: Performance comparison of variable pitch main power supply fault detection.

Fault detection model	F1-score	FPR	FNR
PSO_LDM	95.54% (±0.0057)	9.39% (±0.1086)	3.43% (±0.0296)
GA_LDM	89.38% (±0.0779)	13.19% (±0.0868)	9.84% (±0.1456)
GS_LDM	91.17% (±0.0081)	11.31% (±0.1059)	8.05% (±0.0159)
STA_LDM	96.49% (±0.0142)	5.07% (±0.1091)	2.97% (±0.0143)

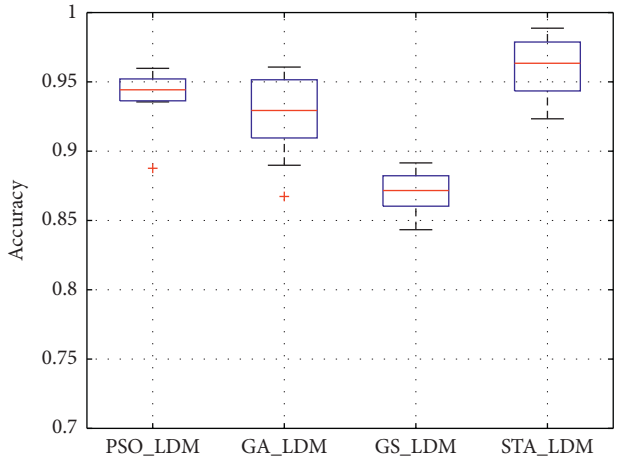


Figure 4: Boxplot of variable pitch blade server drive temperature over-limit fault detection accuracy.

Table 4: Performance comparison of variable pitch blade server drive temperature over-limit fault detection.

Fault detection model	F1-score	FPR	FNR
PSO_LDM	96.24% (±0.0092)	5.44% (±0.0166)	2.73% (±0.0371)
GA_LDM	95.68% (±0.0457)	6.71% (±0.1048)	3.04% (±0.0018)
GS_LDM	87.37% (±0.0135)	8.75% (±0.0191)	13.64% (±0.0713)
STA_LDM	98.72% (±0.0241)	1.10% (±0.0049)	1.16% (±0.0118)

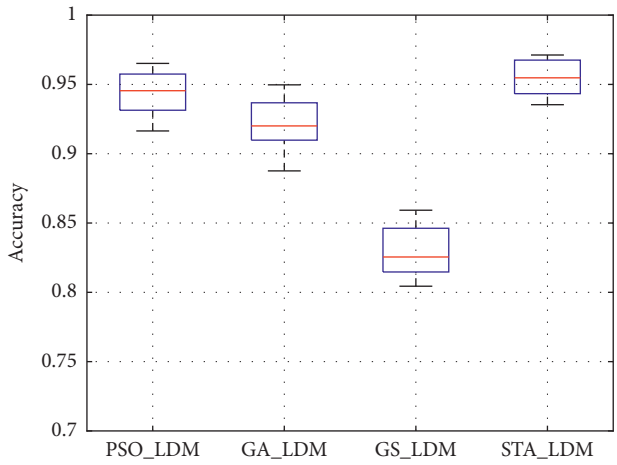


Figure 5: Boxplot of variable pitch system emergency stop fault detection accuracy.

Table 5: Performance comparison of variable pitch system emergency stop fault detection.

Fault detection model	F1-score	FPR	FNR
PSO_LDM	96.55% (±0.0107)	2.73% (±0.1102)	4.23% (±0.0544)
GA_LDM	95.50% (±0.0105)	6.69% (±0.0174)	3.98% (±0.0137)
GS_LDM	81.85% (±0.0134)	11.72% (±0.0753)	15.16% (±0.0096)
STA_LDM	97.19% (±0.0275)	2.21% (±0.0024)	3.65% (±0.0168)

higher than the values in terms of the methods of parameter adjustment through PSO, GA, and GS. FNR and FPR were the lowest among the four parameter adjustment methods.

6. Conclusion

Concerning the problem of dependent parameter selection of the fault detection model, this paper introduces the STA to improve LDM in terms of the parameter optimization of the classification algorithm. First, in order to meet the structure need of the optimization problem, the three parameters in LDM were regarded as a three-dimensional vector form, a state in the STA. In addition, a new state candidate assembly was generated by alternately using the four transformation operators. Second, the accuracy of the fault detection model output is used as a fitness function to support parameter updating and optimization. Finally, for verifying the effectiveness of the wind turbine variable pitch system fault detection method based on the improved LDM, the paper introduced the GS method, PSO, and GA for comparison on parameter optimization. The evaluation indexes were accuracy, F1-score, FPR, and FNR. The experimental data were variable pitch main power supply fault data, variable pitch blade server drive temperature over-limit fault data, and variable pitch system emergency stop fault data.

Experimental results showed that the fault detection model which used the STA for parameter optimization had higher accuracy and lower FPR and FNR than the other three optimization algorithms, which proved that the improved LDM has stronger capability of detecting wind turbine variable pitch system fault.

On account of the vulnerability of the wind turbine to be affected by the environment and load while running, it is incomprehensive to use a single detection method in the process of fault detection. As a result, it is indispensable to study a hybrid fault detection method based on various fault detection methods and technologies in the future.

Data Availability

The data used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data, [6/12 months] after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors contributed equally to this work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant no. 61403046), the Natural Science Foundation of Hunan Province, China (grant no. 2019JJ40304), the Changsha University of Science and Technology "The Double First Class University Plan" International Cooperation and Development Project in

Scientific Research in 2018 (grant no. 2018IC14), the Research Foundation of Education Bureau of Hunan Province (grant No.19K007), the Science and Technology Progress and Innovation Plan Project of Hunan Provincial Department of Transportation in 2018 (grant no. 201843), the Key Laboratory of Renewable Energy Electric-Technology of Hunan Province, the Key Laboratory of Efficient and Clean Energy Utilization of Hunan Province, the Innovative Team of Key Technologies of Energy Conservation, the Emission Reduction and Intelligent Control for Power-Generating Equipment and System, the CSUST, Hubei Superior and Distinctive Discipline Group of Mechatronics and Automobiles (grant no. XKQ2020009), the National Training Program of Innovation and Entrepreneurship for Undergraduates (grant no. 202010536016), and the Major Fund Project of Technical Innovation in Hubei (grant no. 2017AAA133), Hubei Natural Science Foundation Youth Project (2020CFB320).

References

[1] X. Xu, D. Niu, B. Xiao, X. Guo, L. Zhang, and K. Wang, "Policy analysis for grid parity of wind power generation in China," *Energy Policy*, vol. 138, Article ID 111225, 2020.

[2] M. Nazir, A. Q. Khan, G. Mustafa, and M. Abid, "Robust fault detection for wind turbines using reference model-based approach," *Journal of King Saud University–Engineering Sciences*, vol. 29, no. 3, pp. 244–252, 2017.

[3] S. Cho, Z. Gao, and T. Moan, "Model-based fault detection, fault isolation and fault-tolerant control of a blade pitch system in floating wind turbines," *Renewable Energy*, vol. 120, pp. 306–321, 2018.

[4] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 293–311, 2003.

[5] Y.-Y. Ko, "A simplified structural model for monopile-supported offshore wind turbines with tapered towers," *Renewable Energy*, vol. 156, pp. 777–790, 2020.

[6] R. Isermann, "Model-based fault-detection and diagnosis -status and applications," *Annual Reviews in Control*, vol. 29, no. 1, pp. 71–85, 2005.

[7] S. Ibaraki, S. Suryanarayanan, and M. Tomizuka, "Design of luenberger state observers using fixed-Structure<tex>$ cal H_ infty $</tex>Optimization and its application to fault detection in lane-keeping control of automated vehicles," *IEEE/ASME Transactions on Mechatronics*, vol. 10, no. 1, pp. 34–42, 2005.

[8] M. Zhong, S. X. Ding, Q.-L. Han, and Q. Ding, "Parity space-based fault estimation for linear discrete time-varying systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 7, pp. 1726–1731, 2010.

[9] B. Qiu, Y. Lu, L. Sun, X. Qu, Y. Xue, and F. Tong, "Research on the damage prediction method of offshore wind turbine tower structure based on improved neural network," *Measurement*, vol. 151, Article ID 107141, 2020.

[10] H. Chen, S. Jing, X. Wang, and Z. Wang, "Fault diagnosis of wind turbine gearbox based on wavelet neural network," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 37, no. 4, pp. 977–986, 2018.

[11] X. Zhu and Z. Gao, "An efficient gradient-based model selection algorithm for multi-output least-squares support vector regression machines," in *Pattern Recognition Letters*, vol. 111, pp. 16–22, 2018.

[12] H. Chih-Wei and L. Chih-Jen, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[13] J. Hang, J. Zhang, and M. Cheng, "Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine," *Fuzzy Sets and Systems*, vol. 297, pp. 128–140, 2016.

[14] L. Wenyi, W. Zhenfeng, H. Jiguang, and W. Guangfeng, "Wind turbine fault diagnosis method based on diagonal spectrum and clustering binary tree SVM," *Renewable Energy*, vol. 50, pp. 1–6, 2013.

[15] M. Tang, "Cost-sensitive large margin distribution machine for fault detection of wind turbines," *Cluster Computing*, vol. 22, no. 3, pp. 7525–7537, 2019.

[16] M. E. Barrios Aguilar, D. V. Coury, R. Reginatto, and R. M. Monaro, "Multi-objective PSO applied to PI control of DFIG wind turbine under electrical fault conditions," *Electric Power Systems Research*, vol. 180, Article ID 106081, 2020.

[17] Y. Zhang, H. Zheng, J. Liu, J. Zhao, and P. Sun, "An anomaly identification model for wind turbine state parameters," *Journal of Cleaner Production*, vol. 195, pp. 1214–1227, 2018.

[18] H. Yan, H. Mu, X. Yi, Y. Yang, and G. Chen, "Fault diagnosis of wind turbine based on PCA and GSA-SVM," in *Proceedings of the Prognostics and System Health Management Conference*, (PHM-Paris), Beijing, China, pp. 13–17, 2019.

[19] J.-T. Yin, Y.-F. Xie, Z.-W. Chen, T. Peng, and C.-H. Yang, "Weak-fault diagnosis using state-transition-algorithm-based adaptive stochastic-resonance method," *Journal of Central South University*, vol. 26, no. 7, pp. 1910–1920, 2019.

[20] R. Murugesan, J. Solaimalai, and K. Chandran, "Computer-aided controller design for a nonlinear process using a Lagrangian-based state transition algorithm," *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 977–996, 2020.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] T. Zhang and Z. Zhou, "Large margin distribution machine," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 313–322, New York, NY, USA, 2014.

[23] Z. Zhou, "Large margin distribution learning," in *Artificial Neural Networks in Pattern Recognition,* , pp. 1–11, Springer Verlag, Montreal, Canada, 2014.

[24] X. Zhou, C. Yang, and W. Gui, "State transition algorithm," *Journal of Industrial Management Optimization*, vol. 8, no. 4, pp. 1039–1056, 2013.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3673085, 10 pages
https://doi.org/10.1155/2020/3673085

Research Article

Traffic Flow Anomaly Detection Based on Robust Ridge Regression with Particle Swarm Optimization Algorithm

Mingzhu Tang,[1,2] **Xiangwan Fu**,[1] **Huawei Wu**,[2] **Qi Huang**,[3] **and Qi Zhao**[1]

[1]*School of Energy and Power Engineering, Changsha University of Science & Technology, Changsha 410114, China*
[2]*Hubei Key Laboratory of Power System Design and Test for Electrical Vehicle, Hubei University of Arts and Science, Xiangyang 441053, China*
[3]*School of Transport Management, Hunan Communication Polytechnic, Changsha 410132, China*

Correspondence should be addressed to Huawei Wu; whw_xy@163.com and Qi Huang; huang.qiqi.813@163.com

Received 31 July 2020; Accepted 2 September 2020; Published 30 September 2020

Academic Editor: Yong Chen

Traffic flow anomaly detection is helpful to improve the efficiency and reliability of detecting fault behavior and the overall effectiveness of the traffic operation. The data detected by the traffic flow sensor contains a lot of noise due to equipment failure, environmental interference, and other factors. In the case of large traffic flow data noises, a traffic flow anomaly detection method based on robust ridge regression with particle swarm optimization (PSO) algorithm is proposed. Feature sets containing historical characteristics with a strong linear correlation and statistical characteristics using the optimal sliding window are constructed. Then by providing the feature sets inputs to the PSO-Huber-Ridge model and the model outputs the traffic flow. The Huber loss function is recommended to reduce noise interference in the traffic flow. The L_2 regular term of the ridge regression is employed to reduce the degree of overfitting of the model training. A fitness function is constructed, which can balance the relative size between the k-fold cross-validation root mean square error and the k-fold cross-validation average absolute error with the control parameter η to improve the optimization efficiency of the optimization algorithm and the generalization ability of the proposed model. The hyperparameters of the robust ridge regression forecast model are optimized by the PSO algorithm to obtain the optimal hyperparameters. The traffic flow data set is used to train and validate the proposed model. Compared with other optimization methods, the proposed model has the lowest RMSE, MAE, and MAPE. Finally, the traffic flow that forecasted by the proposed model is used to perform anomaly detection. The abnormality of the error between the forecasted value and the actual value is detected by the abnormal traffic flow threshold based on the sliding window. The experimental results verify the validity of the proposed anomaly detection model.

1. Introduction

Traffic flow anomaly detection plays an essential role in the traffic field. Traffic jams have become a common thing in big cities and have received considerable critical attention. The traffic flow anomaly detection model can detect the abnormal traffic flow and can be achieved by constructing a traffic flow forecast model, which is helpful to avoid traffic congestion in time. The accurate forecast of traffic flow can not only provide a basis for real-time traffic control but also provide support for the alleviation of traffic jams and the effective use of traffic networks, and the forecast result of traffic flow can directly affect the accuracy of traffic anomaly detection. Useful information can be extracted from massive traffic flow data through the traffic flow forecast model so as to quickly forecast the short-term traffic flow in the future and detect the traffic flow abnormalities in time, thus improving the traffic operation efficiency.

In recent years, many experts and scholars have studied traffic flow forecasting. The ARIMA model is a classic time series model that is often used in traffic flow forecasts. Kumar and Vanajakshi proposed a SARIMA-based traffic flow forecast scheme, which effectively solved the problem of massive data required for model training [1]. Shahriari et al. combined bootstrap with the ARIMA model, which overcame the shortcomings of nonparametric methods lacking

theoretical support and improved the forecast accuracy of the model [2]. Luo et al. combined the improved SARIMA model with the genetic algorithm and used the real traffic flow to test the model. The model forecast results were good [3]. The ARIMA model forecasts the traffic flow based on historical values. If the model training data contain noise, the model's performance will be greatly reduced.

The neural network model can fit complex data relationships, which can learn the nonlinear relationships implicit in traffic flow. Qu et al. proposed a batch learning method to solve the time-consuming problem of the traffic flow neural network prediction model, which effectively reduced the training time of the neural network [4]. Zhang et al. used the spatiotemporal feature extraction algorithm to extract the temporal and spatial features in traffic flow. The features were input into the recurrent neural network for modeling and forecast, which effectively improved the forecast performance of the model [5]. Zhang et al. proposed a multitask learning deep learning model to forecast the traffic network flow. The nonlinear Granger causality analysis was used to select features for the model. The Bayesian optimization algorithm was used to optimize the model parameters. The forecast performance was better than that of the single deep learning model [6]. Do et al. used temporal and spatial attention mechanisms to help neural network models fully explore the temporal and spatial characteristics of the traffic flow, which not only effectively improved the prediction performance of the model but also enhanced the interpretability of the model [7]. The use of neural network models can cause overfitting easily with a calculation cost much higher than that of the traditional traffic flow forecast model. As neural networks can fit nonlinear relationships of data, it is easy to use the wrong noise as the implicit nonlinear relationship in the data, which will reduce the generalization ability of the model.

The support vector regression machine can fit data based on the strategy of structural risk minimization, which is a common model in the field of traffic flow forecasts. Wang et al. proposed an adaptive traffic flow forecast framework, which used the Bayesian optimization algorithm to optimize the parameters of the support vector machine model. The forecast performance was better than that of the SARIMA model [8]. Luo et al. used the discrete Fourier transform to extract the trend information in traffic flow and used the support vector machines for error compensation, which improved the forecast accuracy of the model [9]. The support vector regression machine solves the optimization problem based on quadratic programming. When the sample size is large, the model training time will be greatly increased. The support vector regression machine is very sensitive to the noise in the data. When the support vector regression machine selects the noise as the support vector, the forecast performance of the model will be poor.

Traffic flow anomaly detection plays an important role in the field of urban traffic control. Many studies have done related work in the field of traffic flow anomaly detection. Djenouri et al. proposed a framework for detecting temporal and spatial traffic anomalies. The KNN algorithm was applied to the space-time traffic database, and the traffic flows at ten different locations were experimented. Experimental results showed that the performance of the proposed framework is better than the baseline model [10]. Yujun et al. proposed a hybrid model that contained the Poisson mixture model and coupled hidden Markov model. The proposed model considered the spatial correlation of traffic flow and the degree of traffic congestion. Semisynthetic and real traffic anomaly data were used to verify the validity of the model [11]. Zhang et al. employed the dictionary-based compression theory to identify the spatial and temporal characteristics of traffic flow and used anomaly index to quantify the degree of traffic anomalies [12]. The proposed method can clearly detect the location of traffic flow spatial anomalies. Noise in traffic data may lead to false detection results of traffic anomaly detection models, which may affect the normal operation of traffic networks.

Influenced by factors such as mechanical damage, line aging, signal loss, and environmental interference, the data detected by the traffic flow sensor contain a lot of noise. Huber loss function is a mixture of L_1 and L_2 loss functions, which is insensitive to noise [13], the L_2 regular term of the ridge regression can effectively avoid overfitting caused by model training [14]. To improve the generalization performance of the model, the sum of $\mathrm{RMSE}_{k_{cv}}$ and $\eta * \mathrm{MAE}_{k_{cv}}$ on the training set based on k-fold cross-validation is constructed as the fitness function and the PSO algorithm is used to optimize the model hyperparameters. The PSO algorithm originated from the research on the foraging process of birds [15]. It has a simple structure. Each particle in the particle swarm has three main parameters: position, velocity, and fitness. In recent years, many pieces of literature have achieved good results using the particle swarm optimization algorithm [16–20].

To solve the problem of noise in traffic flow data, a Huber-Ridge traffic flow anomaly detection model with the particle swarm optimization (PSO) algorithm is proposed. The Huber-Ridge model is used to reduce the negative impact of noise in the data. Huber-Ridge model performance depends on model hyperparameters. Therefore, it is very important to determine the optimal model hyperparameters. A PSO algorithm based on the proposed fitness function is used to search for the optimal hyperparameters of the model so that the model has the best performance.

The remaining part of the paper proceeds as follows: Section 2 introduces the theoretical information of the Huber-Ridge algorithm; Section 3 proposes the data preprocessing steps and the steps using PSO algorithm to optimize the Huber-Ridge model parameters; Section 4 illustrates the model evaluation indexes; Section 5 presents the experimental content which contains the comparison of the forecast models and the results of traffic flow anomaly detections; Section 6 is conclusions.

2. Huber-Ridge Algorithm

2.1. Huber Function. The combination of the Huber function with the L_1 loss function and the L_2 loss function can effectively avoid the interference of noise in the data during the data fitting [21]. Its robustness is better than that of L_1

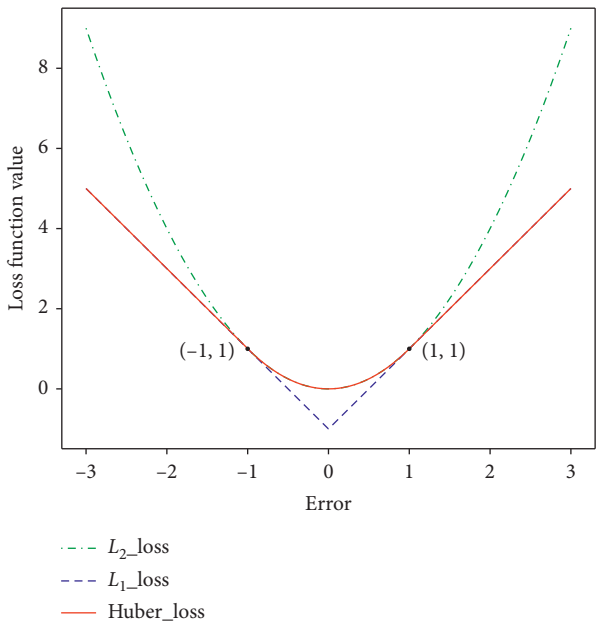


Figure 1: When the threshold M is 1, comparison of Huber function with L_1 loss function and L_2 loss function.

and L_2 loss functions. The definition of the Huber loss function is

$$\phi_{\text{hub}}(u) = \begin{cases} u^2, & |u| \le M, \\ M(2|u| - M), & |u| > M. \end{cases} \tag{1}$$

The definitions of L_1 loss function and L_2 loss function are shown in equations (2) and (3):

$$\phi_{L_1}(u) = M(2|u| - M), \tag{2}$$

$$\phi_{L_2}(u) = u^2, \tag{3}$$

where u is the error between the actual value and the estimated value, and M is the threshold. When the threshold M is 1, the comparison of the Huber loss function, the L_1 loss function, and the L_2 loss function is shown in Figure 1. Compared with the L_1 loss function, when u is smaller than the threshold M, the Huber loss function penalizes the model for making large errors. Compared with the L_2 loss function, when u is greater than the threshold M, the Huber loss function penalizes the model for making small error Therefore, the Huber loss function is quadratic for smaller errors and is linear for larger errors.

2.2. Ridge Regression Model. The ridge regression model is first proposed by Hoerl and Kennard. The ridge regression objective function adds the L_2 regular term based on the least square objective function [22]. Its definition is as follows:

$$\hat{w}_j = \text{argmin}_w \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} (w_j)^2 \right), \tag{4}$$

where $\sum_{j=1}^{k} (w_j)^2$ is the L_2 regular term and λ is the ridge parameter, which is the weight of the L_2 regular term.

For the linear regression model $y = wx + \varepsilon$, the least square estimation of the regression coefficient is defined as follows:

$$\hat{w} = \left(x^T x\right)^{-1} x^T y, \tag{5}$$

where x is the independent variable matrix and y is the dependent variable vector.

The mean square error of the least square estimation is defined as follows:

$$\hat{w}_{\text{mse}} = E\left(||w - \hat{w}||^2\right) = \sigma^2 tr\left(x^T x\right)^{-1} = \sigma^2 \sum_{i=1}^{q} \frac{1}{k_i}. \tag{6}$$

If there is a linear correlation between independent variables, the matrix $x^T x$ is a singular matrix. Some characteristic roots k_i of the singular matrix are close to zero, resulting in a large $\hat{w}_{\text{mse}}$. This indicates that there is a large error between the least-squares estimated value and the actual value. The addition of the disturbance term $\lambda I (\lambda > 0)$ on the matrix $x^T x$ will weaken the singularity, thereby reducing $\hat{w}_{\text{mse}}$. The least square estimation with the disturbance term added is the ridge estimation. The ridge estimate is defined as follows:

$$\hat{w}(\lambda) = \left(x^T x + \lambda I\right)^{-1} x^T y, \tag{7}$$

where λ is the ridge parameter and I is the identity matrix. $\hat{w}(\lambda)$ indicates the ridge estimation of the regression parameter w when the ridge parameter is λ. When $\lambda = 0$, the ridge estimation is the least square estimation. In the case of linear correlation of independent variables, the ridge estimation provides improved efficiency in parameter estimation problems, that is, biased but has lower variance than the least square estimator.

2.3. Huber-Ridge Regression. Owen uses the Huber loss function to replace the least-squares loss function and converted the ridge regression to the Huber-Ridge regression [23]. The definition of the Huber-Ridge model is as follows:

$$\hat{w}_j = \text{argmin}_w \left(\phi_{\text{hub}}(u) + \frac{\lambda}{2} \sum_{j=1}^{k} (w_j)^2 \right), \tag{8}$$

where w is the weight vector of the regression when the objective function is the smallest, w_j represents the estimate for each regression coefficient, $\sum_{j=1}^{k} (w_j)^2$ is the L_2 regular term, and $\lambda/2$ is the weight of the L_2 regular term, which is used to balance the relationship between the Huber loss function and the L_2 regular term. The Huber loss function can help the model avoid the influence of the data noise. The L_2 regular term helps the model have a proper sparsity and avoid overfitting of the model. The Huber-ridge regression combines the robustness of the Huber regression to noise with the regularization of the Ridge regression, which not only ensures the robustness of the model but also makes the regression model more stable.

$\sum_{j=1}^{k}(w_j)^2$ can be considered as $||w||_2^2$, which is the L_2 norm square of the weight vector w. The objective function $f(w)$ is defined as follows:

$$f(w) = \phi_{\text{hub}}(u) + \frac{\lambda}{2}||w||_2^2, \tag{9}$$

where u is the error. The objective function $f(w)$ is used to take the partial derivative of the weight vector w and let it to be zero. It can be obtained that the expression of the weight vector w is at the minimum value of the objective function in the direction of the weight vector w. The solution process of equation (9) is as follows:

$$\frac{\partial f(w)}{\partial w} = \frac{\partial \phi_{\text{hub}}(u)}{\partial u}\frac{\partial u}{\partial w} + \frac{\mathrm{d}(\lambda/2)||w||_2^2}{\mathrm{d}w} = 0, \tag{10}$$

where $u = xw - y$, xw is the estimated value, and y is the actual value. The first term of equation (10) can be simplified as

$$\frac{\partial \phi_{\text{hub}}(u)}{\partial u}\frac{\partial u}{\partial w} = \frac{\partial \phi_{\text{hub}}(u)}{\partial u}x, \tag{11}$$

$$\frac{\partial \phi_{\text{hub}}(u)}{\partial u} = \left[\frac{\partial \phi_{\text{hub}}(u_1)}{\partial u_1}, \frac{\partial \phi_{\text{hub}}(u_2)}{\partial u_2}, \ldots, \frac{\partial \phi_{\text{hub}}(u_n)}{\partial u_n}\right]^T = [h(u_1), h(u_2), \ldots, h(u_n)],$$
$$\frac{\partial \phi_{\text{hub}}(u_i)}{\partial u_i} = \begin{cases} 2|u_i|, & |u_i| \le M, \\ 2M, & |u_i| > M. \end{cases} \tag{12}$$

Let $\omega(u) = \partial h(u)/\partial u$, equation (11) can be simplified as

$$x^T\frac{\partial \phi_{\text{hub}}(u)}{\partial u} = x^T\varphi u,$$
$$\varphi = \text{diag}[\omega(u_1), \omega(u_2), \ldots, \omega(u_n)]. \tag{13}$$

The second term of equation (10) can be simplified as

$$\frac{\mathrm{d}(\lambda/2)w_2^2}{\mathrm{d}w} = \frac{\mathrm{d}(\lambda/2)\left(w^Tw\right)^2}{\mathrm{d}w} = \lambda w. \tag{14}$$

In summary, the solution process of equation (10) is as follows:

$$x^T\varphi u + \lambda w = 0,$$
$$x^T\varphi(xw - y) + \lambda w = 0, \tag{15}$$

$$w = \left(x^T\varphi x + \lambda I\right)^{-1}x^T\varphi y, \tag{16}$$

where I is the identity matrix. The optimal threshold M and the ridge parameter λ can be found in a fixed interval through the optimization algorithm. The weight vector w can be obtained by substituting the threshold value M, the ridge parameter λ, and the sample data into equation (16).

3. PSO-Huber-Ridge Model

3.1. PSO Algorithm. The core idea of the PSO algorithm comes from the foraging process of birds. For the PSO algorithm, the candidate solution of the optimization problem is a particle in the hyperparameter space. Each particle has its corresponding fitness value, speed, and position. The speed of the particle determines the direction and the displacement of the particle to look for the candidate solution. The PSO algorithm can find the optimal solution by iterating a group of initialized random particles.

For the PSO algorithm, there are m particles in the D-dimensional space. The speed of each particle can be expressed as $\overrightarrow{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, and the position of each particle can be expressed as $\overrightarrow{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iD})$, where $i \in [1, 2, \ldots, m]$. In the loop iteration, each particle represents a candidate solution. The corresponding fitness value can be obtained through the fitness function. The individual optimal particle and the global optimal particle can be selected based on the fitness value. The personal optimal particle ($p$best) is expressed as $\overrightarrow{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iD})$, and the global optimal particle ($g$best) is expressed as $\overrightarrow{p}_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$. Before the next iteration, each particle will update its speed and position through equations (17)–(19):

$$\overrightarrow{v}_i^{(t+1)} = \omega * \overrightarrow{v}_i^{(t)} + c_1 * r_1 * \left(\overrightarrow{p}_i - \overrightarrow{s}_i^{(t)}\right) + c_2 * r_2 * \left(\overrightarrow{p}_g - \overrightarrow{s}_i^{(t)}\right), \tag{17}$$

$$v_{ij}^{(t+1)} = \begin{cases} v_{\text{Max}}, & \left|v_{ij}^{(t+1)}\right| > v_{\text{Max}}, \\ v_{ij}^{(t+1)}, & \text{otherwise}, \end{cases} \tag{18}$$

$$\overrightarrow{s}_i^{(t+1)} = \overrightarrow{s}_i^{(t)} + \overrightarrow{v}_i^{(t+1)}, \tag{19}$$

$$i \in [1, 2, \ldots, m],$$
$$j \in [1, 2, \ldots, D], \tag{20}$$

where ω is the inertia factor ($\omega > 0$), c_1 is the local learning factor, and c_2 is the global learning factor ($c_1, c_2 > 0$). r_1 and r_2 are random numbers uniformly distributed between $[0, 1]$. t and $t + 1$ represent the number of iterations. v_{Max} represents the maximum speed of the particle.

For equation (17), where $\omega * \overrightarrow{v}_i^{(t)}$ is called the memory item, which refers to the influences of the speed on the particle when it is updated; $c_1 * r_1 * (\overrightarrow{p}_i - \overrightarrow{s}_i^{(t)})$ is called the

self-cognition term, which means that when the particle is updated, it is biased toward the individual optimal particle; $c_2 * r_2 * (\overrightarrow{p}_g - \overrightarrow{s}_i^{(t)})$ is called the group-cognition term, which means that when the particles are updated, they are biased toward the group optimal particle. It represents the result of collaboration among multiple particles.

3.2. Fitness Function. The PSO algorithm can find the optimal hyperparameters for the model based on the fitness function. The smaller the particle fitness value, the lower the forecast error of the hyperparameters. To improve the generalization ability of the model, the k-fold cross-validation [24] is added to the fitness function. The fitness function is defined as the sum of RMSE and MAE of k-fold cross-validation on the model training set. The expression equation for the fitness function is as follows:

$$\text{fitness} = \text{RMSE}_{k_{cv}} + \eta * \text{MAE}_{k_{cv}}. \tag{21}$$

$\text{RMSE}_{k_{cv}}$ is a root mean square error based on k-fold cross-validation and its expression is as follows:

$$\text{RMSE}_{k_{cv}} = \sqrt{\frac{\sum_{j=1}^{k_{cv}} \sum_{i=1}^{n} \left(y_{ij} - \widehat{y}_{ij}\right)^2}{nk_{cv}}}. \tag{22}$$

$\text{MAE}_{k_{cv}}$ is based on the average absolute error of k-fold cross-validation, and its expression equation is as follows:

$$\text{MAE}_{k_{cv}} = \sum_{j=1}^{k_{cv}} \sum_{i=1}^{n} \frac{\left|y_{ij} - \widehat{y}_{ij}\right|}{nk_{cv}}, \tag{23}$$

where n is the number of training samples, k_{cv} is the number of cross-validated subsets. $\widehat{y}_{ij}$ and y_{ij} are the model estimated value and the true value, respectively. The smaller the fitness function value, the better the corresponding particle.

The weight of $\text{MAE}_{k_{cv}}$ is η $(\eta > 0)$, which is also the control parameter used to balance the size of $\text{RMSE}_{k_{cv}}$ and $\text{MAE}_{k_{cv}}$. When $0 < \eta < 1$, $\text{MAE}_{k_{cv}}$ has less weight than $\text{RMSE}_{k_{cv}}$; when $1 < \eta < +\infty$, $\text{MAE}_{k_{cv}}$ has more weight than $\text{RMSE}_{k_{cv}}$; when $\eta = 1$, $\text{MAE}_{k_{cv}}$ has the same weight as $\text{RMSE}_{k_{cv}}$. $\text{RMSE}_{k_{cv}}$ has a small penalty for small errors. The degree of $\text{MAE}_{k_{cv}}$ penalty for errors remains unchanged. However, it does not punish large errors as much as $\text{RMSE}_{k_{cv}}$. The fitness function controls the degree of which the fitness function penalizes errors by adjusting the size of the control parameter η. As the control parameter η increases, the degree of penalty for small errors by the fitness function increases. Combining $\text{MAE}_{k_{cv}}$ and $\text{RMSE}_{k_{cv}}$, the problem of insufficient penalty for small errors for $\text{RMSE}_{k_{cv}}$ and insufficient penalty for large errors for $\text{MAE}_{k_{cv}}$ can be improved, which not only increases the penalty for model prediction errors but also improves the generalization ability of the model.

3.3. Data Preprocessing. Good data quality can improve the performance of the model. The missing values and the dimensional differences in the data will reduce the forecast performance of the model. Therefore, it is significant to preprocess the data. The data preprocessing can be divided into the following steps:

(1) Data cleaning. The previous value of the missing value should be used to fill in the missing value.

(2) Construction of model feature sets and output samples. For the traffic flow data set, the historical characteristics based on the linear correlation and the statistical characteristics based on the sliding window should be constructed. The model output sample is the traffic flow at the next time point in the sliding window.

(3) Data standardization. There are dimensional differences between different features. To prevent dimensional errors from reducing the model performance, the data distribution is transformed into a standard distribution with a mean of 0 and a variance of 1 through the standardized equation. The standardized equation is as follows:

$$\begin{cases} x_{ki} = \dfrac{X_{ki} - \overline{X}_i}{\sigma_i}, \\ \sigma_i = \dfrac{\left(X_{1i} - \overline{X}_i\right)^2 + \left(X_{2i} - \overline{X}_i\right)^2 + \cdots + \left(X_{ni} - \overline{X}_i\right)^2}{n}. \end{cases} \tag{24}$$

For the feature matrix, x_{ki} is the standardized data of the k-th row and the i-th column, $\overline{X}_i$ is the mean value of the i-th column, σ_i is the standard deviation of the i-th column, and n is the number of samples.

3.4. PSO-Huber-Ridge Model Optimization Process. The optimization steps of the PSO-Huber-Ridge model are as follows:

Step 1. Start the optimization.

Step 2. Determine the model inputs and outputs. The feature set is used as the model input and the model output the traffic flow.

Step 3. PSO-Huber-Ridge model parameter settings. The number of particles m, the inertial factor ω, the local learning factor c_1, and the global learning factor c_2 are input into the PSO algorithm. Initialize the speed $\overrightarrow{v}$ and the position $\overrightarrow{s}$ of each particle. Set the maximum number of iterations of the PSO algorithm i_{Max} and the value range of the model hyperparameters.

Step 4. $i = i + 1$.

Step 5. Particles update. Use equations (17)~(19) to update the speed $\overrightarrow{v}$ and position $\overrightarrow{s}$ of each particle.

Step 6. Fitness evaluation. Use equation (21) to calculate the fitness value of the particle based on the threshold value M and the ridge parameter λ of each particle.

Step 7. Optimal particle selection. Select the individual optimal particle and the global optimal particle according to the fitness value of the particles.

Step 8. Terminate training judgment. If the number of iterations i does not meet the termination condition ($i > i_{\text{Max}}$), return to Step 4. Otherwise, continue to the next step.

Step 9. Output optimization results. Output the threshold M and the ridge parameter λ in the global optimal particle.

Step 10. End the optimization.

4. Evaluation Indexes

The average absolute error (MAE), root mean square error (RMSE), and average absolute percentage error (MAPE) were used to evaluate the forecast performance of the model. The definition equations of MAE, RMSE, and MAPE are as follows:

$$\begin{aligned}
\text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|, \\
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}, \\
\text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right|,
\end{aligned} \tag{25}$$

where n is the number of samples in the test set, $\widehat{y}_i$ is the model forecast value, y_i is the true value. MAE and RMSE can reflect the forecast error of the model. The value range of MAPE is $[0, +\infty]$. The closer its value is to 0, the better the model performance.

5. Experimental Results and Analysis

5.1. Data Description. The traffic flow data set used in the experiment came from a highway intersection in Changsha City and was collected by a single detector with a data interval of 5 minutes. There were a small number of missing values in the traffic flow data set and the previous value of the missing value was used to fill in the missing points. The data sets containing 5 days of traffic flow were divided into the training set and the test set. The traffic flow from Saturday to Tuesday was used as the training set for the training model. The traffic flow on Wednesday was used as the test set to verify the performance of the trained model.

5.2. Feature Extraction and Selection. Historical characteristics based on the linear correlation from the traffic flow data were selected. The statistical characteristics based on the optimal sliding window were extracted.

The historical characteristics were selected. The Pearson correlation coefficient was used to judge the strength of the linear correlation between the data. The range of the correlation coefficient r was $[-1, 1]$. The closer to 1, the stronger the positive correlation between the data; the closer to -1, the stronger the negative correlation between the data; the closer to 0, the weaker the linear correlation between the data. The historical value of r greater than 0.9 was selected as historical characteristics. See Table 1 for the correlation coefficients of traffic flow with delays of 1–9.

According to Table 1, the historical characteristics with delays of 1–6 were selected as historical characteristics. To fully consider the periodicity of the traffic flow, the historical characteristics at the same time point last week were selected. The set of historical characteristics included the historical values with delays of 1–6 and the historical values at the same time point last week.

The statistical characteristics of the optimal sliding window were extracted. The maximum, minimum, median, mean, standard deviation, skewness, and kurtosis of the data set within the length of the sliding window were taken as the statistical characteristics. The sliding window length L had a value range of $[6, 150]$. The Huber-Ridge model with default hyperparameters ($\lambda = 0.0001, M = 1.35$) was used for the exhaustive operation on the traffic flow training set. The optimal window length was selected with the MAPE evaluation index as the standard. It can be seen from Figure 2 that when the MAPE value was the smallest, the sliding window length was 34 as the optimal sliding window length.

5.3. Experimental Results. The state transition algorithm (STA) [25], grey wolf optimizer (GWO) [26], genetic algorithm (GA) [27], and PSO algorithm were used to optimize the hyperparameters of the Huber-Ridge model. The range of model parameters is shown in Table 2:

The parameter values of the optimization algorithm are shown in Table 3:

The model training was performed using the standardized traffic flow training set. The fitness function based

Table 1: Pearson correlation coefficient of traffic flow with delays of 1–9.

Number of delays	9	8	7	6	5	4	3	2	1	0
R	0.85	0.87	0.89	0.90	0.92	0.93	0.95	0.96	0.97	1.00

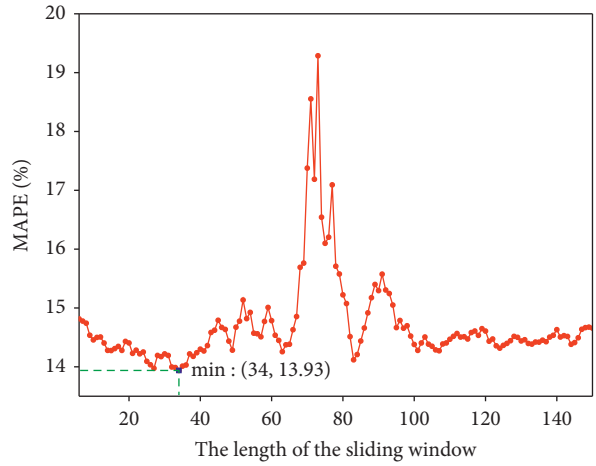


Figure 2: Relationship between sliding window length and MAPE (%).

Table 2: Model hyperparameters and value ranges.

Hyperparameter	Value range
Threshold M	$M \in [1, 4]$
Ridge parameter λ	$\lambda \in [0.0001, 4]$

Table 3: Optimization algorithm parameters.

Optimization algorithm	Parameter value
STA	$m = 30$; $\alpha_{max} = 1$; $\alpha_{min} = 10^{-4}$; $\beta = 1$; $\gamma = 1$; $\delta = 1$; $fc = 2$; maxital = 100
GWO	$m = 30$; $a \in [0, 2]$; $r_1, r_2 \in [0, 1]$; maxital = 100
GA	$m = 30$; $\text{prob}_{mut} = 0.001$; maxital = 100
PSO	$m = 30$; $\omega = 0.5$; $c_1 = 0.5$; $c_2 = 0.5$; $v_{Max} = 2$; maxital = 100

where m represents the number of seeds of each optimization algorithm, the maxital represents the maximum number of the iterations of the optimization algorithms. For the STA: the value range of the rotation factor α is $[\alpha_{max}, \alpha_{min}]$, which decreases in the form of an exponential function with $1/fc$ as the base with the increasing number of iterations; β indicates the translation factor; γ indicates the expansion factor; δ indicates the axesion factor. For the GWO: a is called the convergence factor and decreases from 2 linear to 0 with the increase of iterations; r_1 and r_2 are random numbers evenly distributed over an interval $[0, 1]$. For the GA: prob_{mut} represents the mutation probability, and the Partial-Mapped crossover is used as the crossover operator. For the PSO algorithm: ω indicates the inertia factor; c_1 indicates the local learning factor; c_2 indicates the global learning factor; v_{Max} represents the maximum speed of the particle.

Table 4: Comparison of optimization results between 4 model parameters.

	Threshold M	Ridge parameter λ	Fitness function value
STA-huber-ridge	1.54822021	3.64338165	0.203787303
GWO-huber-ridge	1.67576409	3.48931938	0.230480862
GA-huber-ridge	1.54444195	3.47396986	0.203815397
PSO-huber-ridge	1.55445167	3.99935861	0.203780642

on 10-fold cross-validation was used. The control parameter η of the fitness function was taken as 1. The performances of the STA-Huber-Ridge model, the GWO-Huber-Ridge model, the GA-Huber-Ridge model, and the PSO-Huber-Ridge model were compared and analyzed using RMSE, MAE, and MAPE evaluation functions. The optimization results of the four model parameters are shown in Table 4. The iterative comparison of their fitness values is shown in Figure 3.

It can be seen from Table 1 and Figure 3 that the fitness value of the STA algorithm dropped rapidly in the early stage of the iteration and then fell into the search for the local

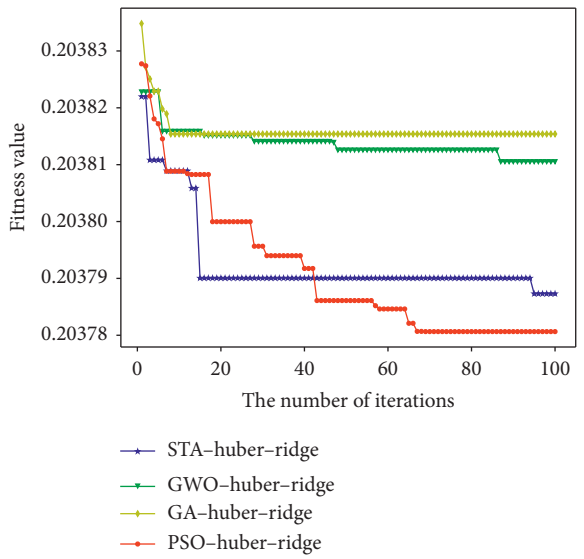


Figure 3: Comparison of four optimization iterations.

Table 5: Comparison of forecast results of 4 models.

	MAE	RMSE	MAPE (%)
STA-huber-ridge	7.06437	9.31449	13.9243
GWO-huber-ridge	7.06517	9.31559	13.9235
GA-huber-ridge	7.06418	9.31511	13.9238
PSO-huber-ridge	7.06393	9.31346	13.9230

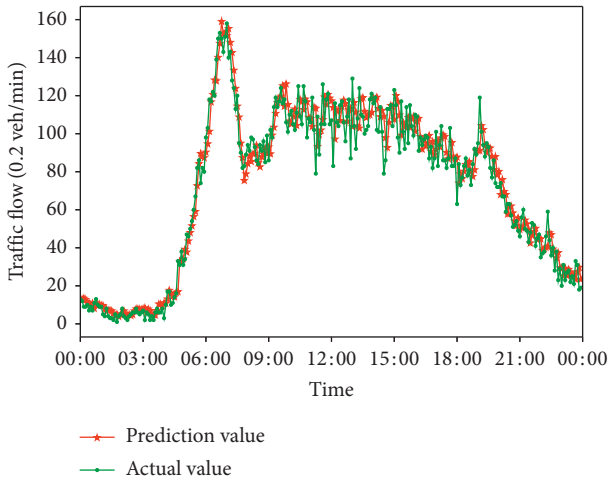


Figure 4: Forecast results of traffic flow by PSO-huber-ridge model.

optimum; after that, it dropped slowly in the later stage. The state transition algorithm used four transform operators to search. The search range was large and the early convergence was fast. However, transform operators with fixed values limited the global search capability of the state transition algorithm [28]. The fitness value of the GWO algorithm decreased slowly in the iterative process. The global optimization efficiency was not high. The GWO algorithm may easily fall into the local optimum and be unsuccessful in finding the global best [29]. The control parameters of the GWO algorithm decreased linearly with the iterative process, which cannot satisfy the complex search process [30]. The fitness value of GA stagnated in the early stage of the iteration and fell into the search for the local optimum. This is because the genetic algorithm has a premature phenomenon [31], making it difficult to jump out of the local optimum. Compared with the GWO, GA, and STA optimization algorithms, the PSO algorithm has a better iterative update strategy. It updates the particle position based on the individual experience of particles and the global experience

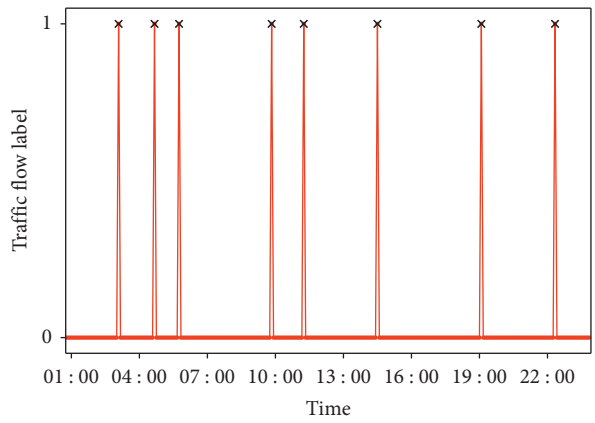


Figure 5: Traffic flow anomaly detection based on PSO-huber-ridge model.

of the particle swarm so that it will not all into the search for the local optimum easily.

The forecast evaluation results of the four models are shown in Table 5. The forecast result of the PSO-Huber-Ridge model is shown in Figure 4.

It can be seen from Table 5 that the PSO-Huber-Ridge model had the lowest MAE, RMSE, and MAPE; that is, the forecast performance of the PSO-Huber-Ridge model was the best. It can be seen from Figure 4 that the PSO-Huber-ridge model can well forecast the trend of the traffic flow at most time points.

Based on the error between the predicted value of the PSO-Huber-Ridge model and the actual value, the anomaly detection was performed on the traffic flow using the threshold (mean $\pm 2\sigma$) by calculating the mean value (mean) and variance (σ) of error data in a sliding window with a length of 10. If the forecast error at the next time point of the sliding window was greater than the anomaly detection threshold, the traffic flow at this time point was defined as an abnormal flow. The abnormal warnings would be reported to relevant traffic departments to avoid possible traffic jams. The label for abnormal traffic flow was defined as 1 and the label for normal traffic flow was defined as 0. The traffic flow anomaly detection based on the PSO-Huber-Ridge model is shown in Figure 5. It can be seen from Figure 5 that the proposed model can well detect the abnormal traffic flow in each period time.

6. Conclusions

To solve the problem of the large data noises in traffic flow, the traffic flow anomaly detection based on PSO-Huber-Ridge model is proposed. The strong robustness of the Huber function enables it to effectively reduce the influence of noise in traffic flow data on model training. The addition of the L_2 regular term of the ridge regression in the objective function can reduce the risk of model overfitting. The sum of $\mathrm{RMSE}_{k_{cv}}$ and $\mathrm{MAE}_{k_{cv}}$ based on 10-fold cross-validation is constructed as the fitness function to improve the generalization ability of the model. The optimal model parameters can be obtained through the particle swarm optimization algorithm so as to improve the model performance. Compared with the STA-Huber-Ridge, GA-Huber-Ridge, and GWO-Huber-Ridge models, the experimental results show that the PSO-Huber-Ridge model has the best model forecast performance. The traffic flow anomaly detection is performed using the traffic flow forecasted by the PSO-Huber-Ridge model. The error between the forecasted and actual traffic flow at a certain time point is large, which indicates that the regular pattern of traffic flow at that time point is different from that of history and may cause traffic congestion. The anomaly detection is performed on the traffic flow using the threshold (mean $\pm 2\sigma$). The experimental results verify the validity of the proposed traffic flow anomaly detection model.

The information contained in the traffic flow is complex. The PSO-Huber-Ridge model is limited to explore the linear information in the traffic flow. The nonlinear information needs further analysis and exploration. When extracting statistical features in feature engineering, the optimal sliding window is determined by the method of exhaustion. Its disadvantage is that it takes a long time and is not easy to apply. Using an adaptive method to extract features will greatly reduce the time of feature engineering. The Huber loss function reduces the negative impact of the data noise on the model training by reducing the penalty for large errors. Combining the Huber function with outlier detection method in data preprocessing can further improve the robustness of the model. Using adaptive feature extraction to mine linear and nonlinear information on the basis of improving model robustness is the next step.

Data Availability

The data used to support the findings of this study are currently under embargo, while the research findings are commercialized. Requests for data, 6/12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors contributed equally to this work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 61403046), the Natural Science Foundation of Hunan Province, China (Grant no. 2019JJ40304), Changsha University of Science and Technology "The Double First Class University Plan" International Cooperation and Development Project in Scientific Research in 2018 (Grant no. 2018IC14), the Research Foundation of Education Bureau of Hunan Province (Grant no. 19K007), Hunan Provincial Department of Transportation 2018 Science and Technology Progress and

Innovation Plan Project (Grant no. 201843), the Key Laboratory of Renewable Energy Electric-Technology of Hunan Province, the Key Laboratory of Efficient and Clean Energy Utilization of Hunan Province, Innovative Team of Key Technologies of Energy Conservation, Emission Reduction and Intelligent Control for Power-Generating Equipment and System, CSUST, Hubei Superior and Distinctive Discipline Group of Mechatronics and Automobiles (Grant no. XKQ2020009), National Training Program of Innovation and Entrepreneurship for Undergraduates (Grant no. 202010536016), Major Fund Project of Technical Innovation in Hubei (Grant no. 2017AAA133), and Hubei Natural Science Foundation Youth Project (Grant no. 2020CFB320).

References

[1] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *European Transport Research Review*, vol. 7, no. 3, p. 21, 2015.

[2] S. Shahriari, M. Ghasri, S. A. Sisson, and T. Rashidi, "Ensemble of ARIMA: combining parametric and bootstrapping technique for traffic flow prediction," *Transportmetrica A: Transport Science*, vol. 16, no. 3, pp. 1552–1573, 2020.

[3] X. Luo, L. Niu, and S. Zhang, "An algorithm for traffic flow prediction based on improved SARIMA and GA," *KSCE Journal of Civil Engineering*, vol. 22, no. 10, pp. 4107–4115, 2018.

[4] L. Qu, W. Li, W. Li, D. Ma, and Y. Wang, "Daily long-term traffic flow forecasting based on a deep neural network," *Expert Systems with Applications*, vol. 121, pp. 304–312, 2019.

[5] W. Zhang, Y. Yu, Y. Qi, F. Shu, and Y. Wang, "Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning," *Transportmetrica A: Transport Science*, vol. 15, no. 2, pp. 1688–1711, 2019.

[6] K. Zhang, L. Zheng, Z. Liu, and N. Jia, "A deep learning based multitask model for network-wide traffic speed prediction," *Neurocomputing*, vol. 396, pp. 438–450, 2020.

[7] L. N. N. Do, H. L. Vu, B. Q. Vo, Z. Liu, and D. Phung, "An effective spatial-temporal attention based neural network for traffic flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 12–28, 2019.

[8] D. Wang, C. Wang, J. Xiao, Z. Xiao, W. Chen, and V. Havyarimana, "Bayesian optimization of support vector machine for regression prediction of short-term traffic flow," *Intelligent Data Analysis*, vol. 23, no. 2, pp. 481–497, 2019.

[9] X. Luo, D. Li, and S. Zhang, "Traffic flow prediction during the holidays based on DFT and SVR," *Journal of Sensors*, vol. 2019, Article ID 6461450, 10 pages, 2019.

[10] Y. Djenouri, A. Belhadi, J. C. Lin, and A. Cano, "Adapted *K*-nearest neighbors for detecting anomalies on spatio-temporal traffic flow," *IEEE Access*, vol. 7, pp. 10015–10027, 2019.

[11] Y. Chen, J. Pu, J. Du, Y. Wang, and Z. Xiong, "Spatial-temporal traffic outlier detection by coupling road level of service," *IET Intelligent Transport Systems*, vol. 13, no. 6, pp. 1016–1022, 2019.

[12] Z. Zhang, Q. He, H. Tong, J. Gou, and X. Li, "Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 284–302, 2016.

[13] P. Petrus, "Robust Huber adaptive filter," *IEEE Transactions on Signal Processing*, vol. 47, no. 4, pp. 1129–1133, 1999.

[14] D. W. Marquardt and R. D. Snee, "ridge regression in practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.

[15] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[16] Z. Zhang, J. Yin, N. Wang, and Z. Hui, "Vessel traffic flow analysis and prediction by an improved PSO-BP mechanism based on AIS data," *Evolving Systems*, vol. 10, no. 3, pp. 397–407, 2019.

[17] C. Luo, C. Huang, J. Cao et al., "Short-term traffic flow prediction based on least square support vector machine with hybrid optimization algorithm," *Neural Processing Letters*, vol. 50, no. 3, pp. 2305–2322, 2019.

[18] Q. Ma, "Design of BP neural network urban short-term traffic flow prediction software based on improved particle swarm optimization," *AIP Conference Proceedings*, vol. 2073, no. 1, Article ID 020085, 2019.

[19] W. Cai, J. Yang, Y. Yu, Y. Song, T. Zhou, and J. Qin, "PSO-ELM: a hybrid learning model for short-term traffic flow forecasting," *IEEE Access*, vol. 8, pp. 6505–6514, 2020.

[20] L. Lin, J. C. Handley, Y. Gu, L. Zhu, X. Wen, and A. W. Sadek, "Quantifying uncertainty in short-term traffic prediction and its application to optimal staffing plan development," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 323–348, 2018.

[21] P. J. Huber, "Robust estimation of a location parameter,"‘-Robust estimation of a location parameter," in *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz and N. L. Johnson, Eds., pp. 492–518, Springer New York, New York, NY, USA, 1992.

[22] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[23] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemporary Mathematics*, vol. 443, pp. 59–72, 2006.

[24] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of *k*-fold cross validation in prediction error estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569–575, 2010.

[25] X. Zhou, C. Yang, C. Yang, and W. Gui, "State transition algorithm," *Journal of Industrial & Management Optimization*, vol. 8, no. 4, pp. 1039–1056, 2012.

[26] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[27] D. Whitley, "A genetic algorithm tutorial," *Statistics and Computing*, vol. 4, no. 2, pp. 65–85, 1994.

[28] X. Zhou, J. Long, C. Xu, and G. Jia, "An external archive-based constrained state transition algorithm for optimal power dispatch," *Complexity*, vol. 2019, Article ID 4727168, 11 pages, 2019.

[29] W. Long, J. Jiao, X. Liang, and M. Tang, "An exploration-enhanced grey wolf optimizer to solve high-dimensional numerical optimization," *Engineering Applications of Artificial Intelligence*, vol. 68, pp. 63–80, 2018.

[30] W. Long, J. Jiao, X. Liang, and M. Tang, "Inspired grey wolf optimizer for solving large-scale function optimization problems," *Applied Mathematical Modelling*, vol. 60, pp. 112–126, 2018.

[31] S. Yu and S. Kuang, "Fuzzy adaptive genetic algorithm based on auto-regulating fuzzy rules," *Journal of Central South University of Technology*, vol. 17, no. 1, pp. 123–128, 2010.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9392467, 14 pages
https://doi.org/10.1155/2020/9392467

Research Article

An Integrated Method for Fire Risk Assessment in Residential Buildings

Hongfu Mi,[1] **Yaling Liu**,[1] **Wenhe Wang**,[1] **and Guoqing Xiao**[2]

[1]*College of Safety Engineering, Chongqing University of Science and Technology, Chongqing 401331, China*
[2]*College of Chemistry and Chemical Engineering, Southwest Petroleum University, Chengdu 610500, China*

Correspondence should be addressed to Hongfu Mi; mimihh5@163.com

Received 12 May 2020; Revised 10 August 2020; Accepted 17 August 2020; Published 26 August 2020

Guest Editor: Esam Hafez Abdelhameed

Building fires are characterized by high uncertainty, so their fire risk assessment is a very challenging task. Many indexes and parameters related to building fires are ambiguous and uncertain; as a result, a flexible and robust method is needed to process quantitative or qualitative data and update existing information when new data are available. This paper presents a novel model to deal with the uncertainty of the residential building fire risk and systematically optimize its performance effectiveness. The model includes fuzzy theory, evidence reasoning theory, and expected utility methods. Fuzzy analysis hierarchy process is applied to analyze the residential building fire risk index system and determine the weights of the risk indexes, while the evidence reasoning operator is used to synthesize them. Three buildings were selected as a case study to illustrate the proposed fire risk model. The results show that the fire risk level of three buildings corresponds to "moderate" or below which is consistent with the previous study. These results also truly reflect the actual situation of fire safety in these residential buildings. The application of this model provides a powerful mathematical framework for cooperative modeling of the fire risk assessment system and allows data to be analyzed step by step in a systematic manner. It is expected that the proposed model could provide managers and researchers with flexible and transparent tools to effectively reduce the fire risk in the system.

1. Introduction

With the acceleration of industrialization, urbanization, and marketization in China, building construction industry has developed rapidly. Particularly, the structure and function of buildings are becoming more complex, and various new technologies and techniques are emerging constantly, which have led to the increasingly severe situation of building fires. According to the statistics provided by the Ministry of Public Security in 2013, a total of 388,821 fires were recorded in China, in which 52% (202,299) of fires occurred in buildings, resulting in 3410 civilian deaths or injuries and 3760 million Chinese yuan (CNY) direct property losses. Nowadays, building fire is considered to be an enormous threat to people's life and production in China, and a growing concern is how to take appropriate measures to reduce the fire risk, minimize the damage and loss caused by fire in buildings, and guarantee building fire safety. Therefore, it is urgent to establish a suitable fire risk assessment model, and it provides information through quantitative or qualitative analysis results to make decisions on whether to take steps to reduce the risk [1, 2].

There are mainly four conventional types of fire risk analysis methods: checklist, description, index, and probability method [3]. However, most of these approaches have prescriptive drawbacks which make them difficult to quantitative fire risk analysis due to the inability to deal with the uncertainties associated with the fire risk factors of the system. With the improvement of performance-based fire-protection design, some fire risk analysis models and corresponding software have emerged, such as FiRECAM™ (Fire Risk Evaluation and Cost Assessment Model) [4, 5], FIERAsystem (Fire Evaluation and Risk Assessment system) [6], CESARE-RISK (Centre for Environment Safety and Risk Engineering, RISK) [7, 8], and Crisp II (Computation of Risk Indices by Simulation Procedures) [9]. However, these models should depend on some strict constraints, such as a

large number of input data, specific fire scenarios, and the large amount of calculations. Consequently, researchers are concentrating on developing flexible fire risk analysis tools based on systematic safety theory. For example, Ibrahim et al. presented a fire risk method based on the analytical hierarchy method (AHP) for heritage buildings [10]. Lo developed a fire risk ranking system for existing buildings using the fuzzy set approach [11]. Liu et al. built a fire risk analysis system for commercial buildings by using the structure entropy weight method [12]. Xin and Huang proposed scenario cluster methods in the process of the fire risk analysis model for residential buildings [2]. Briefly speaking, these methods reveal two main challenges in an uncertain environment associated with the fire risk factors of the system. The first challenge faced by these methods is the lack of the ability to process a variety of data suitable for fire risk reasoning mechanisms, and the second is the lack of the ability to analyze the interdependence of risk factors. In this paper, a fire risk analysis model integrated fuzzy theory, and evidential reasoning (ER) theory is presented for residential buildings. Compared with the traditional fuzzy reasoning approach, ER has the advantage of avoiding losing useful information; therefore, it can be applied to model complex systems. The framework of this model is organised as follows. Section 2 illustrates the methodology of the research. Section 3 presents a case study to verify the feasibility of the methodology. Sections 4 and 5 discuss the empirical results and conclusions.

2. Methodology

Quantitative risk assessment (QRA) techniques are usually used for assessing uncertainties in building fires. However, due to the lack of fire accident statistics, an effective solution is to integrate expert judgments into the QRA process. QRA consists of four main procedures: hazard identification, occurrence probability calculation, consequence severity assessment, and risk quantification [13, 14]. In order to process the complex system structure and promote a flexible implementation method, different decision-making techniques can be used, such as fuzzy analytic hierarchy process, fuzzy set theory, and evidence reasoning method. Due to the fact that fuzzy logic could provide a flexibility way to represent the vague information resulting from the lack of data or knowledge. Therefore, the fuzzy set theory has a wide application in different fields such as reliability engineering, system safety, and risk assessment [15].

The proposed framework, shown in Figure 1, allows step-by-step analysis of the utility tunnel fire risk in a transparent way, as described as follows:

(1) Identifying fire risk factors and establishing the hierarchical structure of the index system

(2) Using fuzzy analytic hierarchy process (FAHP) to calculate the weights of indexes

(3) Applying the belief degree structure based on the fuzzy set theory to measure the fire risk

(4) Aggregating the result of the fire risk using the evidence reasoning (ER) algorithm

(5) Using the expected utility method to obtain a clear result of the fire risk

(6) Sensitivity analysis

2.1. Identifying Fire Risk Factors and Establishing the Hierarchical Structure of the Index System. In order to make better decisions on fire control protection and emergency evacuation measures, a structured and systematic approach is needed. It is better to describe the fire risk problem in a hierarchical structure so that decision makers could have a thorough understanding of the system, especially when it is a complex system with multilevel structural indexes.

According to NFPA550 Guidelines, to achieve fire safety, reducing the fire risk mainly starts from two aspects: one is to prevent the occurrence of fire, and the other is to control the impact of fire [16]. In this paper, fire risk factors of these two aspects are, respectively, defined as disaster-causing factors and loss-controlling factors. Disaster-causing factors may cause the fire risk to be transformed into disaster before fire occurs, while loss control factors signified various fire protection and management measures to control the development process of fire and mainly involved four aspects: passive measures, active measures, fire management, and fire brigade fighting.

Based on the characteristics of residential building fire and the literature review [6, 17–20], the factors influencing the risk of building fires are analyzed from the two aspects of disaster-causing factors and loss-controlling factors. A general hierarchical structure (presented in Figure 2) is finally established after theoretical preparation, the initial construction of the index system, the optimization of the index system, and the determination of the index system.

2.2. Fuzzy Analytic Hierarchy Process (FAHP). The traditional analytic hierarchy process (AHP) constructs the judgment matrix by comparing the two factors with the 1–9 scale method. However, due to the subjectivity of human judgment, different people will get different conclusions. Using the triangular fuzzy number to scale the two-pair comparison of factors can consider the uncertainty of experts in analysis and judgment, which give the range of expert's judgment in the form of intervals to reduce subjectivity. In 1996, Chang [21] applied triangular fuzzy numbers to construct judgment matrices and combined with the extent analysis method to calculate the weights of each index in the hierarchical structure. Finally, the traditional AHP is transformed into the FAHP in the fuzzy environment, which can provide more practical results [22].

2.2.1. Triangular Fuzzy Number. Suppose the triangular fuzzy number is M, and its membership function $\mu_M: R \longrightarrow [0, 1]$ is equal to

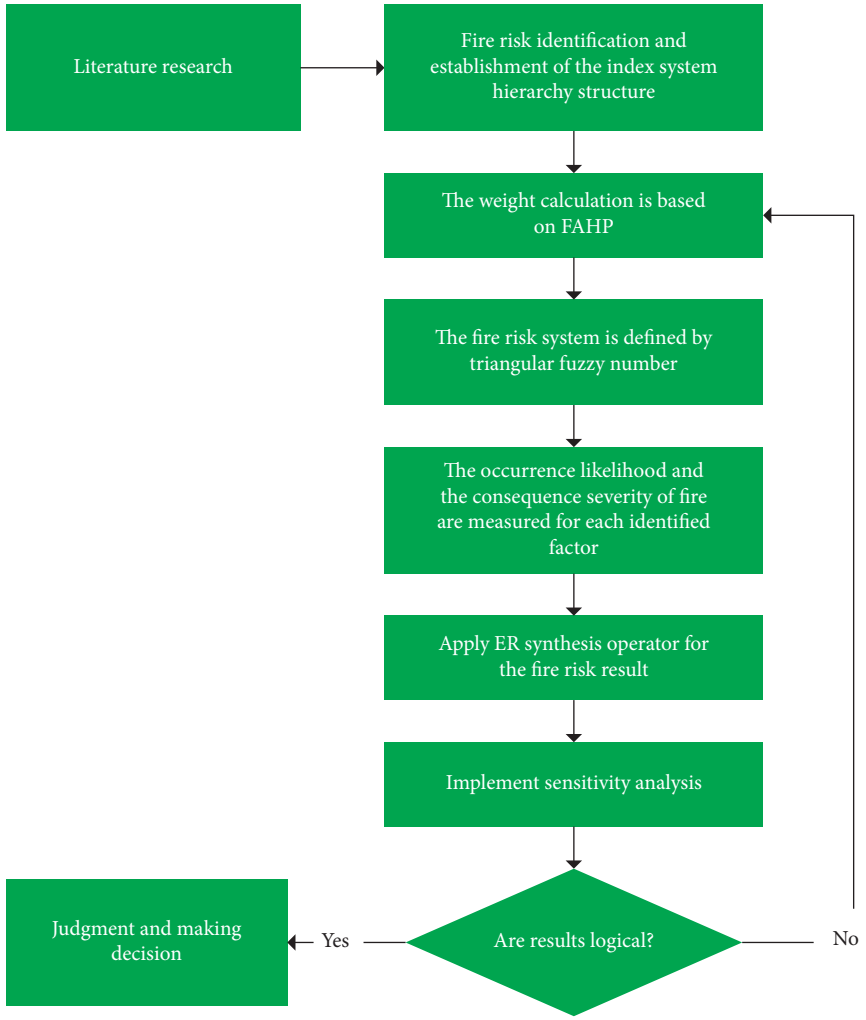


Figure 1: The procedure for fire risk assessment in residential buildings.

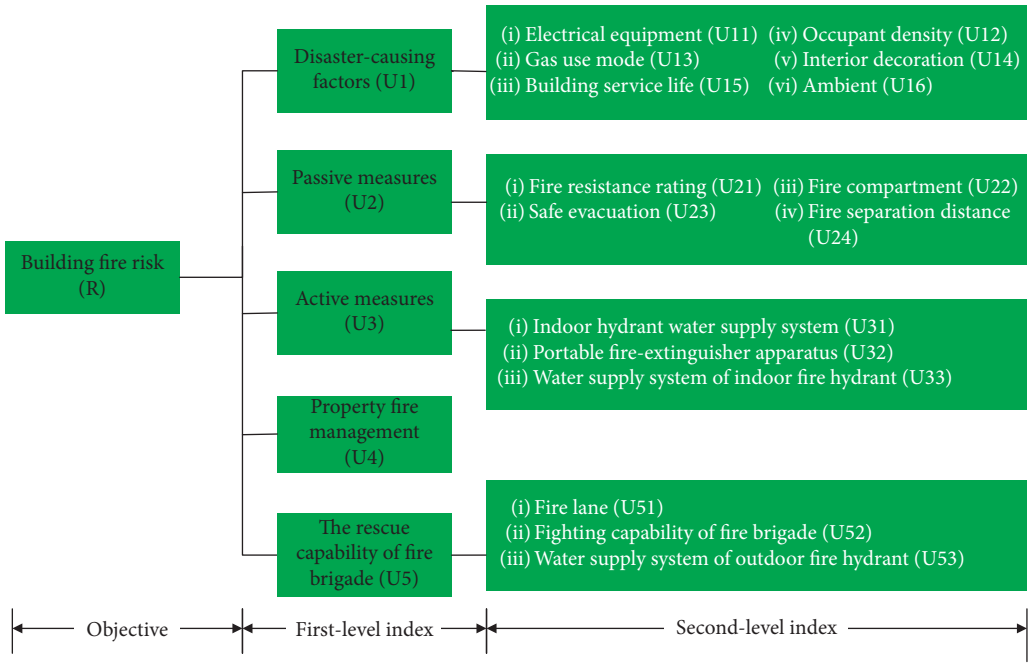


Figure 2: The hierarchical structure for the residential building fire risk model.

$$\mu_M(x) = \begin{cases} \dfrac{x-l}{m-l}, & l \le x \le m, \\ \dfrac{x-u}{m-u}, & m \le x \le u, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Herein, $l \le m \le u$, l and u represent the lower and upper boundary value of triangular fuzzy number M, respectively, and m represents the median value of triangular fuzzy number M. Generally, triangular fuzzy number M can be abbreviated as (l, x, m). Let $M_1 = (l_1, x_1, m_1)$ and $M_2 = (l_2, x_2, m_2)$ be triangular fuzzy numbers; then, the possibility degree of $M_1 \ge M_2$ is defined as follows:

$$V(M_1 \ge M_2) = \begin{cases} 1, & m_1 \ge m_2, \\ \dfrac{l_2 - u_1}{(m_1 - u_1) - (m_2 - l_2)}, & m_1 < m_2, l_2 \le u_1, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

2.2.2. Fuzzy Synthetic Extent. Consider $X = \{x_1, x_2, \ldots, x_n\}$ as a set of analytic objects and $U = \{u_1, u_2, \ldots, u_n\}$ as a target set; we can get the extent value of the i-th object satisfying the j-th goal, in which the sign is $M_{E_i}^j$. Then, the value of synthetic extent of the i-th object is defined as [21, 23]

$$S_i = \sum_{j=1}^{m} M_{E_i}^j \left(\sum_{i=1}^{n} \sum_{j=1}^{m} M_{E_i}^j \right)^{-1}. \quad (3)$$

2.2.3. The Procedure of the FAHP. In the evaluation of the fire risk, the determination of the weight of each fire risk factor is particularly important. The weight represents the relative importance of each factor in the overall evaluation. Only when the weight of each factor is obtained, the fire risk assessment can be carried out. The steps of determining the weight by the FAHP method are as follows:

(1) According to the objective of fire risk assessment, the hierarchical system structure is established, which is composed of fire risk factors.

(2) The judgment matrix is constructed by triangular fuzzy numbers (according to Table 1) through a pairwise comparison of the index system by experts [24, 25].

(3) According to equation (3), the value S_i of synthetic extent S_i of each factor is obtained.

(4) The possibility degree $d'(A_i)$ is calculated such that factor A_i is more important than others:

$$d'(A_i) = \underbrace{\min}_{j=1,2,\ldots,n, j \ne i} V(S_i \ge S_j), \quad i = 1, 2, \ldots, n. \quad (4)$$

Then, the weight vector is obtained:

$$W' = (d'(A_1), d'(A_2), \ldots, d'(A_n))^T. \quad (5)$$

Finally, the normalized weight vector is obtained.

2.3. Application of the Belief Structure for the Fire Risk Calculation. After identifying fire risk factors and establishing the hierarchical structure of the index system, another important task of risk management is to assess the risk, which is an effective way to prevent or reduce the effect of the fire [2]. In this paper, the fire risk of residential buildings is defined as the result of comprehensive measurement associated to the occurrence likelihood and the consequence severity of the fire. The formula is as follows:

$$P = L \otimes S, \quad (6)$$

where P is the magnitude of the fire risk presented by various potential fire hazards, L refers to the occurrence likelihood of potential fire hazards or fire risk factors, S implies the consequence severity of potential fire hazards or fire risk factors, and $\otimes$ represents the interconnection relationship between L and S.

2.3.1. Fuzzy Linguistic Variables for the Fire Risk. After defining the fire risk, it is necessary to transform the factors into the same form of fuzzy evaluation grade. Due to the uncertainty, analysts tend to use linguistic variable terms rather than precise numerical values to evaluate the fire risk. Therefore, this paper uses a ranking form of fuzzy linguistic variables to represent the fire risk profile of each factor.

A belief degree is generally used to describe the level of expectations for trust events, and it must be less than or equal to 1 to express the degree of which answer is considered true. Individual differences of belief degree depend on assessor's expertise and the knowledge to understand the assessment system. A belief structure can solve the problems of fuzziness, uncertainty, and imprecision in human decision-making. Therefore, this paper presents a model that combines fuzzy linguistic variables and a belief degree to construct a belief structure with the same set of assessment grades [26]. These sets' form of each factor could be expressed as follows:

$$\begin{aligned} R_L &= [R_{L1}, R_{L2}, R_{L3}, R_{L4}, R_{L5}] = \{\text{highly unlikely, unlikely slight, likely, reasonably likely, highly likely}\}, \\ R_S &= [R_{S1}, R_{S2}, R_{S3}, R_{S4}, R_{S5}] = \{\text{negligible, slight, moderate, serious, catastrophic}\}, \\ R &= [R_1, R_2, R_3, R_4, R_5] = \{\text{very low, low, medium, high, very high}\}. \end{aligned} \quad (7)$$

Table 1: Relative importance described by the triangular fuzzy numbers.

Relative importance in qualitative description	Description	Triangular fuzzy number
Equally important	Both indexes contribute equally to the target fire risk	(1, 1, 2)
Between equally and slightly important	Between the front and the back	(1, 2, 3)
Slightly important	Based on the objective judgment and expert experience, it is considered that one index contributes slightly more to the target fire risk than another	(2, 3, 4)
Between slightly and strongly important.	Between the front and the back	(3, 4, 5)
Strongly important	Based on the objective judgment and expert experience, it is considered that the contribution of one index to the target fire risk is better than another	(4, 5, 6)
Between strongly and very strongly important	Between the front and the back	(5, 6, 7)
Very strongly important	The index's contribution to the target fire risk is significantly better than another	(6, 7, 8)
Between very strongly and absolutely important.	Between the front and the back	(7, 8, 9)
Absolutely important	There is evidence that one index is definitely better than another for the target fire risk	(8, 9, 9)

Among them, R_L, R_S, and R represent the evaluation grade variables of the occurrence likelihood of fire, consequence severity of fire, and fire risk, respectively.

2.3.2. Fire Risk Level Based on a Belief Structure. Because of the complexity and uncertainty of the system, the type of membership function is not the dominant factor in the risk assessment analysis of the system [27]. Therefore, as listed in Table 2 and Figure 2, this paper applies the triangular membership function which is the most commonly used one to describe the subjective linguistic variables [15] and adopts the five-phase method, adjusted and modified from Ngai and Wat [28] to represent the occurrence likelihood of fire (L) and the consequence severity of building fire (S), respectively. Suppose that the occurrence likelihood of building fire (L) and the consequence severity of building fire (S) for each factor are independent of each other; they are denoted by triangular fuzzy numbers $\mathrm{FTN}_L = (a_L, b_L, c_L)$ and $\mathrm{FTN}_S = (b_S, b_S, c_S)$. Then, the desired fire risk of each factor can be expressed as

$$\mathrm{FTN}_{LS} = \mathrm{FTN}_L \otimes \mathrm{FTN}_S = (a_L \otimes a_S, b_L \otimes b_S, c_L \otimes c_S). \quad (8)$$

Fuzzy risk P with a belief structure can be obtained through the following steps:

(1) According to formula (8), calculate FTN_{LS} of each factor

(2) Map the calculated FTN_{LS} to the FTN_P membership curve, and obtain the intersection points of each fuzzy language level variable (note: if there is more than one intersection point on a certain fuzzy language level variable, take the intersection point with the largest longitudinal coordinate value), as shown in Figures 3 and 4

(3) Obtain a set of intersection values (β_P), which denote five nonstandardized linguistic variable levels of risk P in the form of fuzzy sets

(4) Normalize β_P, and obtain the basic belief degree β of each factor related to its fire risk

As listed in Table 2, if a single factor judged by experts' knowledge and experience takes a fire risk value of that the occurrence likelihood of building fire corresponds to (0.5,0.75,1), the consequence severity of building fire corresponds to (0.25,0.50,0.75). The corresponding value of FTN_{LS} will be (0.125,0.375,0.75). Then, map FTN_{LS} to FTN_P to get the set of intersection values (β_P), shown in Figure 4. Finally, the basic belief degree β is obtained after the normalization of β_P, which denotes that five nonstandardized linguistic variables of very low, low, general, high, and very high correspond to 0.25, 0.75, 0.8, 0.4, and 0, respectively.

It is noteworthy that the triangular fuzzy numbers for the occurrence likelihood (L) and the consequence severity (S) of building fire judged by experts cannot be used directly as input data for the synthesis of fire risk results by the evidential reasoning algorithm. They need to convert to five standardized linguistic variable terms before synthesizing the fire risk of each factor [29].

2.4. Synthesizing Assessment Result Using the Evidence Reasoning Algorithm. The theory of evidential reasoning was first proposed by Dempsterin 1967 [30]. Then, in 1976, Shafer further expanded and improved Dempster's work to form a complete and systematic theory [31]. Subsequently, in commemoration of Dempster and Shafer's contribution to the theory of evidence reasoning, the theory was often called Dempster–Shafer theory or D-S theory for abbreviation. D-S theory can be used to deal with uncertain, imprecise, and or inaccurate information. It was originally used as an approximate reasoning tool for information synthesis in expert systems [32]. Later, it was applied to the decision-making judgment of uncertain problems [33]. Due to the uncertainty of the changing system environment and qualitative descriptive information and to consider the influence of the weight in the synthesis of evidence, evidence reasoning algorithm (ER algorithm) was proposed [34].

After knowing the basic belief degree β and the weight ω of each factor, suppose $m_{n,i}$ is a basic probability mass,

Table 2: Linguistic variables described by the triangular membership number.

Likelihood of building fire (L)	Severity of building fire (S)	Triangular fuzzy number
Highly unlikely (HU)	Negligible (NE)	(0.00, 0.00, 0.25)
Unlikely slight (US)	Slight (SL)	(0.00, 0.25, 0.50)
Likely (LI)	Moderate (MO)	(0.25, 0.50, 0.75)
Reasonably likely (RL)	Serious (SE)	(0.50, 0.75, 1.00)
Highly likely (HL)	Catastrophic (CA)	(0.75, 1.00, 1.00)

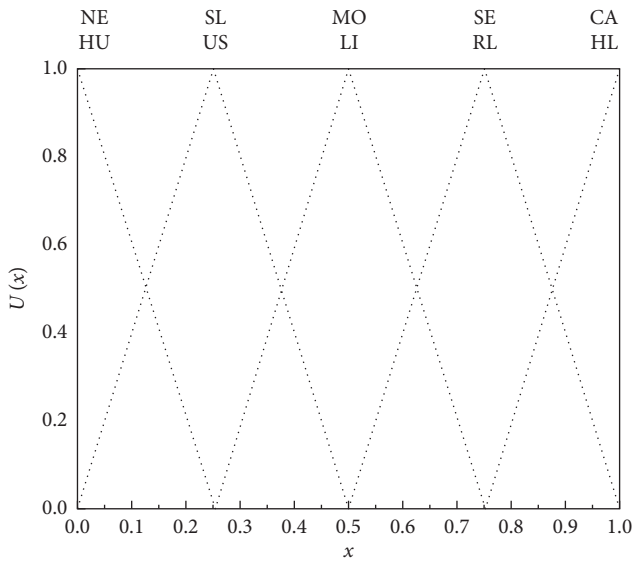


Figure 3: Triangular fuzzy membership function.

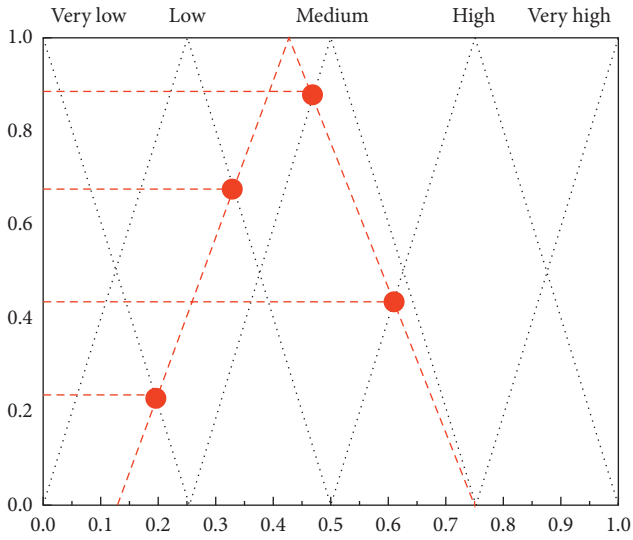


Figure 4: An example of mapping $\mathbf{FTN_{LS}}$ to $\mathbf{FTN_P}$.

denoting the degree to which the i-th basic factor e_i supports the general factor y to be evaluated as the n-th grade:

$$m_{n,i} = \omega_i \beta_{n,i}, \quad n = 1, \ldots, N. \tag{9}$$

The unassigned probability mass $m_{H,i}$ is composed of two parts, which represent the unassigned mass function $\overline{m}_{H,i}$ due to the weight and the unassigned mass function $\tilde{m}_{H,i}$ due to the lack of information and incompleteness:

$$m_{H,i} = 1 - \sum_{n=1}^{N} m_{n,i} = 1 - \omega_i \sum_{n=1}^{N} \beta_{n,i}, \tag{10}$$

$$m_{H,i} = \overline{m}_{H,i} + \tilde{m}_{H,i}, \tag{11}$$

$$\overline{m}_{H,i} = 1 - \omega_i, \tag{12}$$

$$\tilde{m}_{H,i} = \omega_i \left(1 - \sum_{n=1}^{N} \beta_{n,i} \right). \tag{13}$$

Suppose $m_{n,I(i+1)}$ represent the combined masses of i basic factors synthesized on the n-th evaluation grade. Suppose $m_{H,I(i+1)}$ represent the unassigned probability mass to the first i basic factors. The formula is as follows:

$$\{H_n\}: m_{n,I(i+1)} = K_{I(i+1)} \left[m_{n,I(i)} m_{n,i+1} + m_{H,I(i)} m_{n,i+1} + m_{n,I(i)} m_{H,i+1} \right], \tag{14}$$

$$\{H\}: \overline{m}_{H,I(i+1)} = K_{I(i+1)} \left[\overline{m}_{H,I(i)} + \overline{m}_{H,i+1} \right], \tag{15}$$

$$\{H\}: \tilde{m}_{H,I(i+1)} = K_{I(i+1)} \left[\tilde{m}_{H,I(i)} \tilde{m}_{H,i+1} + \overline{m}_{H,I(i)} \tilde{m}_{H,i+1} + \tilde{m}_{H,I(i)} \overline{m}_{H,i+1} \right], \tag{16}$$

$$K_{I(i+1)} = \left[1 - \sum_{t=1}^{N} \sum_{\substack{j=1 \\ j \neq 1}}^{N} m_{t,I(i)} m_{j,i+1} \right]^{-1}, \quad i = 1, \ldots, L-1, \tag{17}$$

where $K_{I(i+1)}$ represents the normalizing factor, which reflects the degree of conflict between the indicators (evidence). Suppose that there is a total of L basic factors for evaluation objectives; then, $m_{n,I(L)}$, $\overline{m}_{H,I(L)}$, and $\tilde{m}_{H,I(L)}$ are obtained by iteration calculation. After that, the combined belief degree can be obtained by the following normalization process:

$$\{H_n\}: \beta_n = \frac{m_{n,I(i)}}{1 - \overline{m}_{H,I(L)}}, \tag{18}$$

$$\{H\}: \beta_H = \frac{\tilde{m}_{H,I(L)}}{1 - \overline{m}_{H,I(L)}}, \tag{19}$$

where β_H represents the unassigned belief degree to the general factor y after aggregation. β_n and β_H represent the comprehensive belief degree to the evaluation object.

2.5. Obtaining a Clear Result Using the Expected Utility Method. In fact, the belief degree vector obtained in the former evaluation is the trust distribution of risk under the identification framework, and the result cannot be shown clearly. For example, the identification framework of a building fire risk (i.e., assessment set) is recorded as "very low," "low," "general," "high," and "very high." Suppose that the combined degree of the belief vector is (0,0.45,0.5,0.05,0), calculated by the above formula, which means that the construction risk level corresponds to a "low" level of 45%, a "general" level of 50%, and a "high" level of 5%. However, this information cannot clearly indicate the magnitude of the fire risk. Therefore, the concept of utility value is introduced in [35] as follows:

$$u(y) = \sum_{n=1}^{N} \beta_n u(H_n), \tag{20}$$

where $u(H_n)$ represents the utility of the evaluation grade H_n. In order to further clarify the level of the fire risk corresponding to the utility value, it is necessary to classify the grade of the fire risk. This paper presents the classification as shown in Table 3.

Quantitative evaluation results (utility values) can be obtained by processing the above methods. However, if the basic attribute (factor) information is incomplete or the expert's information about the factor is uncertain, the result obtained by the ER algorithm is also uncertain. [34, 36–38] refer to the concept of utility interval and conquer this problem through minimum utility $u_{\min}(y)$, maximum utility $u_{\max}(y)$, and average utility $u_{\text{avg}}(y)$:

$$\begin{aligned} u_{\min}(y) &= (\beta_1 + \beta_H) u(H_1) + \sum_{n=2}^{N} \beta_n u(H_n), \\ u_{\max}(y) &= \sum_{n=1}^{N-1} \beta_n u(H_n) + (\beta_N + \beta_H) u(H_N), \\ u_{\text{avg}}(y) &= \frac{u_{\max}(y) + u_{\min}(y)}{2}. \end{aligned} \tag{21}$$

2.6. Verification of the Model Using Sensitivity Analysis. Due to the influence of external factors, input values obtained from different experts or the same experts in different periods are different. Consequently, the uncertainty is inherent in fire risk assessment. In this paper, a sensitivity analysis method is introduced for studying and predicting the disturbance degree of the model output value (risk magnitude) caused by the change of the input value of each index. Sensitivity analysis is a systematic analysis method, which identifies weak points or areas in the system with the insight of managers in quantitative evaluation and continuously improves the design of the system and improves the stability of the system [39].

If the validated model is reliable and its reasoning process is logically feasible, then the sensitivity analysis of the model at least satisfies the following three theorems:

(1) A slight increase/decrease in the degrees of belief at any linguistic variables of the lowest-level factors will result in increase/decrease in the fire risk level of the output of the model

(2) If the belief degree at the lowest preference linguistic variable of the lowest-level factors increases by p and q (meanwhile, the belief degree at the highest preference linguistic variable decreases by p and q $(1 > q > p)$) and the utility values of the model output are u_p and u_q, then u_p should be greater than u_p

(3) In the lowest-level factors, the total influence of x factors on the output of the model is always greater than that of $x - y\,(y \in x)$ factor sets

3. Case Study

Three residential buildings marked from BUILDING-1 to BUILDING-3 were selected as a case study to illustrate the proposed fire risk model. This paper takes BUILDING-1, for example, to describe the calculation process of the model step by step. Based on the hierarchical structure of the fire risk model in Figure 2 and the available information in [40], the fire risk of BUILDING-1can be assessed through the following steps.

3.1. Develop a Generic Fire Risk Model for BUILDING-1. At this phase, the identified fire risk factors and a generic fire risk model are presented in Figure 1. The index system of fire risk assessment mainly consists of three levels, including the total target risk, the first-level factor set, and the second-level factor set. According to Wang et al. [41], fuzzy linguistic terms for risk expression are used for effective information processing in the range of 4 to 7. Therefore, this study uses five linguistic terms to denote the assessment of fire risk based on the viewpoint of experts in the field.

3.2. Determine the Weights of Each Factor. Given the hierarchical structure of fire risk in Figure 2, the weight calculations for fire risk factors are conducted. The weight calculations of factors U1, U2, U3, U4, and U5 are taken as an example. Firstly, the judgment matrix is constructed through the pairwise comparison of these five factors by experts (according to Table 1) and presented in Table 4**.** Then, according to equations (3) and (4), the value of synthetic extent S_i and the possibility degree $d'(A_i)$ of each factor are obtained, respectively. Finally, the normalized weight vector for five factors is obtained. Using a similar way, the weights of all factors can be calculated and listed in Table 5.

3.3. Application of the Belief Structure for Fire Risk in BUILDING-1. According to the actual situation of BUILDING-1 fire safety, the occurrence likelihood of fire (L) and the consequence severity of fire (S) for each bottom index should be scored, and the scoring standards are mainly based on the code for fire-protection design of buildings (GB50016-2014) [42], code for fire prevention in design of

Table 3: Classification of the building fire risk level.

Fire risk level	Risk interval	Risk description and measures
Very low	(0.00, 0.06]	Risks are negligible.
Low	(0.06, 0.25]	Risks are acceptable, but if cost-effectiveness is reasonable, measures can be taken to reduce risks.
Moderate	(0.25, 0.44]	Risks are tolerable and, if feasible, measures must be taken to reduce them.
High	(0.44, 0.72]	Measures must be taken to reduce risks.
Very high	(0.72, 1.00]	Risk is unacceptable. Measures must be taken to reduce the risk and control it effectively.

Table 4: Triangular fuzzy judgment matrix of indexes U1–U5.

R	U1	U2	U3	U4	U5
Expert 1	(1,1,2)	(2,3,4)	(2,3,4)	(2,3,4)	(1,2,3)
Expert 2	(1,1,2)	(2,3,4)	(1,2,3)	(1,2,3)	(3,4,5)
Expert 3	(1,1,2)	(3,4,5)	(2,3,4)	(2,3,4)	(1,2,3)
Expert 4	(1,1,2)	(2,3,4)	(2,3,4)	(1,2,3)	(1,2,3)
Expert 5	(1,1,2)	(2,3,4)	(2,3,4)	(2,3,4)	(3,4,5)
Expert 6	(1,1,2)	(3,4,5)	(1,2,3)	(2,3,4)	(1,2,3)
U1	(1,1,2)	(2.33,3.33,4.33)	(1.67,2.67,3.67)	(1.67,2.67,3.67)	(1.67,2.67,3.67)
Expert 1	(0.25,0.33,0.5)	(1,1,2)	(0.33,0.5,1)	(1,1,2)	(1,1,2)
Expert 2	(0.25,0.33,0.5)	(1,1,2)	(0.25,0.33,0.5)	(1,1,2)	(1,2,3)
Expert 3	(0.2,0.25,0.33)	(1,1,2)	(0.5,1,1)	(1,1,2)	(0.5,1,1)
Expert 4	(0.25,0.33,0.5)	(1,1,2)	(0.33,0.5,1)	(1,1,2)	(0.5,1,1)
Expert 5	(0.25,0.33,0.5)	(1,1,2)	(0.25,0.33,0.5)	(1,1,2)	(1,2,3)
Expert 6	(0.2,0.25,0.33)	(1,1,2)	(0.5,1,1)	(1,1,2)	(1,1,2)
U2	(0.23,0.31,0.44)	(1,1,2)	(0.36,0.61,0.83)	(1,1,2)	(0.83,1.33,2)
Expert 1	(0.25,0.33,0.5)	(1,2,3)	(1,1,2)	(1,1,2)	(1,2,3)
Expert 2	(0.33,0.5,1)	(2,3,4)	(1,1,2)	(1,2,3)	(1,2,3)
Expert 3	(0.25,0.33,0.5)	(1,1,2)	(1,1,2)	(1,2,3)	(1,1,2)
Expert 4	(0.25,0.33,0.5)	(1,2,3)	(1,1,2)	(1,1,2)	(1,1,2)
Expert 5	(0.25,0.33,0.5)	(2,3,4)	(1,1,2)	(1,2,3)	(1,2,3)
Expert 6	(0.33,0.5,1)	(1,1,2)	(1,1,2)	(1,2,3)	(1,2,3)
U3	(0.28,0.39,0.67)	(1.33,2,3)	(1,1,2)	(1,1.67,2.67)	(1,1.67,2.67)
Expert 1	(0.33,0.5,1)	(0.5,1,1)	(0.5,1,1)	(1,1,2)	(1,1,2)
Expert 2	(0.25,0.33,0.5)	(0.5,1,1)	(0.5,1,1)	(1,1,2)	(1,1,2)
Expert 3	(0.33,0.5,1)	(0.5,1,1)	(0.33,0.5,1)	(1,1,2)	(0.5,1,1)
Expert 4	(0.25,0.33,0.5)	(0.5,1,1)	(0.5,1,1)	(1,1,2)	(0.5,1,1)
Expert 5	(0.33,0.5,1)	(0.5,1,1)	(0.5,1,1)	(1,1,2)	(1,1,2)
Expert 6	(0.33,0.5,1)	(0.5,1,1)	(0.33,0.5,1)	(1,1,2)	(1,1,2)
U4	(0.28,0.39,0.67)	(0.5,1,1)	(0.33,0.5,1)	(1,1,2)	(0.83,1,1.67)
Expert 1	(0.33,0.5,1)	(0.5,1,1)	(0.33,0.5,1)	(0.5,1,1)	(1,1,2)
Expert 2	(0.2,0.25,0.33)	(0.33,0.5,1)	(0.33,0.5,1)	(0.5,1,1)	(1,1,2)
Expert 3	(0.33,0.5,1)	(1,1,2)	(0.5,1,1)	(1,1,2)	(1,1,2)
Expert 4	(0.33,0.5,1)	(1,1,2)	(0.5,1,1)	(1,1,2)	(1,1,2)
Expert 5	(0.2,0.25,0.33)	(0.33,0.5,1)	(0.33,0.5,1)	(0.5,1,1)	(1,1,2)
Expert 6	(0.33,0.5,1)	(0.5,1,1)	(0.33,0.5,1)	(0.5,1,1)	(1,1,2)
U5	(0.29,0.42,0.78)	(0.61,0.83,1.33)	(0.39,0.67,1)	(0.67,1,1.33)	(1,1,2)

interior decoration of buildings (GB 50222-2017) [43], guidance on building fire risk assessment for property insurance, and CIB W14 Workshop Report [44]. For example, the detailed scoring rules of index U15 (building service life) and index U4 (property fire management) are shown in Table 6. According to these rules, it is easy to obtain the value of FTN_L and FTN_s of each bottom index. Accordingly, by utilising equation (8), the fire risk of each bottom index is presented in Table 7 in the form of FTN_{LS}. Then, FTN_{LS} is mapped to FTN_P for obtaining the intersection point. Finally, the basic belief degree β is obtained after the normalization of $\beta_{\mathbf{p}}$, and the results are shown in Table 8.

3.4. Synthesizing Assessment Result Using the Evidence Reasoning Algorithm. On the premise that the weight of each index was obtained, the aggregation calculations for U11, U12, U13, U14, U15, and U16 were implemented according to the D-S operator (equations (9)–(20)); then, the aggregation result of disaster-causing factor U1 is obtained. Similarly, the aggregation results of passive measures U2, active measures U3, property fire management U4, the rescue capability of fire brigade U5 and the objective fire risk R can also be obtained, and the results of the first-level index are presented in Table 9.

Table 5: Weights of fire risk factors.

Fire risk factors	Abbreviation	Weights
Disaster-causing factors	U1	0.364
Passive measures	U2	0.149
Active measures	U3	0.244
Property fire management	U4	0.122
The rescue capability of fire brigade	U5	0.121
Electrical equipment	U11	0.276
Occupant density	U12	0.236
Gas use mode	U13	0.222
Interior decoration	U14	0.117
Building service life	U15	0.110
Ambient	U16	0.049
Fire resistance rating	U21	0.476
Fire compartment	U22	0.205
Safe evacuation	U23	0.265
Fire separation distance	U24	0.054
Indoor hydrant water supply system	U31	0.331
Portable fire-extinguisher apparatus	U32	0.366
Safety monitoring system	U33	0.303
Fire lane	U51	0.317
Fighting capability of fire brigade	U52	0.367
Water supply system of outdoor fire hydrant	U53	0.316

Table 6: Detailed scoring rules of U15 and U4.

Indexes	Detailed scoring rules	Score (L)	Score (S)	Remarks
U15-building service life	(1) $0 <$ service life $< n/5$	(0.00, 0.00, 0.25)	(0.00, 0.00, 0.25)	n is the design life
	(2) $n/5 <$ service life $\leq 2n/5$	(0.00, 0.25, 0.50)	(0.00, 0.25, 0.50)	
	(3) $2n/5 <$ service life $\leq 3n/5$	(0.25, 0.50, 0.75)	(0.25, 0.50, 0.75)	
	(4) $3n/5 <$ service life $\leq 4n/5$	(0.50, 0.75, 1.00)	(0.50, 0.75, 1.00)	
	(5) Service life $< 4n/5$	(0.75, 1.00, 1.00)	(0.75, 1.00, 1.00)	
U4-property fire management	(1) Four aspects are perfect	(0.25, 0.50, 0.75)	(0.25, 0.50, 0.75)	Including four aspects: (1) There are full-time fire safety management personnel with prejob training (2) The system of fire management is established, and the responsibility is clear (3) The hidden danger is checked and recorded every day (4) Regular inspection of fire facilities and timely maintenance
	(2) Any aspects are not perfect	(0.25, 0.50, 0.75)	(0.25, 0.50, 0.75)	
	(3) All four aspects are not perfect	(0.50, 0.75, 1.00)	(0.50, 0.75, 1.00)	

Table 7: The fire risk of each factor.

Fire risk factors	FTN_L	FTN_S	FTN_{LS}
U11	(0,0,0.25)	(0,0.25,0.5)	(0,0,0.13)
U12	(0.25,0.5,0.75)	(0.25,0.5,0.75)	(0.06,0.25,0.56)
U13	(0,0.25.0.5)	(0.25,0.5,0.75)	(0,0.06,0.25)
U14	(0.5,0.75,1)	(0.5,0.75,1)	(0.25,0.56,1)
U15	(0.25,0.5,0.75)	(0.25,0.5,0.75)	(0.06,0.25,0.56)
U16	(0,0,0.25)	(0,0,0.25)	(0,0,0.06)
U21	(0.25,0.5,0.75)	(0.25,0.5,0.75)	(0.06,0.25,0.56)
U22	(0.25,0.5,0.75)	(0,0.25.0.5)	(0,0.13,0.38)
U23	(0.25,0.5,0.75)	(0.25,0.5,0.75)	(0.06,0.25,0.56)
U24	(0,0,0.25)	(0,0,0.25)	(0,0,0.06)
U31	(0.25,0.5,0.75)	(0.25,0.5,0.75)	(0.06,0.25,0.56)
U32	(0.25,0.5,0.75)	(0,0.25.0.5)	(0,0.13,0.38)
U33	(0,0.25.0.5)	(0,0.25.0.5)	(0,0.06,0.25)
U4	(0.25,0.5,0.75)	(0.25,0.5,0.75)	(0.06,0.25,0.56)
U51	(0.25,0.5,0.75)	(0,0,0.25)	(0,0,0.19)
U52	(0.25,0.5,0.75)	(0,0.25.0.5)	(0,0.13,0.38)
U53	(0.25,0.5,0.75)	(0,0.25.0.5)	(0,0.13,0.38)

Table 8: Intersection results of fire risk factors.

Fire risk factors	β_P					β				
	VL	**L**	**M**	**H**	**VH**	**VL**	**L**	**M**	**H**	**VH**
U11	1	0.33	0	0	0	0.75	0.25	0	0	0
U12	0.25	0.75	0.8	0.4	0	0.11	0.34	0.36	0.18	0
U13	0.67	0.75	0.25	0	0	0.4	0.45	0.15	0	0
U14	0	0.44	0.89	0.73	0.36	0	0.18	0.37	0.3	0.15
U15	0.43	1	0.56	0.11	0	0.2	0.48	0.27	0.05	0
U16	0.8	0.57	0	0	0	0.58	0.42	0	0	0
U21	0.43	1	0.56	0.11	0	0.2	0.48	0.27	0.05	0
U22	0.67	0.75	0.25	0	0	0.4	0.45	0.15	0	0
U23	0.43	1	0.56	0.11	0	0.2	0.48	0.27	0.05	0
U24	1	0.2	0	0	0	0.83	0.17	0	0	0
U31	0.43	1	0.56	0.11	0	0.2	0.48	0.27	0.05	0
U32	0.67	0.75	0.25	0	0	0.4	0.45	0.15	0	0
U33	0.8	0.57	0	0	0	0.58	0.42	0	0	0
U4	0.43	1	0.56	0.11	0	0.2	0.48	0.27	0.05	0
U51	1	0.43	0	0	0	0.7	0.3	0	0	0
U52	0.67	0.75	0.25	0	0	0.4	0.45	0.15	0	0
U53	0.67	0.75	0.25	0	0	0.4	0.45	0.15	0	0

Table 9: Aggregation of fire risk factors.

Fire risk factors	Very low	Low	Moderate	High	Very high
Disaster-causing factors U1	0.3875	0.3515	0.1753	0.0721	0.0136
Passive measures U2	0.2555	0.4841	0.2234	0.0370	0
Active measures U3	0.3898	0.4679	0.1277	0.0146	0
Property fire management U4	0.2045	0.4773	0.2652	0.0530	0
The rescue capability of fire brigade U5	0.5099	0.4015	0.0886	0	0
Objective fire risk R	0.3644	0.4332	0.1572	0.0383	0.0049

3.5. The Target Fire Risk Assessment Using the Expected Utility Method. From Table 9, the objective fire risk R corresponding to five-level linguistic terms can be expressed as $R = \{\text{VL } (0.36), \text{L } (0.43), \text{M } (0.16) \text{ H } (0.04), \text{VH } (0.005)\}$, which cannot reveal the magnitude of the target fire risk in a clear way. Thus, the final fire risk (FR) is evaluated using equation (21), and the result is 0.2205, shown in Table 10. According to Table 3, it could be observed that the objective fire risk R is acceptable, but if cost-effectiveness is reasonable, measures can be taken in this building to reduce its fire risks.

3.6. Sensitivity Analysis. In order to verify the model, the degrees of belief at the lowest preference linguistic variable of the lowest-level factors should increase by 10%, 20%, and 30% (meanwhile, the degrees of belief at the highest preference linguistic variable decrease by 10%, 20%, and 30%). The model output data are tabulated in Table 11, and the graphic display results are listed in Figure 5. It is obvious that all the results are consistent with theorems 1 and 2, respectively. According to theorem 3, if the model is logically reasonable and feasible, the belief degree at the lowest level of the hierarchy structure associated with x factors will always be smaller than the one associated with $x - y$ ($y \in x$) factors. This can be illustrated by comparing the results of different input data, such as if the belief degree at the lowest preference linguistic variable associated with all the lowest-level factors increases by 10% (simultaneously, the one at the highest preference linguistic variable decreases by 10%), the output utility value is 0.1717. However, if the belief degree at the lowest preference linguistic variable associated with U11, U12, U13, U14, U15, U16, U21, U22, U23, U24, U31, U32, and U33 factors increases by 10% (simultaneously, the one at the highest preference linguistic variable decreases by 10%), the output utility value is 0.1826. Considering that 0.1717 is smaller than 0.1826, it can be concluded that the verified model satisfies theorem 3.

4. Results and Discussion

Based on the results of the case study in Table 12 and Figure 6, it can be observed that the fire risk level of three buildings corresponds to "moderate" or below. However, it is noteworthy to mention that some aspects should be paid attention to.

In the aspect of disaster-causing factor U1: since the service life of the three residential buildings is less than 10 years and there is no dangerous disaster-causing factor in the internal and external environment of these buildings, the fire risk corresponding to U1 of these buildings is all acceptable.

In the aspect of passive measures U2: U2 of BUILDING-2 and BUILDING-3 was higher than that of BUILDING-1. This was mainly due to obstruction of safe evacuation in the

Table 10: Utility value for measuring the building fire risk.

Rating H_n	Very low	Low	Moderate	High	Very high
V_n	1	2	3	4	5
$u(H_n)$	0	0.25	0.5	0.75	1
β_n	0.3664	0.4332	0.1572	0.0383	0.0049
$\beta_n \times u(H_n)$	0	0.1083	0.0786	0.0287	0.0049
FR = $\sum_{n=1}^{N} \beta_n \times u(H_n) = 0.2205$					

Table 11: Increase/decrease model input data.

Increase the input data at the lowest preference linguistic variable; meanwhile, decrease the input data at the highest preference linguistic variable			
Fire risk factors	10%	20%	30%
U12	0.2116	0.2054	0.1997
U11	0.2123	0.2069	0.2026
U4	0.2128	0.2076	0.2041
U32	0.2138	0.2094	0.2050
U31	0.2142	0.210	0.2069
U21	0.2146	0.2108	0.2082
U14	0.2155	0.2121	0.2089
U13	0.2163	0.2132	0.2102
U33	0.2168	0.2143	0.2118
U52	0.2170	0.2152	0.2134
U23	0.2172	0.2153	0.2138
U53	0.2174	0.2160	0.2147
U22	0.2179	0.2168	0.2156
U51	0.2180	0.2168	0.2160
U16	0.2181	0.2170	0.2161
U15	0.2182	0.2173	0.2164
U24	0.2191	0.2184	0.2177

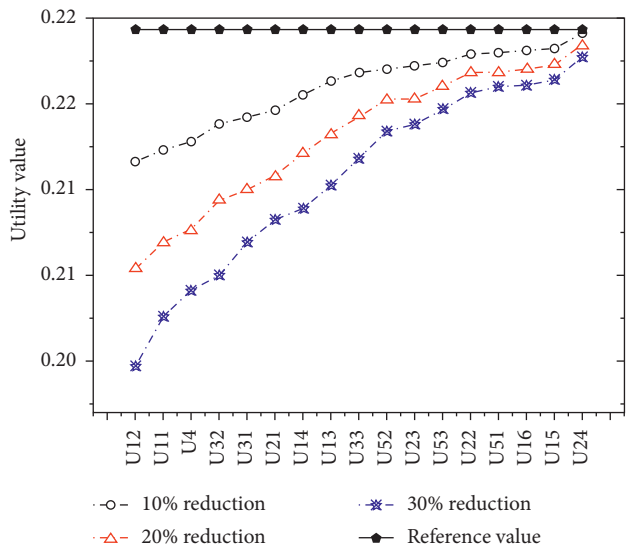


Figure 5: Sensitivity analysis of model output data.

stairwell of BUILDING-2 and BUILDING-3, such as some evacuation passageways are littered with debris and some safety evacuation signs are missing, which mean that the residents may fail to evacuate from these buildings in case of a fire.

Table 12: Fire risk levels of three buildings.

Target building	U1	U2	U3	U4	U5	U
BUILDING-1	0.2432	0.2604	0.1918	0.2916	0.1447	0.2205
BUILDING-2	0.2264	0.3804	0.1921	0.4034	0.2691	0.2511
BUILDING-3	0.2398	0.3779	0.3367	0.6042	0.2558	0.3073

In the aspect of active measures U3 and property fire management U4: U3 and U4 of BUILDING-3 were higher than those of BUILDING-1 and BUILDING-2. This was mainly due to the lack of regular maintenance of fire-fighting equipment in BUILDING-3. It can be assured that if there is no regular maintenance and inspection, the reliability of fire-fighting equipment will be reduced. In BUILDING-3, it was found that some fire-fighting equipment were rusty or even abandoned, such as safety monitoring device was out of use, and the fire extinguisher was out of the service date range. In addition, U4 of BUILDING-2 was higher than that of BUILDING-1. This is mainly because that, in BUILDING-2, there is no prejob training of safety management personnel, and daily fire hazard investigation is not carried out.

In the aspect of the rescue capability of fire brigade U5: BUILDING-2 and BUILDING-3 are all located in CBD of the city. It means that the traffic around the buildings is congested, and the nearby fire brigade may not be able to arrive in time. In particular, BUILDING-2 is further away from the fire brigade than BUILDING-3.

Furthermore, the fire risk of residential buildings is determined by many factors in the complex external environment. It is noteworthy from the analysis that a slight change will lead to the corresponding change in the output value of the model. According to Figure 5, it is obvious that the fire risk model is more sensitive to occupant density U12, electrical equipment U11, property fire management U4, portable fire-extinguisher apparatus U32, and indoor hydrant water supply system U31 than other factors. In other words, the uncertainty of these factors has a relatively large influence on the disturbance of the model system. Therefore, the most effective way to reduce the fire risk of residential buildings is to control these five indicators at first. The analysis results are also consistent with the actual fire prevention measures.

In the previous studies [40, 45], grey correlation method and fuzzy clustering method were applied for fire risk assessment in these buildings of China, and the results of these studies are in accordance with the results of our research, which indicated that the presented model is logically feasible and can still maintain its specific function under turbulence or uncertainty conditions.

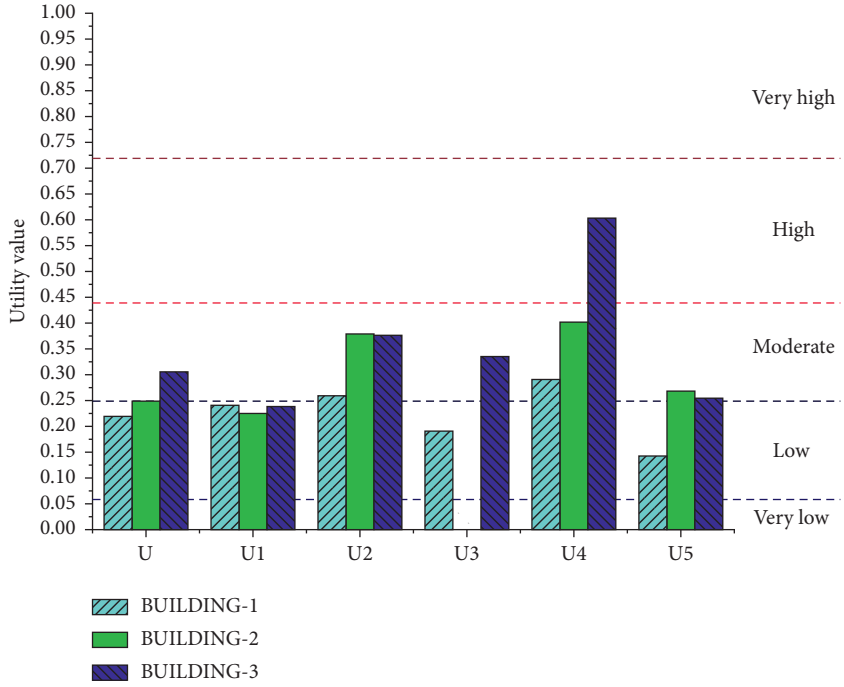


Figure 6: Utility values of the main factors.

5. Conclusions

This study proposes a novel model which combines evidence theory, fuzzy theory, and sensitivity analysis technique for assessing the building fire risk using inaccurate input data in order to optimize system operating efficiency by a standardized fuzzy linguistic term. This model is different from the traditional risk assessment model and characterized with flexible data acquisition capability and unified input and output modes. Therefore, it is easy to deal with the uncertainty of the fire risk problem in the complex system.

Furthermore, the model adopts a series of processes, such as weight calculation based on the FAHP, two-dimensional measurement of the fire risk based on triangular fuzzy numbers, construction of the belief structure, factor aggregation via the evidential reasoning algorithm, and assessment results using the expected utility method, to effectively address uncertainties of subjective estimation. In summary, the proposed model has the following advantages for fire risk analysis on the complex system: (1) this model presents a managerial view to analysts in a reasonable, reliable, and transparent way so that they can collaborate with experts' suggestion or on-site investigation to model complex systems under external uncertainties. (2) The model provides an effective tool for researchers to make full use of limited information to assess the fire risk of the whole system and improve its operational flexibility. (3) The model has strong flexibility, has high robustness, and is easy to program. It can be used as a computer tool for fire risk assessment of complex systems under high uncertainty.

This research proposes a quantitative fire risk assessment model which could provide building fire managers and researchers with flexible and transparent tools to effectively reduce the fire risk under the disturbance of fire risk uncertainty of the system. It should be noted that, in our study, the index scoring rules are mainly based on codes and standards, which lead to conservative results. Therefore, the acceptable level of fire risk based on performance-based codes needs to be determined in the future [46].

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC, 51704054 and 51874255), Key Technologies for Prevention and Control of Serious and Extraordinary Accidents of Ministry of Emergency Management (no. Chongqing-0001-2018AQ), the Natural Science Foundation of Chongqing (cstc2019jcyj-msxmX0462), and Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJQN201801531 and KJQN201801519).

References

[1] J. M. Watts and J. R. Hall, "Introduction to fire risk analysis," in *SFPE Handbook of Fire Protection Engineering*, pp. 2817–2826, Springer, New York, NY, USA, 2016.

[2] J. Xin and C. Huang, "Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management," *Fire Safety Journal*, vol. 62, pp. 72–78, 2013.

[3] G. Hadjisophocleous and Z. Fu, "Literature review of fire risk assessment methodologies," *International Journal on Engineering Performance-Based Fire Codes*, vol. 6, no. 1, pp. 28–45, 2004.

[4] D. Yung, "Cost-effective fire-safety upgrade options for a Canadian government office building," in *Proceedings of the 1996 International Conference on Performance-Based Codes and Fire Safety Design Methods*, Ottawa, Canada, 1996.

[5] J. J. H. P. Watts, "Systematic methods of evaluating fire safety: a review," *Hazard Prevention*, vol. 18, no. 2, pp. 24–27, 1981.

[6] N. Benichou, "FIERAsystem: a fire risk assessment tool to evaluate fire safety in industrial buildings and large spaces," *Journal of Fire Protection Engineering*, vol. 15, no. 3, pp. 145–172, 2005.

[7] Y. He and V. Beck, "A computer model for smoke spread in multi-storey buildings," in *Proceedings of the 8th International Symposium on Transport Phenomena in Combustion*, San Francisco, CA, USA, July 1995.

[8] V. Beck, "CESARE-RISK: a tool for performance-based fire engineering design," in *Proceedings of 2nd International Conference on Performance-Based Codes and Fire Safety Design Methods*, Maui, HI, USA, 1998.

[9] J. Fraser-Mitchell, "An object-oriented simulation (crisp 11) for fire risk assessment," *Fire Safety Science*, vol. 4, pp. 793–804, 1994.

[10] M. N. Ibrahim, K. Abdul-Hamid, M. S. Ibrahim, A. Mohd-Din, R. M. Yunus, and M. R. Yahya, "The development of fire risk assessment method for heritage building," *Procedia Engineering*, vol. 20, pp. 317–324, 2011.

[11] S. M. Lo, "A fire safety assessment system for existing buildings," *Fire Technology*, vol. 35, no. 2, pp. 131–152, 1999.

[12] F. Liu, S. Zhao, M. Weng, and Y. Liu, "Fire risk assessment for large-scale commercial buildings based on structure entropy weight method," *Safety Science*, vol. 94, pp. 26–40, 2017.

[13] F. I. Khan, R. Sadiq, and T. Husain, "Risk-based process safety assessment and control measures design for offshore process facilities," *Journal of Hazardous Materials*, vol. 94, no. 1, pp. 1–36, 2002.

[14] M. Kalantarnia, F. Khan, and K. Hawboldt, "Dynamic risk assessment using failure assessment and Bayesian theory," *Journal of Loss Prevention in the Process Industries*, vol. 22, no. 5, pp. 600–606, 2009.

[15] A. John, D. Paraskevadakis, A. Bury, Z. Yang, R. Riahi, and J. Wang, "An integrated fuzzy risk assessment for seaport operations," *Safety Science*, vol. 68, pp. 180–194, 2014.

[16] National Fire Protection Association, *NFPA 550: Guide to the Fire Safety Concepts Tree*, National Fire Protection Association, Quincy, MA, USA, 2007.

[17] M. Omidvari, N. Mansouri, and J. Nouri, "A pattern of fire risk assessment and emergency management in educational center laboratories," *Safety Science*, vol. 73, pp. 34–42, 2015.

[18] G. Chen and X. Zhang, "Fuzzy-based methodology for performance assessment of emergency planning and its application," *Journal of Loss Prevention in the Process Industries*, vol. 22, no. 2, pp. 125–132, 2009.

[19] A. Grassi, R. Gamberini, C. Mora, and B. Rimini, "A fuzzy multi-attribute model for risk evaluation in workplaces," *Safety Science*, vol. 47, no. 5, pp. 707–716, 2009.

[20] M. Kobes, I. Helsloot, B. de Vries, and J. G. Post, "Building safety and human behaviour in fire: a literature review," *Fire Safety Journal*, vol. 45, no. 1, pp. 1–11, 2010.

[21] D.-Y. Chang, "Applications of the extent analysis method on fuzzy AHP," *European Journal of Operational Research*, vol. 95, no. 3, pp. 649–655, 1996.

[22] P. J. M. van Laarhoven and W. Pedrycz, "A fuzzy extension of Saaty's priority theory," *Fuzzy Sets and Systems*, vol. 11, no. 1-3, pp. 229–241, 1983.

[23] D.-Y. Chang, "Extent analysis and synthetic decision," *Optimization Techniques and Applications*, vol. 1, no. 1, pp. 352–355, 1992.

[24] M. An, S. Huang, and C. J. Baker, "Railway risk assessment—the fuzzy reasoning approach and fuzzy analytic hierarchy process approaches: a case study of shunting at Waterloo depot," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 221, no. 3, pp. 365–383, 2007.

[25] J. J. Buckley, "Fuzzy hierarchical analysis," *Fuzzy Sets and Systems*, vol. 17, no. 3, pp. 233–247, 1985.

[26] Y. Li and X. Liao, "Decision support for risk analysis on dynamic alliance," *Decision Support Systems*, vol. 42, no. 4, pp. 2043–2059, 2007.

[27] D. J. C. E. P. Simon, "Fuzzy sets and fuzzy logic," *Theory and Applications*, vol. 9, no. 4, pp. 1332-1333, 1996.

[28] E. W. T. Ngai and F. K. T. Wat, "Fuzzy decision support system for risk analysis in e-commerce development," *Decision Support Systems*, vol. 40, no. 2, pp. 235–255, 2005.

[29] R. Sadiq, E. Saint-Martin, and Y. Kleiner, "Predicting risk of water quality failures in distribution networks under uncertainties using fault-tree analysis," *Urban Water Journal*, vol. 5, no. 4, pp. 287–304, 2008.

[30] L. Krishnasamy, F. Khan, and M. Haddara, "Development of a risk-based maintenance (RBM) strategy for a power-generating plant," *Journal of Loss Prevention in the Process Industries*, vol. 18, no. 2, pp. 69–81, 2005.

[31] G. Shafer, *A Mathematical Theory of Evidence*, Vol. 42, Princeton University Press, Princeton, NJ, USA, 1976.

[32] R. L. D. Mantaras, *Approximate Reasoning Models*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1990.

[33] R. R. Yager, "On the determination of strength of belief for decision support under uncertainty-part II: fusing strengths of belief," *Fuzzy Sets and Systems*, vol. 142, no. 1, pp. 129–142, 2004.

[34] J.-B. Yang and D.-L. Xu, "On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 32, no. 3, pp. 289–304, 2002.

[35] Y. Jian-Bo and M. G. Singh, "An evidential reasoning approach for multiple-attribute decision making with uncertainty," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 1, pp. 1–18, 1994.

[36] J.-B. Yang, "Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties," *European Journal of Operational Research*, vol. 131, no. 1, pp. 31–61, 2001.

[37] Y.-M. Wang, J.-B. Yang, and D.-L. Xu, "Environmental impact assessment using the evidential reasoning approach," *European Journal of Operational Research*, vol. 174, no. 3, pp. 1885–1913, 2006.

[38] D.-L. Xu and J.-B. Yang, "Intelligent decision system for self-assessment," *Journal of Multi-Criteria Decision Analysis*, vol. 12, no. 1, pp. 43–60, 2003.

[39] S. Contini, S. Scheer, and M. Wilikens, "Sensitivity analysis for system design improvement," in *Proceedings International Conference on Dependable Systems and Networks, 2000*, New York, NY, USA, June 2000.

[40] X. Wu, *Residential Fire Risk Assessment and Control Strategies*, Central South University, Changsha, China, 2014.

[41] J. Wang, J. B. Yang, and P. Sen, "Safety analysis and synthesis using fuzzy sets and evidential reasoning," *Reliability Engineering & System Safety*, vol. 47, no. 2, pp. 103–118, 1995.
[42] GB50016-2014, *Code for Fire Protection Design of Buildings*, China Planning Press, Beijing, China, 2007.
[43] GB 50222-2017, *Code for Fire Prevention in Design of Interior Decoration of Buildings*, Ministry of Construction of the People's Republic of China, Beijing, China, 2001.
[44] P. H. Thomas, "Design guide: structure fire safety CIB W14 workshop report," *Fire Safety Journal*, vol. 10, no. 2, pp. 77–137, 1986.
[45] L. Zhang, *High-Rise Building Fire Risk Evaluation and Intelligent Alarm System Research*, Beijing Institute of Technology, Beijing, China, 2015.
[46] V. Kodur, P. Kumar, and M. M. Rafi, "Fire hazard in buildings: review, assessment and strategies for improving fire safety," *PSU Research Review*, vol. 4, no. 1, pp. 1–23, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3753274, 14 pages
https://doi.org/10.1155/2020/3753274

Research Article

Application of Model-Based Deep Learning Algorithm in Fault Diagnosis of Coal Mills

Yifan Jian,[1] Xianguo Qing,[1] Yang Zhao,[1] Liang He,[1] and Xiao Qi[2]

[1]*Science and Technology on Reactor System Design Technology Laboratory, Nuclear Power Institute of China, Chengdu, China*
[2]*Energy and Electricity Research Center, Jinan University, Zhuhai, Guangdong, China*

Correspondence should be addressed to Xiao Qi; qixiao@jnu.edu.cn

Received 5 May 2020; Revised 3 July 2020; Accepted 20 July 2020; Published 14 August 2020

Guest Editor: Mohammed Abouheaf

The coal mill is one of the important auxiliary engines in the coal-fired power station. Its operation status is directly related to the safe and steady operation of the units. In this paper, a model-based deep learning algorithm for fault diagnosis is proposed to effectively detect the operation state of coal mills. Based on the system mechanism model of coal mills, massive fault data are obtained by analyzing and simulating the different types of faults. Then, stacked autoencoders (SAEs) are established by combining the said data with the deep learning algorithm. The SAE model is trained by the fault data, which provide it with the learning and identification capability of the characteristics of faults. According to the simulation results, the accuracy of fault diagnosis of coal mills based on SAE is high at 98.97%. Finally, the proposed SAEs can well detect the fault in coal mills and generate the warnings in advance.

1. Introduction

The coal mill is one of the important auxiliary equipment of coal-fired units, and its operating status is directly related to the safe and stable operation of the units. When a fault occurs in the coal mill, the fuel supply of the boiler cannot be guaranteed which creates the mismatch between boiler energy output and the turbine power output. Under this situation, a quick load rejection operation will occur, which directly leads to fire extinguishing in the furnace. The fault in the coal mill will cause large economic loss to power generation enterprises and decrease the safety and stability of the power system. Therefore, it is of great necessity to guarantee the normal operation through effective fault warning and diagnosing of coal mill.

Agrawal et al. [1] divided the fault diagnosis methods into three categories: model-, signal-, and historical operation data-based fault diagnosis methods. Model-based fault diagnosis methods need to establish the mathematical model of the coal mill. Odgaard and Mataji [2] used a simplified energy balance equation to monitor and diagnose abnormal energy flow in the coal mill. Andersen et al. [3] designed a Kalman filter to estimate the moisture in the coal that enters and exists a coal mill to determine whether the energy in the coal mill is in normal condition. Based on the multisegment model of coal mills established by Wei et al. [4], Guo et al. [5] realized the monitoring of the state of coal mills by identifying the abnormal variation in the model parameters. Model-based fault diagnosis methods analyze the mathematical model of the actual object for fault diagnosis, and thus, the physical meaning is clear. However, establishing the exact model in practical application is difficult [6–10]. Thus, the operability of these methods is poor.

Signal-based fault diagnosis systems are widely used to evaluate the health of mechanical equipment. Many signals of high frequency change during the operation of the coal mill, such as current of coal mill, outlet primary air flow of coal mill, and differential pressure of primary air. Su et al. [11] designed a system that records the vibration signals of the coal mill and shifts them to energy amplitudes by use of wavelet analysis. Whether the coal mill is in coal interruption or coal choking, other fault operations can be determined by analyzing the relationship between the vibration signals and the amount of coal in the mill. Kisić et al. [12] proposed a method to detect the wear degree of grinding roller and analyzed the multivariate control chart on

the frequency spectrum to find the appropriate time to replace the worn parts. Collura et al. [13] utilized model identification and signal processing techniques to develop a coal mill performance monitoring tool based on real-time detection of the fineness of pulverized coal. Compared with the model-based fault diagnosis method, the signal-based fault diagnosis method does not need to establish complex object model. Only through the analysis of collected data can a fault in the system be found. However, these methods often need to install a large number of sensors to collect signal, thereby resulting in high implementation and maintenance costs.

Fault diagnosis based on historical operation data is mainly done by analyzing the differences between the normal operation data and fault operation data to determine the health status of coal mill. Han and Jiang [14] proposed a fault diagnosis method based on fuzzy decision clustering and used a single-layer neural network to realize three kinds of fault identification of coal mill. Qin et al. [15] utilized the abnormal operation data of the coal mill to establish an expert system to determine the operating status of the coal mill by comparing the trend of the model output with the expert system. A data-based fault analysis method is a data-driven approach, and even researchers who are unfamiliar with the system can use relevant algorithms for analysis. However, fault types and fault data in the mass historical data of the thermal power units are incomplete and a data-driven method requires analyzing a large amount of fault data. Thus, selecting the fault data from the vast amount of historical data one by one is difficult [16–18].

The model-based fault diagnosis method needs to establish an accurate model of the coal mill in order to obtain good fault diagnosis results. However, the coal mill is a complex object with multiparameter coupling. It is difficult to establish an accurate mathematical model. The premise of applying signal-based fault diagnosis methods is to be able to measure the monitored parameters. Therefore, it is necessary to install a large number of new sensors on the shell of the coal mill. However, when the coal mill was initially constructed, it usually did not consider reserving the mechanical interface for new sensors. So, it is not easy to install new sensors on the shell of the coal mill. The fault diagnosis method based on historical operation data firstly needs to obtain a large amount of fault operation data of the coal mill. However, the fault data of the coal mill is usually mixed with the normal operation data, which is difficult to classify and identify. Based on the above analysis, the existing methods are difficult to achieve good application results for the fault diagnosis of coal mills. Although the above three types of traditional methods have shortcomings, combining their advantages can find a simpler and more effective method to solve the fault diagnosis of coal mills. The basic idea is to obtain fault simulation data based on a simplified model and use big data analysis for fault identification. In recent years, the rapid development of deep learning algorithms has provided the possibility of big data analysis. Guo et al. [19] constructed an adaptive convolution neural network, which greatly improves accuracy of fault diagnosis of motor bearing. Duan et al. [20] used a deep learning algorithm to study the missing traffic data to implement the interpolation of missing data. In this study, a model-based data-driven fault diagnosis method is proposed to obtain a fault diagnosis method with simple operation, low cost, and high accuracy. First, on the basis of the simplified coal mill model, massive fault data of the coal mill are obtained by analyzing the fault principle and simulating the fault operation status of the coal mill. This method solves the difficulty in obtaining a large amount of fault data manually from the massive data. Then, stacked autoencoders (SAEs) with multilayer neural networks are established on the basis of the theory of deep learning algorithm. The numerous fault data obtained by the steps above are used in the training of the networks to fully motivate the nonlinear characteristics of deep neural networks. Thus, the built network can accurately learn the essential characteristics of all kinds of faults and then achieve the early warning and diagnosis of the fault in coal mills. The above method can greatly improve the fault diagnosis accuracy of the coal mill, and at the same time, it can also provide fault warning to the operator, which is of great significance for ensuring the safe operation of the power plant and ensuring the safety of the equipment.

The rest of the paper is organized as follows. Section 2 introduces the working principle of the coal mill and its nonlinear dynamic model. Section 3 analyzes the mechanism of two typical coal mill faults and obtains a large number of fault data by simulation experiments. Section 4 introduces the working principle of the SAEs and makes certain improvements to the model. Section 5 is the simulation analysis that aims to verify the effectiveness of the proposed method in the fault diagnosis of the coal mill. Section 6 elaborates the conclusions of the study.

2. Brief Introduction of the Coal Mill System

2.1. Working Principle. MPS-type medium-speed coal mill [21] is a roller-type coal mill designed and manufactured by Babcock, Germany. Such mills are characterized by smooth output, low energy consumption, and long maintenance period. In this study, MPS180-HP-II medium-speed coal mill is used in the analysis. The maximum output is 44.496 t/h, and the fineness of coal powder R90 is 22% (Figure 1). R90 indicates the probability that coal powders cannot pass through a sieve with a pore size of 90 μm.

The raw coal falls into the coal mill through the coal dropping pipe and is milled into coal powder under the squeezing effect of two milling parts (grinding disks and rollers) [4]. The primary air enters the coal mill through the annulus around the grinding disk to dry the coal powders and bring them into the coarse coal separator for separation. The qualified fine coal powders are blown into the boiler for combustion while large ones return into the coal for subsequent milling.

2.2. Mathematical Model of the Coal Mill. The operation of the coal mill involves the mass balance of coal and the energy balance of the entire coal mill. Establishing an effective dynamic mathematical model of coal mills is an important prerequisite for the state monitoring of coal mills. Zeng et al. [22–24] established an MPS medium-speed mill model (equation (1)),

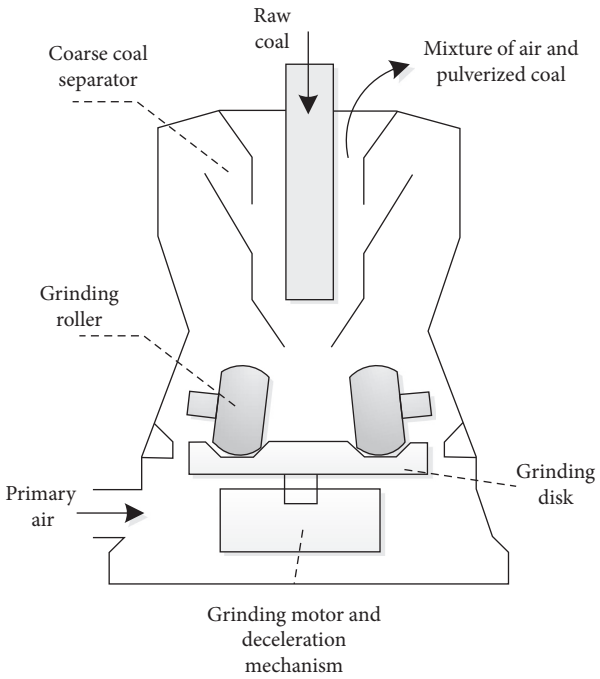


Figure 1: Schematic structure of the MPS medium-speed mill.

which includes three inputs and three outputs based on the mass and energy balance of the primary air and coal moisture in the mill. The proposed method is based on this model, and the symbolic description of the model is shown in Table 1:

$$
\begin{aligned}
\dot{W}_{\text{air}} &= \frac{1}{T_1}\left(-W_{\text{air}} + W_L^{\max}u_L + W_H^{\max}u_H\right),\\
\dot{T}_{\text{in}} &= \frac{1}{T_2}\left[-T_{\text{in}} + \frac{C_L W_L^{\max}u_L T_L + C_H W_H^{\max}u_H T_H}{C_{\text{in}}\left(W_L^{\max}u_L + W_H^{\max}u_H\right)}\right],\\
\dot{M}_c &= W_c - K_{10}M_c,\\
\dot{M}_{pf} &= K_{10}M_c - W_{pf},\\
W_{pf} &= K_{11}\Delta P_{pa}M_{pf},\\
\Delta P_{pa} &= \frac{22.4}{28.8}\cdot\frac{273+T_{\text{in}}}{273}\cdot\left(\frac{W_{\text{air}}}{10}\right)^2,\\
I &= K_6 M_{pf} + K_7 M_c + K_8,\\
\dot{T}_{\text{out}} &= \left[K_1 T_{\text{in}} + K_2\right]W_{\text{air}} + K_3 W_c\\
&\quad - \left[K_4 T_{\text{out}} + K_5\right]\left[W_{\text{air}} + W_c\right] + K_9 I\\
&\quad + K_{12}T_{\text{out}} - K_{14}W_{\text{free}}^{\text{water}},\\
\dot{M}_{pc} &= \frac{1}{M_c + M_{pf}}\left(M_{ar}W_c - W_{\text{free}}^{\text{water}} - M_{pc}W_{pf}\right),\\
W_{\text{free}}^{\text{water}} &= K_{13}\left(W_c M_{ar}\right)T_{\text{out}}\left(1 - e^{\left(W_{\text{air}}/K_{15}\right)}\right).
\end{aligned}
\tag{1}
$$

Table 1: Nomenclature.

Symbol	Meaning
$W_L^{\max}$	Maximum flow of cold air (kg/s)
$W_H^{\max}$	Maximum flow of hot air (kg/s)
u_L	Valve position of cold air
u_H	Valve position of hot air
T_L	Temperature of cold air (°C)
T_H	Temperature of hot air (°C)
ΔP_{pa}	Differential pressure of primary air (mbar)
C_{in}	Specific heat capacity of mixed primary air (kJ/(kg·°C))
C_L	Specific heat capacity of cold air (kJ/(kg·°C))
C_H	Specific heat capacity of hot air (kJ/(kg·°C))
T_{in}	Inlet primary air temperature of coal mill (°C)
W_{air}	Inlet primary air flow of coal mill (kg/s)
W_{pf}	Outlet pulverized coal flow of coal mill (kg/s)
M_{pf}	The mass of pulverized coal in coal mill (kg)
W_c	Inlet coal flow of coal mill (kg/s)
M_c	The mass of raw coal in coal mill (kg)
T_{out}	Outlet temperature of coal mill (°C)
$W_{\text{free}}^{\text{water}}$	Evaporation of coal moisture (kg/s)
M_{pc}	Pulverized coal (%)
M_{ar}	Coal moisture (%)
I	Current of coal mill (A)
T_1	Delay time of the value position to primary air flow (s)
T_2	Delay time of value position to primary air temperature (s)
K_i	Identified model parameters ($i = 1, 2, \ldots, 15$)

In Equation (1), U_L, U_H, and W_c are the control quantities of the model; W_{air}, T_{out}, and W_{pf} are the output quantities of the model; and K_i and T_i ($i = 1, 2, \ldots, 15, j = 1, 2$) are the model parameters to be identified and of which the values are shown in Table 2.

3. Model-Based Coal Mill Fault Simulation

The fault types and fault data in the vast amount of historical data of the thermal power units are incomplete, and selecting the fault data one by one from the massive historical data is difficult. Therefore, effectively obtaining a large number of fault data is the key to solve the fault diagnosis of coal mill. The simulation results in [22] showed that the mathematical model of MPS-type medium speed coal mill presents high precision. In the current study, the coal mill model is used in the analysis and two typical coal mill faults (coal interruption and coal choking) are simulated by analyzing the fault mechanism of coal mill. The simulation experiments obtain a large number of fault data, which can effectively solve the difficulty in obtaining fault data manually from the massive data.

First, a control scheme is designed for the coal mill model. The purpose is to ensure that the simulation experiments are conducted under the closed loop regulation, such that the fault data obtained by the simulation experiments can be significantly close to the real operation status of coal mill. The control scheme is shown in Figure 2. The entire control scheme consists of three controlled, three control, and four state variables as presented in Table 3. The control circuits are composed of three single-loop

Table 2: Identified model parameters.

Parameter	Value
K_1	0.00069
K_2	0.19549
K_3	0.00999
K_4	0.00109
K_5	0.09338
K_6	0.17999
K_7	0.88836
K_8	34.2065
K_9	0.01656
K_{10}	0.41378
K_{11}	0.07005
K_{12}	−0.05987
K_{13}	0.01152
K_{14}	0.26254
K_{15}	14.9867
T_1	10.3004
T_2	3.6765

proportion-integral-differential (PID) controllers. Specifically, PID1, where the setting parameters, respectively, are $K_p = 1$, $K_i = 0.05$, $K_d = 0$, controls the outlet temperature of coal mill by adjusting the valve position of cold air. PID2, where the setting parameters, respectively, are $K_p = 2$, $K_i = 0.5$, $K_d = 0$, controls the inlet primary air flow of coal mill by adjusting the valve position of hot air, and PID3, where the setting parameters, respectively, are $K_p = 0.1$, $K_i = 0.1$, $K_d = 0$, controls the outlet pulverized coal flow of coal mill by adjusting the inlet coal flow of coal mill. After designing the control scheme, the fault operation status of the coal mill can be simulated by adjusting the corresponding controllers.

3.1. Fault Simulation of Coal Interruption. When an obstruction exists in the coal dropping pipe or a fault occurs in the coal feeder, the amount of coal into the coal mill will reduce directly and coal interruption will occur when the case is serious, thereby endangering the stability of the boiler combustion. The process of simulating coal interruption is as follows. When the coal mill is in stable operation, a negative slope signal is superimposed on PID3, such that the mass of coal entering the coal mill is gradually reduced to 0. The data generated in this process can be considered the coal interruption samples. A large number of coal interruption samples can be obtained by adjusting the set value to run the coal mill in other operation status and repeating the steps above to record fault data.

To verify the effectiveness of the simulation experiments of coal interruption, the variables that change significantly during the period of coal interruption are selected and their varying curves are drawn. Figure 3 shows the result of an arbitrary selection of experimental data. Figure 3(a) shows that coal interruption decreases the mass of coal entering the coal mill, which then decreases the outlet pulverized coal flow of the coal mill. Meanwhile, the heat consumption of the inlet primary air flow of coal mill through the coal mill reduces, thereby resulting in an upward trend in the outlet temperature of coal mill; the valve position of cold air is then rapidly opened, thereby making the outlet temperature of coal mill fall (Figure 3(b)). The reduction in the mass of coal stored in the coal mill results in the reduction of the current of coal mill and the differential pressure of primary air. The trend is consistent with that of the curves described in Figures 3(c) and 3(d). The ramp signal is removed after 75 s, and the variables are returned to the original set value under the control of the controller.

The research object in [19] is a MPS-type medium speed coal mill in a power plant in Hainan, China. The current study obtains sets of fault data of coal interruption by looking for the historical operation data of the coal mill and draws the varying curves of key variables as shown in Figure 4. According to the accident analysis, the coal interruption fault occurs because of the malfunctioning of the coal feeder; as a result, the actual supply of coal gradually reduces to 0 (Figure 4(a)). Figure 4 shows that, when coal interruption fault occurs, the outlet temperature of the coal mill rises (Figure 4(b)); however, the current of coal mill (Figure 4(c)) and the differential pressure of primary air (Figure 4(d)) decrease. This changing trend is similar to that of the key variables in the simulation of coal interruption. Therefore, the simulation of coal interruption in this study is reasonable. Accordingly, the data in the rectangular frame in Figure 3 can be recorded as fault samples.

3.2. Fault Simulation of Coal Choking. Coal choking may be caused by too little inlet primary air flow of coal mill, excessive coal feed, or too much moisture in raw coal. The process of simulating coal choking is as follows. A positive step signal is superimposed on PID3 to make the mass of coal in coal mill quickly reach the upper limit, and the data are recorded as the coal choking samples. Similar to previous simulation experiments, a large number of coal choking samples can be obtained by adjusting the set value to run the coal mill in other operation status and repeating the steps above to record fault data. The upper limit of the mass of coal stored in the coal mill is set as 60 kg.

The varying curves of the variables are drawn, and Figure 5 shows the result of an arbitrary selection of experimental data. As shown in the figure, the sudden increase in the set value of the outlet pulverized coal flow of coal mill causes the mass of raw coal in the coal mill to rise continuously (Figure 5(a)) and increases the resistance along the way, which results in the increase in differential pressure of primary air (Figure 5(c)). At the same time, the work load of coal mill increases accordingly, such that the current of the coal mill increases as well (Figure 5(b)). When the thickness of the raw coal reaches a certain degree, the grinding efficiency drops significantly, which then reduces the current of the coal mill. The outlet pulverized coal flow of coal mill is reduced to 0 until the mass of raw coal in the coal mill reaches the upper limit, at which time the primary air pipe is blocked and the pulverized coal cannot be blown out (Figure 5(d)). The analysis above shows that the simulation experiment results are consistent with the fault characteristics of coal choking; therefore, the data in the rectangular frame can be used as fault samples.

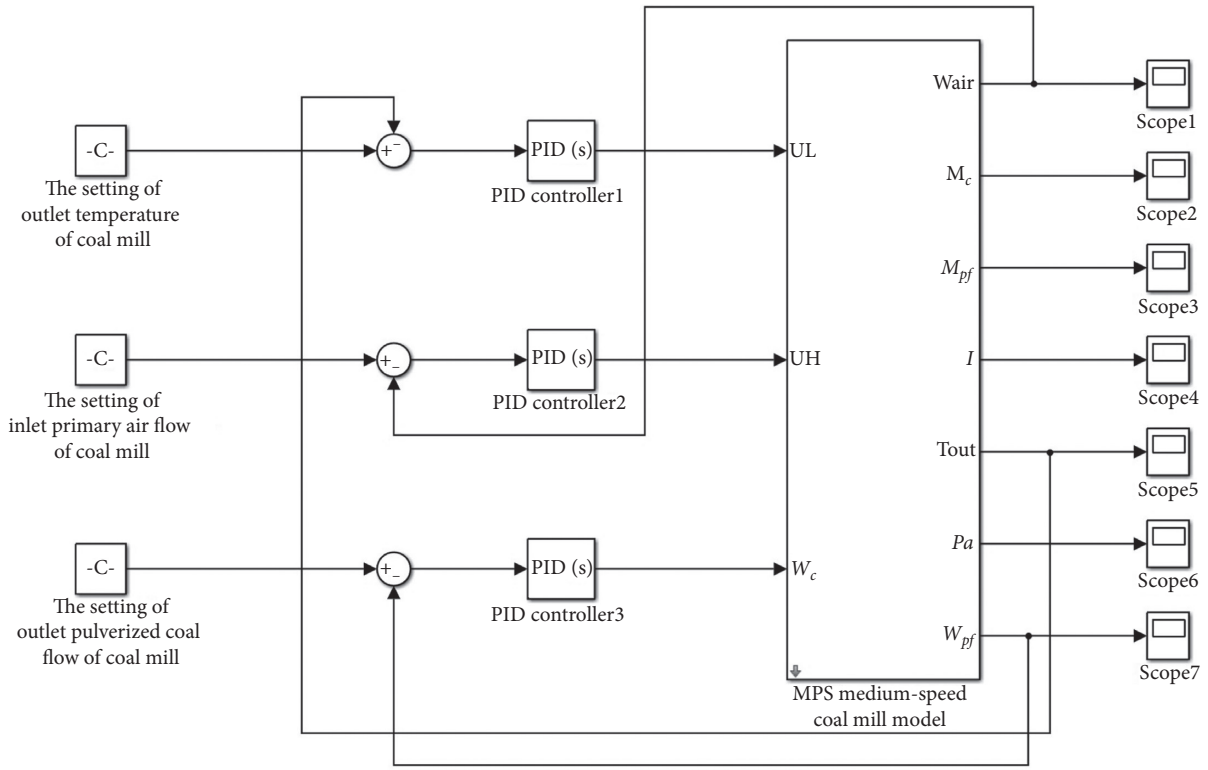


Figure 2: Control scheme for fault simulation of the coal mill.

Table 3: Variables declaration of control scheme.

Variable	Symbol
Controlled variable	T_{out}
	W_{air}
	W_{pf}
Control variable	u_L
	u_H
	W_c
State variable	M_c
	M_{pf}
	I
	ΔP_{pa}

4. Stacked Autoencoders

The fault diagnosis method based on historical operation data is a data-driven approach, which aims to obtain the nonlinear mapping relationship between the data and fault features. When sufficient data are available for learning, the deep neural networks can theoretically approximate any nonlinear function. This section describes a deep neural network called SAE, which is stacked by autoencoders (AEs), for fault diagnosis of coal mill and proposes two ways to improve the network performance.

4.1. Fundamentals of Autoencoder. An AE neural network can be considered a three-layer neural network. This network applies unsupervised learning algorithm to train and adjust the network weight and ultimately sets the network output to be equal to the network input. A typical example is shown in Figure 6, where $\{x_1, x_2, \ldots, x_n; x_i \in R^n\}$ can be treated as a set of unlabeled raw data and $\{x_1', x_2', \ldots, x_n'; x_i' \in R^n\}$ represents the network output. The circles with b are called bias units and correspond to the intercept term.

The transfer process of raw data from the input layer to the hidden layer is called encoding, and the transfer process from the hidden layer to the output layer is called decoding, which can be described by

$$d = S(W_1 x + b_1), \tag{2}$$

$$y = S(W_2 x + b_2), \tag{3}$$

where $S(\cdot)$ represents sigmoid function, W_1 represents the weight matrix between the input and hidden layers, W_2 represents the weight matrix between the hidden and output layers, and b_1 and b_2 represent the bias.

According to the concepts mentioned in this section, AE tries to learn a function $h_{w,b}(x) \approx x$. In other words, AE is trained to learn an approximate function such that the network output is similar to the network input. In fact, by putting constraints into AE, such as limiting the number of nodes in hidden layer, AE can obtain the low-dimensional feature of the data by compressing the high-dimensional input data. In Section 3, a large number of fault data have been obtained by fault simulation experiments. The remaining parts focus on the establishment of a suitable AE network to find the relationship between data and fault characteristics by effective learning of the complex fault data.

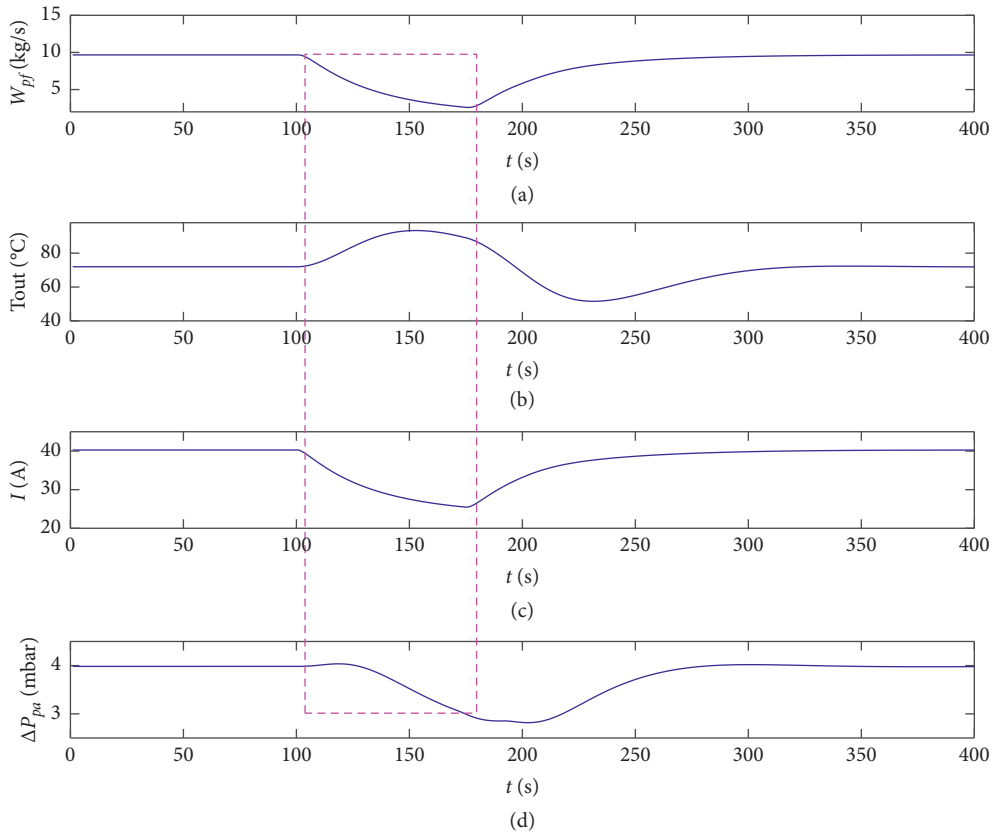


Figure 3: Fault simulation of coal interruption in coal mill. (a) Outlet pulverized coal flow of coal mill; (b) outlet temperature of coal mill; (c) current of coal mill; (d) differential pressure of primary air.

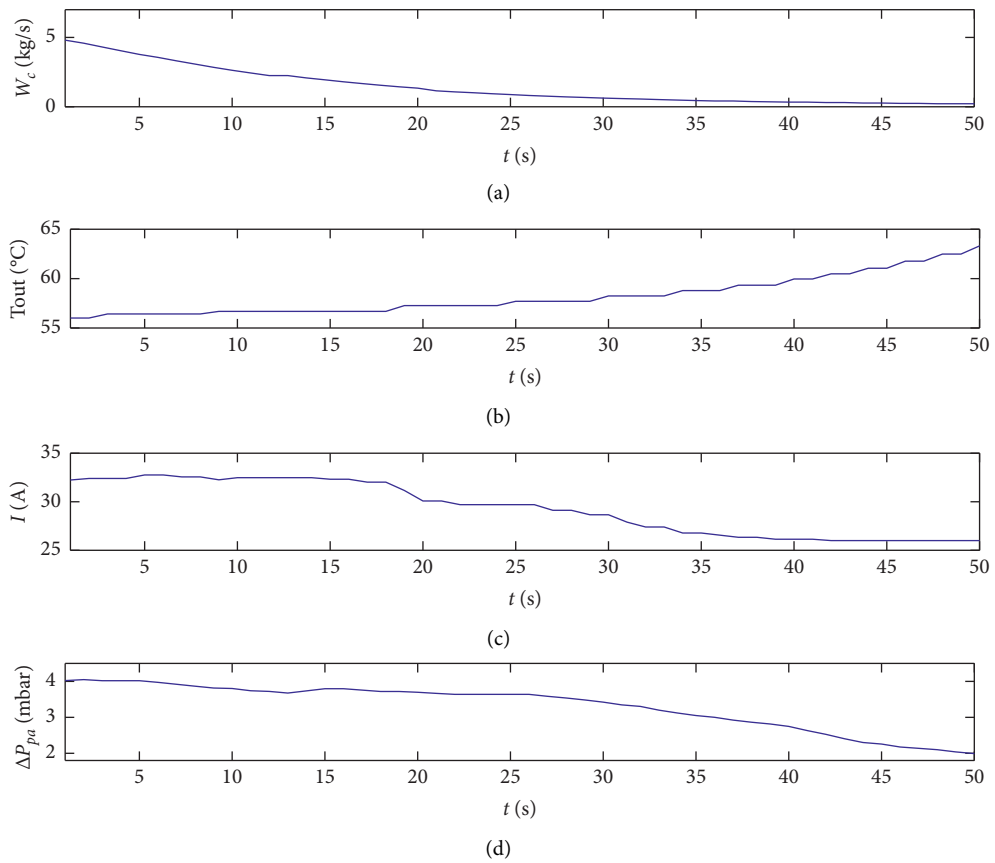


Figure 4: Actual fault data of coal interruption in coal mill. (a) Outlet pulverized coal flow of coal mill; (b) outlet temperature of coal mill; (c) current of coal mill; (d) differential pressure of primary air.

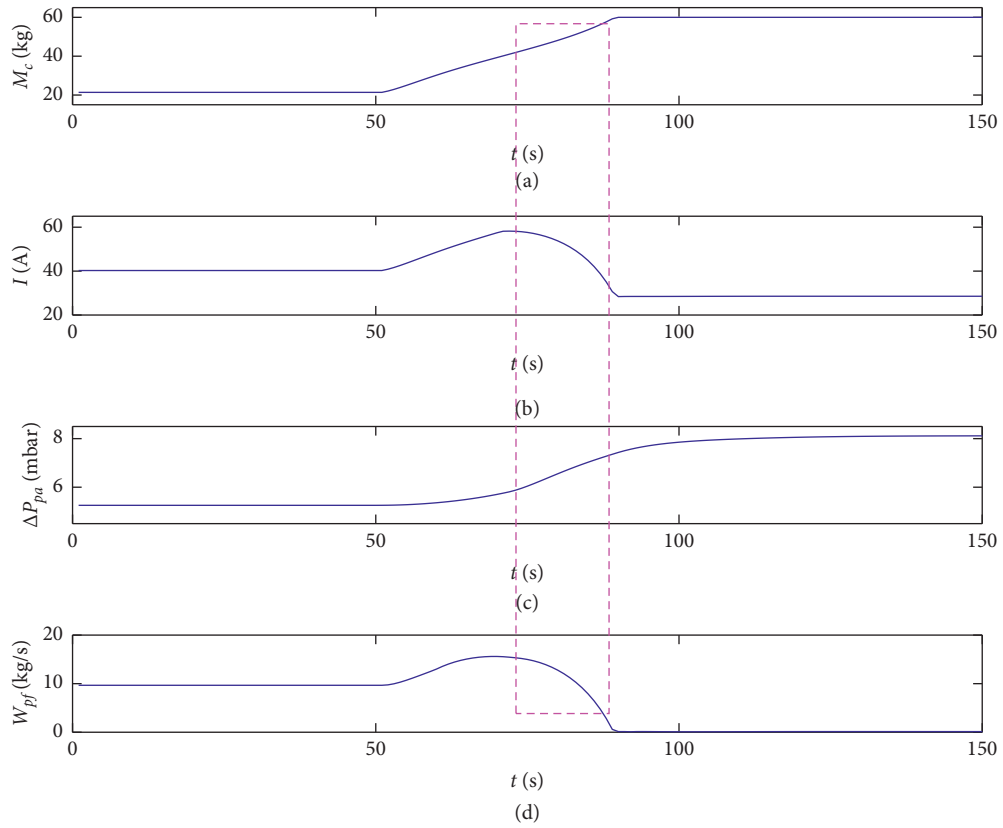


Figure 5: Fault simulation of coal choking in coal mill. (a) The mass of raw coal in coal mill; (b) current of coal mill; (c) differential pressure of primary air; (d) outlet pulverized coal flow of coal mill.

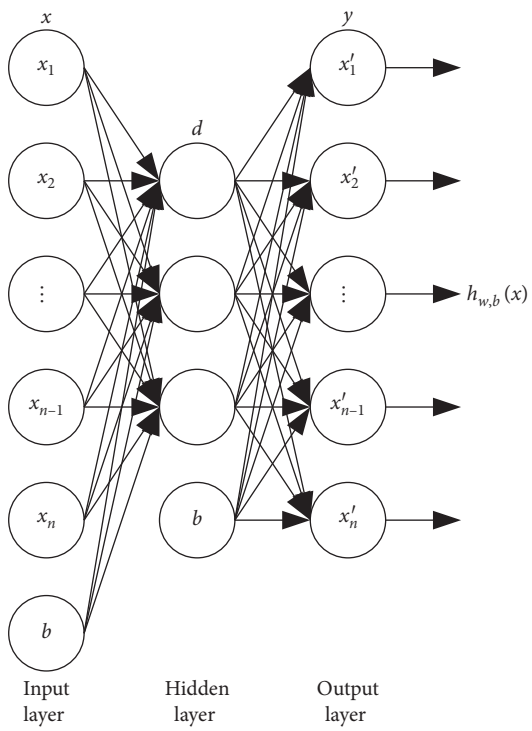


Figure 6: Network structure of AE.

The backpropagation algorithm is used for AE training. A training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ of m training samples is assumed. The network can be trained using batch gradient descent. For a single training example (x, y), the cost function can be defined as

$$J(W, b; x, y) = \frac{1}{2}\left\| h_{w,b}(x) - y \right\|^2. \tag{4}$$

For a training set of m samples, the overall cost function is

$$J(W, b) = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\, h_{w,b}(x) - y^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(W_{ji}^{(l)}\right)^2, \tag{5}$$

where λ represents weight decay coefficient that controls the relative importance of the two terms in equation (5). $W_{ji}^{(l)}$ represents the synaptic weight between the i-th neuron in layer l and j-th neuron in layer $l+1$. n_l represents the number of layers in AE. In other words, n_l can represent the output layer of the network, and s_l represents the number of the total neurons in layer l. The first term in the definition of $J(W, b)$ is an average sum-of-squares error term. The second term is a weight decay term that can decrease the magnitude of the weights and prevent overfitting.

The weight W and bias b are updated with gradient descent as follows:

$$W_{ij}^{'(l)} = W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b),$$
$$b_i^{'(l)} = b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W,b), \tag{6}$$

where α represents the learning rate. The partial derivatives in the equations above are derived as follows:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial W_{ij}^{(l)}} J\left(W,b,x^{(i)},y^{(i)}\right)\right] + \lambda W_{ij}^{(l)},$$
$$\frac{\partial}{\partial b_i^{(l)}} J(W,b) = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial b_i^{(l)}} J\left(W,b,x^{(i)},y^{(i)}\right), \tag{7}$$

where

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)} \delta_i^{(l+1)},$$
$$\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)}, \tag{8}$$

where $a_j^{(l)}$ represents the activation of unit j in layer l and $\delta_i^{(l+1)}$ represents the error term of layer $l+1$, given by

$$\delta_i^l = \left(\sum_{j=1}^{s_l+1} W_{ji}^{(l)} \delta_i^{(l+1)}\right) f'\left(z_i^{(l)}\right), \tag{9}$$

where $z_i^{(l)}$ represents the input weighted sum of unit i in layer l and $f'(\cdot)$ represents partial deflection of sigmoid function. The error term of the output layer n_l is given by

$$\delta_i^{(n_l)} = -\left(y_i - a_i^{(n_l)}\right) f'\left(z_i^{(n_l)}\right),$$
$$f'\left(z_i^{(n_l)}\right) = a_i^{(l)}\left(1 - a_i^{(l)}\right), \tag{10}$$

where $a_i^{(l)}$ represents the activation of unit i of layer l, $a_i^{(n_l)}$ represents the activation of unit i in the output layer, and $z_i^{(n_l)}$ represents the input weighted sum of unit i in the output layer.

Repeating the above equations can make the output of AE equal to the input of AE by minimizing the overall cost function (equation (5)).

4.2. Improvement of AE. As mentioned in Section 4.1, limiting the number of nodes in hidden layer is conducive to helping AE learn the relationship between the input data and fault features, because reducing the number of neurons can simplify the structure of the hidden layer and reduce the dimension of the input data.

Restricting the number of neurons can reduce the dimension of the data, but the network can learn few features in the hidden layer. On the basis of guaranteeing the diversity of the features in the hidden layer, a method called sparse constraint is introduced in this study to improve AE. The main idea is not to reduce the number of neurons but to consider restrictions to limit the activities of the neurons and thus reduce the dimension of the input data. Accordingly, the original overall cost function (equation (5)) should be modified to introduce an additional penalty factor, given by

$$J_{\text{sparse}}(W,b) = J(W,b) + \beta \sum_{j=1}^{s} \text{KL}\left(\rho \parallel \widehat{\rho}_j\right), \tag{11}$$

where

$$\text{KL}\left(\rho \parallel \widehat{\rho}_j\right) = \rho \log \frac{\rho}{\widehat{\rho}_j} + (1-\rho)\log \frac{1-\rho}{1-\widehat{\rho}_j}, \tag{12}$$

where $\sum_{j=1}^{s} \text{KL}(\rho \parallel \widehat{\rho}_j)$ represents the sparsity penalty term, β controls the weight of the sparsity penalty term, $\widehat{\rho}_j$ represents the average activation of unit j in hidden layer, ρ represents a sparsity parameter, and s represents the number of units in one hidden layer.

The penalty term has the following property: if $\widehat{\rho}_j = \rho$, then $\text{KL}(\rho \parallel \widehat{\rho}_j)$; the value increases monotonically with the difference between $\widehat{\rho}_j$ and ρ. Therefore, the activations of hidden units are sufficiently small when ρ is set close to zero.

Random noise is introduced into the input data to make the network learn rich information and thus prevent the AE from learning only the equivalent representation of the original data. The main idea is to set a small number of nodes in the input layer to zero at a small probability. However, the probability of introducing random noise should be appropriate; otherwise, the noise may cause irreversible damage to the input data.

SAEs are deep neural networks consisting of multiple layers of the improved AEs in which the output of each layer is wired to the input of the next layer. The SAE model is connected with a Softmax classifier to complete the construction of the deep neural network (Figure 7). The SAE model can identify the fault in the coal mill by learning the labeled data obtained from the fault simulation experiments of the coal mill.

5. Fault Diagnosis Based on SAE

5.1. Data Preprocessing and Health State Definition. In accordance with the fault simulation method described in Section 3, the simulation experiments are conducted repeatedly, and then 5000 sets of experimental data are obtained including three kinds of data samples; namely, coal mill operates in coal broken condition (coal interruption), full-of-coal condition (coal choking), and normal condition (normal operation). To facilitate the training of the SAE model, the three different operation conditions of the coal mill are labeled. The definition is shown in Table 4. The 5000 sets of data are randomly divided into training data and test data (Table 5). At the same time, two experiments are conducted to validate the effects of the proposed fault diagnosis method. The test samples of two experiments are the same.

5.2. Establishment of the SAE Model. The optimum SAE model has important effects on the accuracy rate of fault identification. In [25, 26], the unsupervised learning effect of SAE is reported to be affected by parameters of model, such

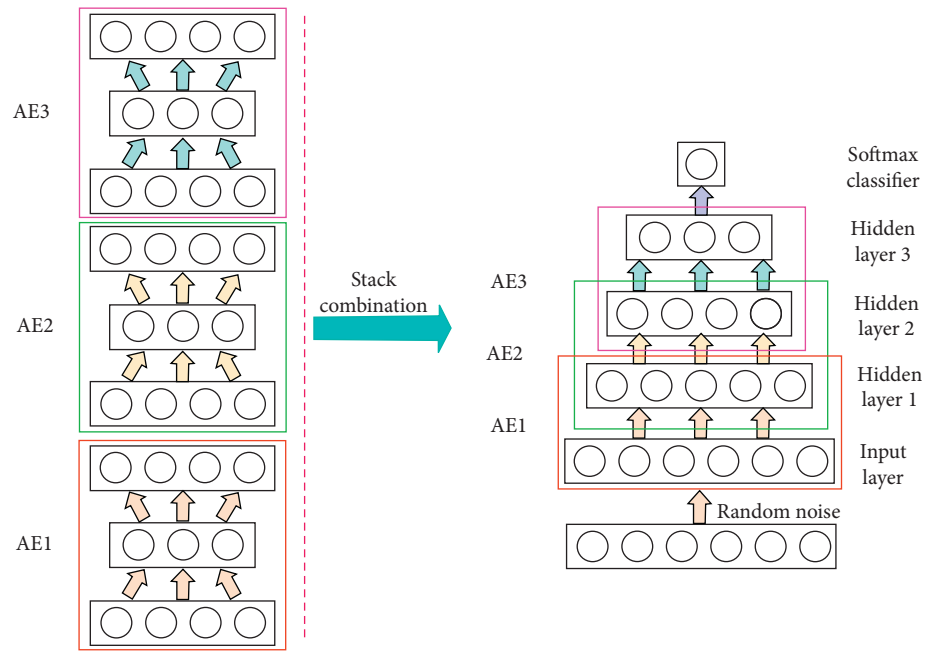


Figure 7: Deep neural network with SAE.

Table 4: Definition of different conditions of the coal mill.

Label	Condition
1	Coal interruption
2	Coal choking
3	Normal operation

Table 5: Training and testing data for fault diagnosis of the coal mill.

Experiment	Sample	Label		
		1	2	3
Experiment 1	Training samples	1500	1500	500
	Testing samples	61	58	75
Experiment 2	Training samples	1500	1500	1500
	Testing samples	61	58	75

as the number of nodes in the input and hidden layers, sparse parameter, and the number of times of network training. The experimental data of Experiment 1 are used as training samples, and relevant experiments are conducted to determine the optimum parameters of the SAE model. The evaluation index is the reconstruction error of the first layer of the SAE model, which is calculated by equation (4), and the experimental results are shown in Figure 8. According to the analysis in Section 3, the significantly changed variables during the fault period of the coal mill include differential pressure of primary air, outlet temperature of coal mill, and current of coal mill. Therefore, the three variables are used as the input nodes of the SAE model.

The SAE model can learn much information when the number of nodes in the input layer is large. However, the number of nodes in the input layer cannot be increased indefinitely because of the computational complexity. Figure 8(a) shows that, when the number of nodes in the input layer increases from 40 to 100, the reconstruction error of the network decreases continuously. If the number of nodes in the input layer increases further, then the reconstruction error will remain unchanged.

The number of nodes in the hidden layer determines the degree to which the model compresses the input data. The degree of compression is high when the number of nodes in the hidden layer is small. An experiment is employed using the first layer of SAE. In the experiment, the input size is set to 120 on the basis of the experiment above to determine the appropriate hidden layer parameters by analyzing the influence on the reconstruction results. As shown in Figure 8(b), when the number of hidden layer nodes is less than the input layer nodes, the reconstruction error fluctuates in a small range. This result indicates that the original data can obtain better compression when the number of hidden layer nodes is small, and this situation is conducive for the model to learn data characteristics. However, when the number of nodes in the hidden layer exceeds the number of nodes in the input layer, the reconstruction error increases rapidly, and the training effect is poor. The reason is that the sparse parameter ρ is set to 0 at this time, and the activities of neurons in the hidden layer cannot be limited, which then leads to the poor compression effect of the SAE model on original data. Combining the constraints of complexity of network structure and computational efficiency, the number of hidden layers is set to three, and the number of nodes in each layer is 100, 50, and 25.

Sparse constraint is introduced to improve the capability of the SAE model to compress input data. Figure 8(c) shows that, when the value of ρ is between 0.05 and 0.15, the reconstruction error of the network continues to decrease, showing that the inhibitory effect on neurons is appropriate. With the increase in the value of ρ, the inhibitory effect on neurons is excessive, and the reconstruction error increases rapidly.

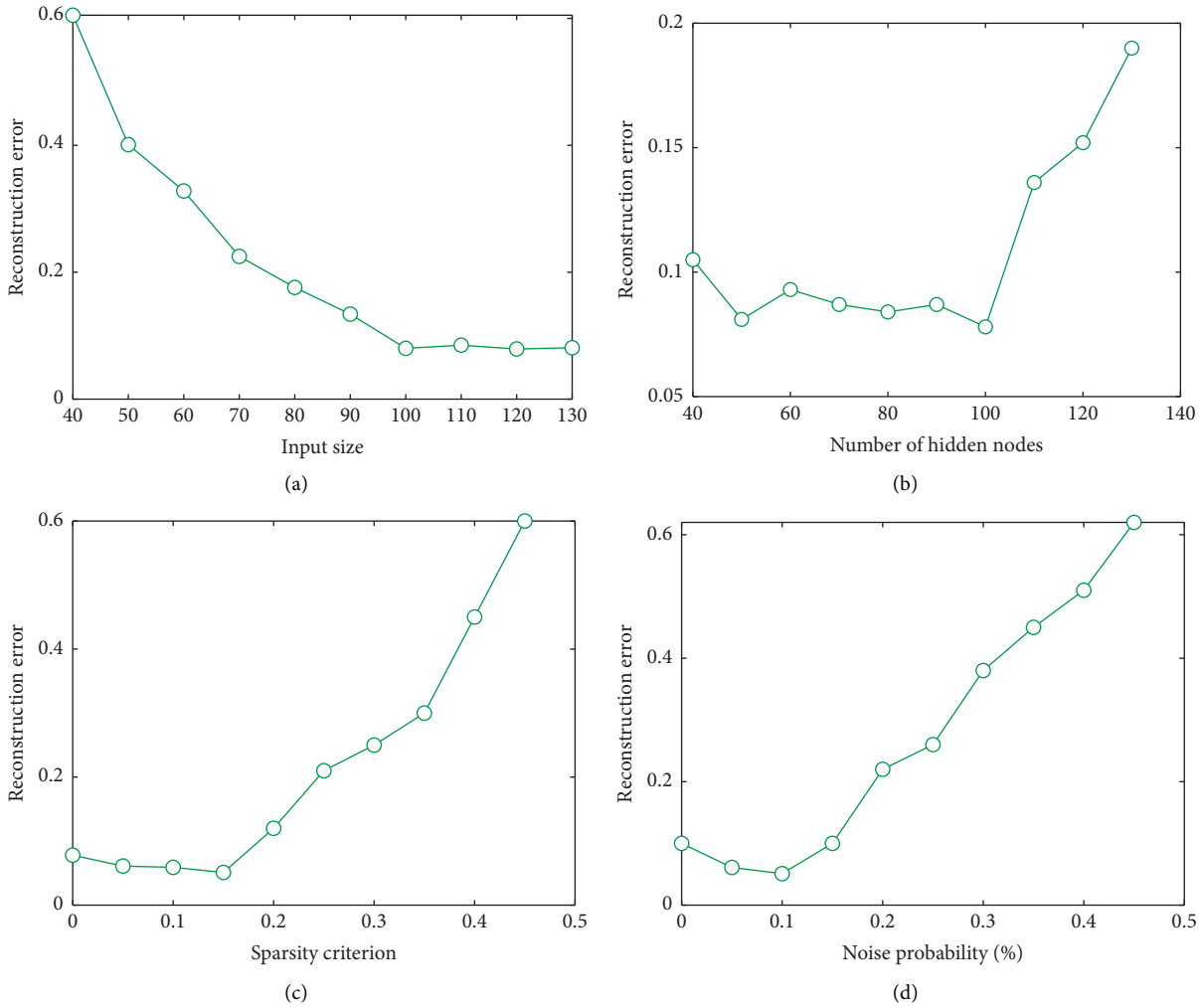


Figure 8: Reconstruction error curves for different SAE model parameters. (a) Input size; (b) number of hidden nodes; (c) sparse parameter; (d) noise probability.

Table 6: SAE network parameters.

Structure parameters	Input neurons	Hidden layer 1	Hidden layer 2	Hidden layer 3	Output layer	Transfer function
	120	100	50	20	4	Sigmoid
Learning parameters	Number of training	Batch size	ρ	Noise probability	β	Learning rate
	500	100	0.15	0.1	0.05	0.2

Random noise is introduced to the input data to prevent the SAE model from learning only the equivalent representation of the original data. As shown in Figure 8(d), when the probability of introducing noise is in the range of 0 to 0.1, reconstruction error decreases with the increase in noise. However, with the increase in noise, the reconstruction error increases rapidly because excessive noise causes nondestructive damage to the raw data.

In combination with the analysis above, the key parameters of the SAE model are shown in Table 6. The allocation of the nodes in the input layer is shown in Table 7. In the table, 1–40 nodes are sampled values of the differential pressure of primary air in four seconds (the sampling time is set to 0.1 s), 41–80 nodes are sampled values of outlet temperature of coal mill, and 81–120 nodes are sampled values of current of coal mill.

Table 7: Assignment situation of input layer nodes.

No.	Variable
1–40	Differential pressure of primary air
41–80	Outlet temperature of coal mill
81–120	Current of coal mill

5.3. Validation of the Proposed Method. After determining the parameters of SAE, the network is trained by the training data obtained from simulation experiments, and then the test data are sent into the network to test the result of fault identification of the SAE model, as shown in Figure 9. The simulation results of Experiment 1 show that all samples of false diagnosis in the 194 groups of test samples are from the normal operation data. Two sets of normal operation

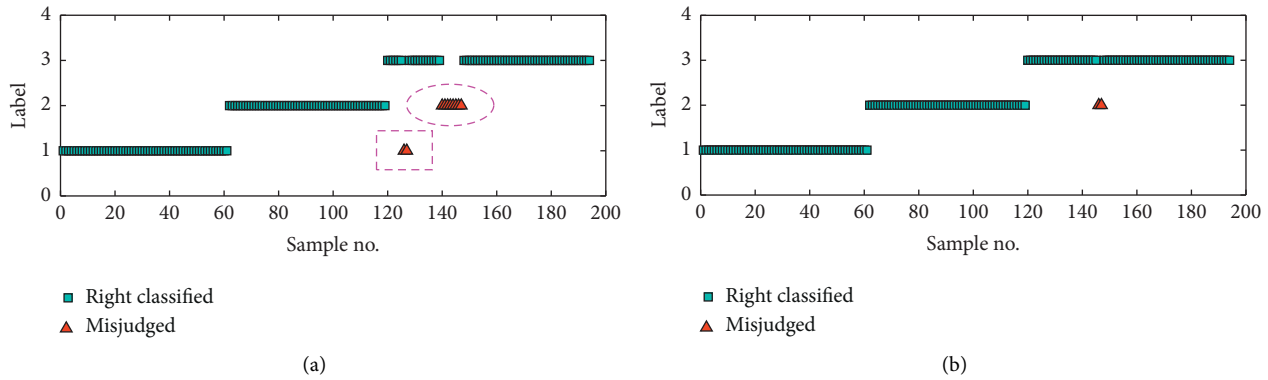


Figure 9: SAE-based classification result. (a) Experiment 1; (b) Experiment 2.

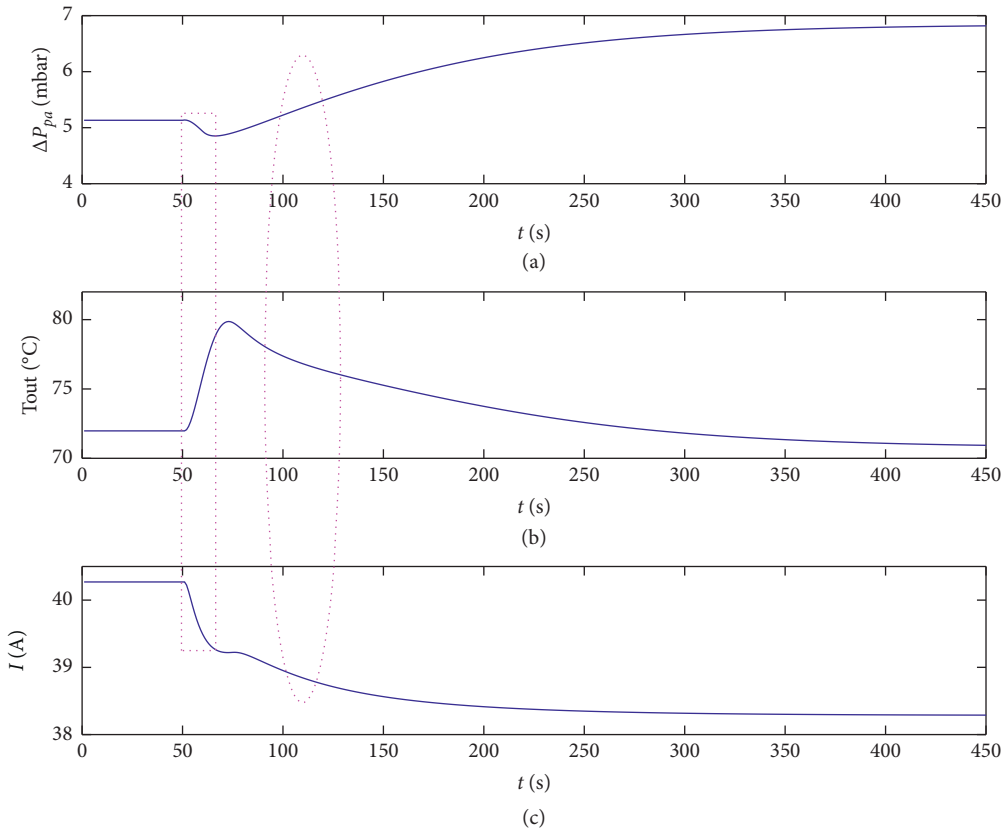


Figure 10: Data curve of test samples. (a) Differential pressure of primary air; (b) outlet temperature of coal mill; (c) current of coal mill.

samples are mistakenly diagnosed as coal interruption, and seven sets of normal samples are mistakenly diagnosed as coal choking. The accuracy rate of fault identification is 95.4%. To analyze the experimental results, the data curves corresponding to the misdiagnosed test samples are plotted and shown in Figure 10. The two sets of misdiagnosed samples contained in the rectangular box in Figure 9(a) correspond to data contained in the rectangular box in Figure 10. The change trend of data during this period indicates that differential pressure of primary air decreases, outlet temperature of coal mill rises, and current of coal mill decreases. These trends are consistent with the characteristics of coal mill when the coal interruption fault occurs, and these characteristics have been described in Section 3.1. Therefore, when normal operation samples are insufficient, SAE cannot fully study the differences between the two types

Table 8: Comparison of the effectiveness of network improvement.

Methods	Fault recognition accuracy (%)
SAE without noise and sparse constraint	84.02
SAE with sparse constraint	85.05
SAE with random noise	89.18
Improved SAE	95.4

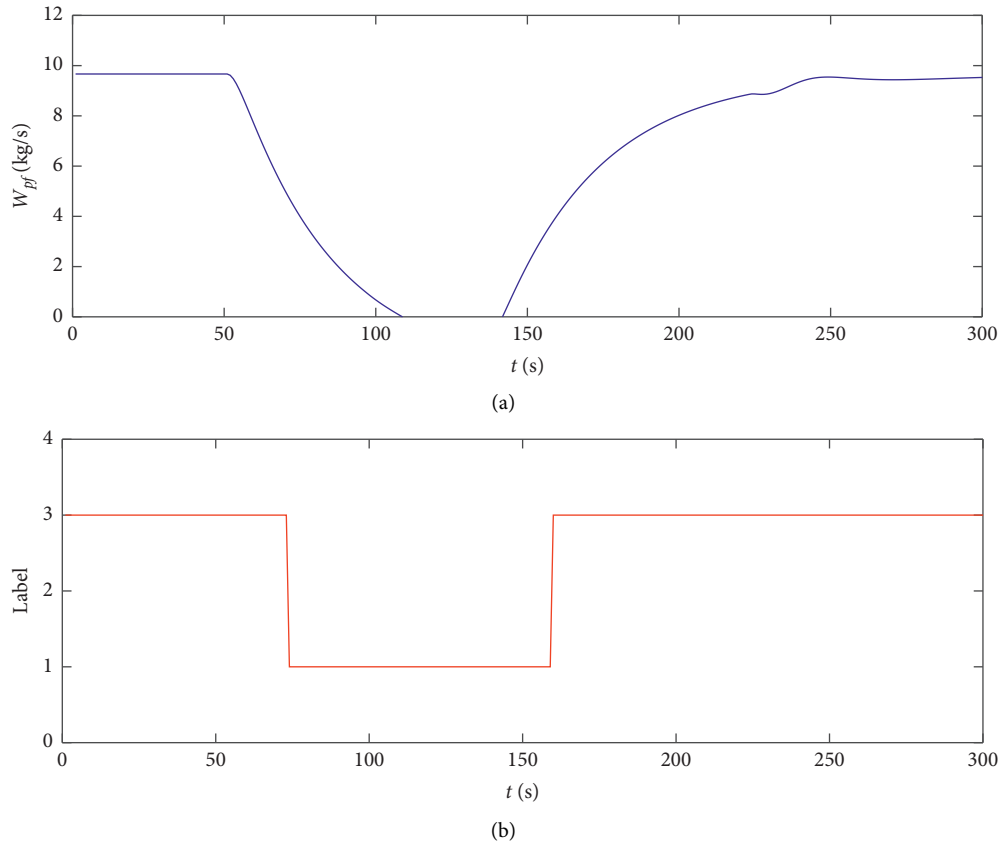


Figure 11: SAE-based fault diagnosis of coal interruption. (a) Outlet pulverized coal flow of coal; (b) early warning signal.

of data and wrongly diagnoses the normal operation data as coal interruption fault.

Similarly, from the data contained in the ellipse box in Figure 10, the change trend of the data in this period is found to be consistent with the characteristics of the coal choking, differential pressure of primary air rises, outlet temperature of coal mill decreases, and current of coal mill decreases. Thus, SAE can mistakenly diagnose the normal operation data as coal choking fault. To improve the accuracy of fault diagnosis of SAE, the training samples of normal operation are increased to 1500 groups, and an experiment (Experiment 2) is conducted again. Figure 9(b) shows that, although two sets of misdiagnosed samples are still present, the accuracy of fault diagnosis of SAE has been improved to 98.7%. Therefore, if the training samples continue to increase, then the accuracy of fault diagnosis of SAE will theoretically be close to 100%.

To illustrate the effectiveness of the method that improves the performance of SAE, the accuracy rate of fault diagnosis of SAE before and after the algorithm improvement is compared. The comparison results are shown in Table 8. Training and test samples are from the data in Experiment 1. Table 8 shows that, when no improvement is implemented in SAE, the fault recognition rate is 84.02%. When sparse constraint is introduced in SAE, the fault recognition rate is increased to 85.05%. After introducing random noise into input data, the fault identification rate increases to 89.18%. When two improved methods are introduced into SAE, the network fault recognition rate further increases to 95.4%. The above analysis shows that the two improved methods proposed in this study can improve the fault recognition capability of SAE.

The coal mill is characterized by a large delay system. Detecting changes in outlet pulverized coal flow of coal mill to find the operation fault in coal mill often cannot establish early warning. Through real-time monitoring of differential pressure of primary air, outlet temperature of coal mill, and current of coal mill, and the three kinds of fast changing signals, the trained SAE can find the operation fault in the coal mill in advance. Coal interruption is taken as an example in this study. As shown in Figure 11, coal interruption fault occurs by artificial simulation. As a result, the outlet pulverized coal flow of coal mill reduces to 0 in 110 s, while the output of the SAE jumps from the normal operation state to coal interruption fault in 75 s. The network has advanced 35 s to predict the fault in the coal mill. With the adjustment of PID, the outlet pulverized coal flow of the coal mill rises gradually and goes back to the safety limit in 160 s. At this point, the output of the SAE returns to normal. Thus, the proposed method based on the deep learning algorithm can play an important role in the fault diagnosis of the coal mill.

6. Conclusions

In this study, a deep learning algorithm based on a data-driven model is proposed for fault diagnosis of coal mills. On the basis of the mechanism model of coal mills, the fault

operation of coal mills is simulated and numerous fault data are obtained. Thus, the difficulty in obtaining the fault data using traditional methods is addressed. The performance of SAE is improved by introducing sparse constraints and random noise in the input layer. At the same time, the accuracy of fault diagnosis of coal mills is effectively improved, thereby enabling the possible prediction of fault in coal mills. The method proposed in the paper greatly improves the accuracy of the fault diagnosis of coal mills, which is of great significance for ensuring the safe operation of power plants. In addition, the proposed method is easy to generalize. Complex mechanical equipment in other industrial fields can use this method for fault diagnosis. The method can reduce the use of sensors for fault diagnosis of large equipment and the investment of human resources, which is essential to improve the economy and safety of the industry.

It should be noted that the paper does not consider online training for SAE. The main reason is that the online training of deep neural networks will take a lot of time, which puts stricter requirements on the performance of computers and optimization algorithms. So, we adopt a simplified method which chooses to directly use the offline training model to achieve fault diagnosis. It greatly saves the cost of calculation. However, using the online training model can continuously optimize the accuracy of the model [27–31] as the data accumulate. Therefore, how to realize the online training of the model will be the focus of our follow-up research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] V. Agrawal, B. K. Panigrahi, and P. M. V. Subbarao, "Review of control and fault diagnosis methods applied to coal mills," *Journal of Process Control*, vol. 32, pp. 138–153, 2015.

[2] P. F. Odgaard and B. Mataji, "Observer-based fault detection and moisture estimating in coal mills," *Control Engineering Practice*, vol. 16, no. 8, pp. 909–921, 2008.

[3] P. Andersen, J. D. Bendtsen, T. S. Pedersen, and B. Mataji, "Coal moisture estimation in power plant mills," in *Proceedings of the 2009 17th Mediterranean Conference on Control and Automation*, pp. 1066–1071, Thessalonik, Greece, 2009.

[4] J.-L. Wei, J. Wang, and Q. H. Wu, "Development of a multisegment coal mill model using an evolutionary computation technique," *IEEE Transactions on Energy Conversion*, vol. 22, no. 3, pp. 718–727, 2007.

[5] S. Guo, J. Wang, J. Wei, and P. Zachariades, "A new model-based approach for power plant tube-ball mill condition monitoring and fault detection," *Energy Conversion and Management*, vol. 80, pp. 10–19, 2014.

[6] Y. Wang, H. R. Karimi, H. Shen, Z. Fang, and M. Liu, "Fuzzy-model-based sliding mode control of nonlinear descriptor systems," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3409–3419, 2019.

[7] Y. Wang, X. Yang, and H. Yan, "Reliable fuzzy tracking control of near-space hypersonic vehicle using aperiodic measurement information," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9439–9447, 2019.

[8] E. A. Garcia, Z. Han, and P. M. Frank, "FDI based on parameter and output estimation: an integrated approach," in *Proceedings of the European Control Conference*, Karlsruhe, Germany, 1999.

[9] T. Hofling and R. Isermann, "Adaptive parity equations and advanced parameter estimation for fault detection and diagnosis," in *Proceedings of the IFAC World Congress*, pp. 55–60, San Francisco, CA, USA, July 1996.

[10] P. Nomikos and J. F. MacGregor, "Monitoring batch processes using multiway principal component analysis," *AICHE Journal*, vol. 40, no. 8, pp. 1361–1375, 1994.

[11] Z.-G. Su, P.-H. Wang, X.-J. Yu, and Z.-Z. Lv, "Experimental investigation of vibration signal of an industrial tubular ball mill: monitoring and diagnosing," *Minerals Engineering*, vol. 21, no. 10, pp. 699–710, 2008.

[12] E. Kisić, V. Petrović, S. Vujnović, Ž. Đurović, and M. Ivezić, "Analysis of the condition of coal grinding mills in thermal power plants based on the T^2 multivariate control chart applied on acoustic measurements," *Facta Universitatis-Series: Automatic Control and Robotics*, vol. 11, no. 2, pp. 141–151, 2012.

[13] S. Collura, D. Possanzini, D. Pestonesi, and M. Gualerci, *Coal Mill Performances Optimization through Non-invasive Online Coal Fineness Monitoring*, Powergen, Vienna, Austria, 2013.

[14] X. Han and X. Jiang, "Fault diagnosis of pulverizing system based on fuzzy decision-making fusion method," in *Fuzzy Information and Engineering*, vol. 2, pp. 1045–1056, Springer Berlin Heidelberg, Berlin, Germany, 2009.

[15] W. Qin, W. Yan, and J. Xu, "Application of fault diagnosis expert system in grinding process," in *Proceedings of the Automation and Logistics (ICAL)*, pp. 290–295, Hong Kong, China, 2010.

[16] Y. Wang, H. R. Karimi, H.-K. Lam, and H. Shen, "An improved result on exponential stabilization of sampled-data fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 6, pp. 3875–3883, 2018.

[17] R. Isermann and P. Balle, "Trends in the application of model based fault detection and diagnosis of technical processes," in *Proceedings of the IFAC World Congress*, pp. 55–60, San Francisco, CA, USA, July 1996.

[18] P. M. Frank, "Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy," *Automatica*, vol. 26, no. 3, pp. 459–474, 1990.

[19] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, pp. 490–502, 2016.

[20] Y. Duan, Y. Lv, Y. L. Liu et al., "An efficient realization of deep learning for traffic data imputation," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 168–181, 2016.

[21] Y. Gao, D. Zeng, J. Liu, and Y. Jian, "Optimization control of a pulverizing system on the basis of the estimation of the outlet coal powder flow of a coal mill," *Control Engineering Practice*, vol. 63, pp. 69–80, 2017.

[22] D.-L. Zeng, Y. Hu, S. Gao, and J.-Z. Liu, "Modelling and control of pulverizing system considering coal moisture," *Energy*, vol. 80, pp. 55–63, 2015.

[23] D. Simon, "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms," *IET Control Theory & Applications*, vol. 4, no. 8, pp. 1303–1318, 2010.

[24] V. Rashtchi, E. Rahimpour, and S. Fazli, "Genetic algorithm application to detect broken rotor bar in three phase squirrel cage induction motors," *International Review of Electrical Engineering*, vol. 6, no. 5, pp. 2286–2292, 2011.

[25] H. Pan and X. W. Wang, "Survey on collaborative filtering recommendation algorithm based on extreme learning machine stacked denoising autoencodes," *Application Research of Computers*, vol. 33, no. 8, 2016.

[26] Y. Jian, X. Qing, H. Liang, Y. Zhao, X. Qi, and M. Du, "Fault diagnosis of motor bearing based on deep learning," *Advances in Mechanical Engineering*, vol. 11, no. 9, pp. 1–9, 2019.

[27] A. Das and F. L. Lewis, "Distributed adaptive control for synchronization of unknown nonlinear networked systems," *Automatica*, vol. 46, no. 12, pp. 2014–2021, 2010.

[28] S. El-Ferik, H. A. Hashim, and F. L. Lewis, "Neuroadaptive distributed control with prescribed performance for the synchronization of unknown nonlinear networked systems," *IEEE Transactions on Systems,Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2135–2144, 2017.

[29] S. El-Ferik, A. Qureshi, and F. L. Lewis, "Neuro-adaptive cooperative tracking control of unknown higher-order affine nonlinear systems," *Automatica*, vol. 50, no. 3, pp. 798–808, 2014.

[30] H. A. Hashim, S. El-Ferik, and F. L. Lewis, "Adaptive synchronisation of unknown nonlinear networked systems with prescribed performance," *International Journal of Systems Science*, vol. 48, no. 4, pp. 885–898, 2017.

[31] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches*, Springer Science & Business Media, Berlin, Germany, 2013.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2091382, 10 pages
https://doi.org/10.1155/2020/2091382

Research Article

Fault Detection of Wind Turbine Pitch System Based on Multiclass Optimal Margin Distribution Machine

Mingzhu Tang,[1,2] **Zijie Kuang**,[1] **Qi Zhao**,[1] **Huawei Wu**,[2] **and Xu Yang**[3]

[1]*School of Energy and Power Engineering, Changsha University of Science & Technology, Changsha 410114, China*
[2]*Hubei Key Laboratory of Power System Design and Test for Electrical Vehicle, Hubei University of Arts and Science, Xiangyang 441053, China*
[3]*State Grid (Beijing) Integrated Energy Service Company Limited, Beijing 100176, China*

Correspondence should be addressed to Huawei Wu; whw_xy@hbuas.edu.cn

Received 11 June 2020; Accepted 19 June 2020; Published 4 August 2020

Guest Editor: Yong Chen

In response to the unbalanced sample categories and complex sample distribution of the operating data of the pitch system of the wind turbine generator system, this paper proposes a method for fault detection of the pitch system of the wind turbine generator system based on the multiclass optimal margin distribution machine. In this method, the power output of the wind turbine generator system is used as the main status parameter, and the operating data history of the wind turbine generator system in the wind power supervisory control and data acquisition (SCADA) system is subject to correlation analysis with the Pearson correlation coefficient, to eliminate the features that have low correlation with the power output status parameter. Secondary analysis is performed to the remaining features, thus reducing the number and complexity of samples. Datasets are divided into the training set for training of the multiclass optimal margin distribution machine fault detection model and test set for testing. Experimental verification was carried out with the operating data of one wind farm in China. Experimental results show that, compared with other support vector machines, the proposed method has higher fault detection accuracy and precision and lower false-negative rate and false-positive rate.

1. Introduction

Wind turbine generator systems are usually used in complex and unstable natural environments and eroded by sunlight, rain, wind, and sand all the year round. In addition, the wind turbine generator systems work at high altitudes, and their main parts are in high-altitude nacelles, which may lead to faults during operation. Long downtime of the wind turbine generator system arising from failure will result in a lot of operation and maintenance costs and part replacement costs, low power generation efficiency of the wind farm, and huge economic losses [1].

The pitch system is a critical part of the wind turbine generator system, mainly consisting of the blades, hubs, and other parts. These parts account for a large proportion in terms of the average maintenance time, material costs, and corresponding technical personnel [2]. Hence, it is of particular importance to guarantee the safe and stable operation of the pitch system of the wind turbine generator system. Timely and efficient status monitoring and fault detection of the pitch system has excellent economic benefits and engineering application values for the wind power industry [3].

The current fault detection of wind turbine generator systems is mainly based on the data analysis of the wind power supervisory control and data acquisition (SCADA) system. A correlation model is built by analyzing the data (e.g., power, vibration, and temperature) generated in the operation of the wind turbine generator system, to obtain the operating status, fault, and other information of the wind turbine generator system and thus detecting faults [4].

Fault detection mainly involves two aspects: feature selection and detection model [5–7]. The status parameters reflecting the faults of the wind turbine generator system are

selected from the SCADA system, and the detection model is built through training, which is used for status monitoring and fault detection of the wind turbine generator system. Pandit and Infield proposed a method of status monitoring of the wind turbine generator system based on the Gaussian process [8]. The wind power curve is predicted according to the data of the SCADA system and used for yaw fault detection of the wind turbine generator system. Liu et al. presented a method of wavelet transform fault detection based on the generative adversarial network, in which the normal operating data of the wind turbine generator system is converted into rough fault data based on the prior knowledge, and a generative adversarial network model is built for fault detection [9]. Ruiming et al. proposed a method based on SCADA data and dynamical network marker, which is constructed as a fault warning signal of a wind turbine [10]. The techniques including the multinode complex network and the correlation and cross-correlation analysis of the denosed method, and the experiment verifies its convenience and robustness. However, the excessive use of the feature parameters based on artificial experience will introduce human influencing factors into fault detection, resulting in interference in the detection process. Due to the particularity of the SCADA system, the operating data of the wind turbine generator system may be missing or abnormal, and it is difficult to extract effective features from a large amount of raw data, which may lead to low efficiency [11]. Furthermore, the current SCADA system is not yet mature, which may involve strong coupling of status parameters. The use of these parameters will lead to redundancy and ultimately model overfitting. Hence, more potential of SCADA data needs to be explored [12].

The support vector machine (SVM), as a machine learning method based on the statistical theory, has good learning performance. It has been successfully applied in many fields such as multiclass recognition and regression forecasting [13–15] and favored by a large number of scholars in the field of fault research on wind turbine generator systems, including fault diagnosis and prediction of wind turbine generator systems via the SVM [16–18]. Liu et al. presented the diagonal spectrum and clustering binary tree are combined with the SVM for fault detection of the gearboxes of wind turbine generator systems [19]. Hang et al. proposed a method of fault diagnosis of wind turbine generator systems based on the multilevel fuzzy SVM classifier, in which the fault feature vectors are extracted from vibration signals utilizing empirical mode decomposition, and the kernel function parameters of the fuzzy clustering algorithm are optimized, and the faults of wind turbine generator systems are diagnosed via the multilevel fuzzy SVM [20]. Saari et al. were to detect and identify wind turbine bearing faults by using fault-specific features extracted from vibration signals. Automatic identification was achieved by training models by using these features as an input for a one-class support vector machine [21]. In the SVM, however, classification is based on the identification of the hyperplane with the minimum margin, leading to low generalization performance and, in the case of complex nonlinear multiclassification, final optimization may become a nondifferentiable nonconvex process [22]. In order to solve this problem, Zhang and Zhou put forward the multiclass optimal distribution machine (mcODM), in which a distribution model is built based on the sample distribution features during fault detection. The sample mean and sample variance are taken into account for higher classification performance [23]. The experiments of multiple datasets have verified the accuracy and generalization performance of this model and the model complexity is relatively low during the optimization.

To resolve the unbalanced samples and complex distribution in fault detection of the pitch system of the wind turbine generator system, a method for fault detection of the pitch system of the wind turbine generator system based on mcODM is proposed. This method mainly consists of three parts. First, the SCADA data of the wind turbine generator system are preprocessed, including data cleaning and normalization. Secondly, the correlation of parameters is analyzed according to the operation mechanism of the wind turbine generator system and the Pearson correlation coefficient, followed by feature selection. Finally, sample sets are built, including the training set for training of the detection model and the test set for testing of this model, using the actual operating data of one wind farm in China as experimental data. Experimental results show that this method has higher accuracy and precision of fault detection and lower false-negative rate and false-positive rate.

2. Pitch System of Wind Turbine Generator System

The pitch system of the wind turbine generator system is used to change the upwind area of the blades when the rotor is facing the wind, thus controlling the rotation torque of the rotor. In combination with the yaw system, the wind turbine generator system can maintain the stable efficiency of power generation under different wind conditions [24]. At present, the pitch systems of wind turbine generator systems are mainly divided into the hydraulic pitch system and electric pitch system.

The hydraulic pitch system is equipped with a set of crank sliding structure to drive all blades for synchronous pitching. This system has a fast response to pitch signals and large pitch torque, which is conducive to the centralized layout and integration. It is mostly used in large-sized wind turbine generator systems. However, it is a nonlinear system that has a relatively complex structure and may be subject to hydraulic oil leakage, jamming, etc. [25].

The electric pitch system is equipped with an independent control mechanism for each blade and composed of the pitch controller, servo driver, and standby power supply, in which the pitch of each blade is controlled separately. Its transmission features a relatively simple structure, stable operation, and high reliability, but has large inertia due to its poor dynamic features. Where the wind speed changes rapidly, frequent pitching may lead to controller overheat and damage to the body [26].

Once the pitch system of the wind turbine generator system fails, the blade pitch will be abnormal and the

rotation torque of the rotor will not be the expected value. If the speed is too low, the wind energy capture rate will be affected. The mechanical energy generated in the rotation will be transferred to the generator through the gearbox transmission chain, resulting in the abnormal speed of the generator and ultimately affecting the power output of the generator. Accordingly, the safe and stable operation of the pitch system is essential for the stable and efficient power generation of the wind turbine generator system.

During fault detection of the pitch system of the wind turbine generator system, an important step is to acquire the status parameters that effectively reflect the features of the pitch system from a lot of SCADA data. Due to the particularity of the SCADA system that involves the complex and diverse parameters of the pitch system, including strong coupling parameters, it is necessary for feature selection to optimize the model complexity to reduce the calculation time and the amount and select the effective status parameters and also to take redundant items into account to delete excess parameters and avoid model overfitting [27, 28].

The method proposed in this paper is for fault detection of the electric pitch systems of large-sized wind turbine generator systems. The experimental data are the actual operating data of a wind farm, and various categories of samples are used. The method involves the typical data category imbalance, complex distribution, and the like.

3. Fault Detection of Pitch System of Wind Turbine Generator System

Fault detection of the pitch system of the wind turbine generator system consists of the preprocessing of the operating data acquired, selection of effective features, and building of the sample sets, including training sets for the training of the detection model and the test set for testing. Figure 1 shows the mcODM-based process for fault detection of the pitch system of the wind turbine generator system.

3.1. Data Cleaning and Preprocessing. In order to obtain the fault samples of the pitch system of the wind turbine generator system, the actual operating data of the wind turbine generator system of one wind farm are used, including the sensor monitoring data during normal operation and at the failure time of the pitch system. Unstable environmental factors and sensor abnormalities under the actual operating conditions will cause information processing errors, data losses, data abnormalities, and other problems. Thus, the obtained raw data are cleaned and preprocessed as follows:

Step 1: delete the "no data" variable in the dataset

Step 2: delete all status variables with a value of "0"

Step 3: according to the fault record of the wind turbine generator system, select the data from 30 min before a fault to 30 min after the fault

Step 4: normalize the sample data by the following formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{1}$$

where X is a status parameter, $X_{\min}$ and $X_{\max}$ represent the minimum and maximum value of the status variable, respectively, and X' represents the normalized value.

Normalization makes the model smoother and more convergent to find the optimal solution.

3.2. Feature Selection. According to the mechanism analysis of the pitch system, when the pitch system fails, the main status parameter that is ultimately affected is the power output of the wind turbine generator system. Hence, the correlation between the power output and other operating parameters of the wind turbine generator system is analyzed based on the Pearson correlation coefficient during feature selection, to delete the parameters that are little correlated to the pitch system.

The Pearson correlation coefficient was proposed by the British statistician Karl Pearson in the 20th century. It reflects the degree of correlation between two variables and calculated by the following formula:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left((X - \mu_X)(Y - \mu_Y)\right)}{\sigma_X \sigma_Y}, \tag{2}$$

where $\operatorname{cov}(X,Y)$ represents the covariance of the two variables, and μ_X/μ_Y and σ_X/σ_Y represent the mean and standard deviation of the two variables, respectively.

The aforesaid formula defines the population correlation coefficient. When the sample size of the variables X and Y is n, the Pearson correlation coefficient is given by

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}, \tag{3}$$

where r represents the degree of linear correlation between the two variables. It ranges from -1 to 1, i.e., $-1 \le r \le 1$, as described as follows: $0 < r < 1$: the two variables are positively correlated. The closer r is to 1, the greater the positive correlation of the variables; $-1 < r < 0$: the two variables are negatively correlated. The closer r is to -1, the greater the negative correlation of the variables; $|r| = 1$: the two variables are linearly correlated; and $r = 0$: the two variables are linearly independent of each other.

In order to further reduce the sample size as well as the computational complexity of the model, and avoid model overfitting, the status variables selected in the first step are subject to a secondary Pearson correlation analysis, to delete some highly correlated parameters and resolve the redundancy.

Following the feature selection of the datasets based on the Pearson correlation coefficient, the normal and fault samples are divided into the training set for model training and test set for model performance testing.

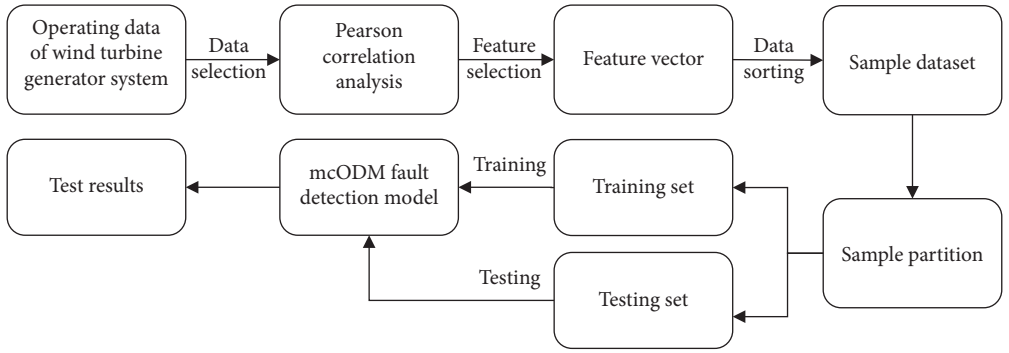


Figure 1: mcODM-based process for fault detection of the pitch system of wind turbine generator system.

3.3. mcODM Algorithm. A feature set $X = [x_1, \ldots, x_k]$ is assumed, corresponding to the category label set $Y = [K]$, where $[K] = \{1, \ldots, k\}$. The training set is $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. The mapping function φ is defined, and the sample set is mapped by the kernel function κ to the high-dimensional space φ: $X \longrightarrow \mathrm{H}$. The corresponding weight vectors are $\omega_1, \ldots, \omega_k$. A scoring function is $\omega_l^{\mathrm{T}}\varphi(x)$ defined for each weight vector ω_l. The feature value of each sample and the corresponding label will maximize the value of the scoring function of the samples, i.e., $h(x) = \mathrm{argmax}_{l\in Y}\omega_l^{\mathrm{T}}\varphi(x)$, thereby leading to a margin definition:

$$\gamma_h(x, y) = \omega_y^{\mathrm{T}}\varphi(x) - \max_{l\neq y}\omega_l^{\mathrm{T}}\varphi(x). \tag{4}$$

When a negative margin is generated in calculation, the category provided by the classifier will be incorrect.

Let $\overline{\gamma}$ represent the mean of margins, the optimal margin distribution machine can be expressed as follows:

$$\begin{aligned} \min_{\omega,\overline{\gamma},\xi_j,\varepsilon_j} \quad & \Omega(\omega) - \eta\overline{\gamma} + \frac{\lambda}{m}\sum_{j=1}^{m}\left(\xi_j^2 + \varepsilon_j^2\right), \\ \text{s.t.} \quad & \gamma_h\left(x_j, y_j\right) \geq \overline{\gamma} - \xi_j, \\ & \gamma_h\left(x_j, y_j\right) \leq \overline{\gamma} + \varepsilon_j, \forall j, \end{aligned} \tag{5}$$

where $\Omega(\omega)$ is a regular term, η and λ are balance parameters, ξ_j and ε_j are the positive and negative deviations of the margin $\gamma_h(x_j, y_j)$ and its mean $\overline{\gamma}$, respectively, and $(1/m)\sum_{j=1}^{m}(\xi_j^2 + \varepsilon_j^2)$ is the variance.

The margin mean can be fixed at 1 by scaling of ω. The deviation of the sample (x_j, y_j) and margin mean will be $|\gamma_h(x_j, y_j) - 1|$. Then, the optimal margin distribution machine can be expressed as follows:

$$\begin{aligned} \min_{\omega,\xi_j,\varepsilon_j} \quad & \Omega(\omega) + \frac{\lambda}{m}\sum_{j=1}^{m}\frac{\xi_j^2 + \tau\varepsilon_j^2}{(1-\theta)^2}, \\ \text{s.t.} \quad & \gamma_h\left(x_j, y_j\right) \geq 1 - \theta - \xi_j, \\ & \gamma_h\left(x_j, y_j\right) \leq 1 + \theta + \varepsilon_j, \forall j, \end{aligned} \tag{6}$$

where $\tau \in [0, 1)$ is a parameter balancing two different deviations (greater or less than the margin means). $\theta \in [0, 1)$ is a zero-loss parameter which can control the number of support vectors, i.e., the sparseness of the solution. $(1-\theta)^2$ is a substitution loss used to change the aforesaid second item into a 0-1 loss function.

The regular term is $\Omega(\omega) = \sum_{l=1}^{k}(\|\omega_l\|_{\mathrm{H}}^2/2)$, and the mcODM is ultimately expressed as follows:

$$\begin{aligned} \min_{\omega_l,\xi_j,\varepsilon_j} \quad & \frac{1}{2}\sum_{l=1}^{k}\|\omega_l\|_{\mathrm{H}}^2 + \frac{\lambda}{m}\sum_{j=1}^{m}\frac{\xi_j^2 + \tau\varepsilon_j^2}{(1-\theta)^2}, \\ \text{s.t.} \quad & \omega_{y_j}^{\mathrm{T}}\varphi\left(x_j\right) - \max_{l\neq y_j}\omega_l^{\mathrm{T}}\varphi\left(x_j\right) \geq 1 - \theta - \xi_j, \\ & \omega_{y_j}^{\mathrm{T}}\varphi\left(x_j\right) - \max_{l\neq y_j}\omega_l^{\mathrm{T}}\varphi\left(x_j\right) \leq 1 + \theta + \varepsilon_j, \forall j, \end{aligned} \tag{7}$$

where λ, τ, and θ are the aforementioned balance parameters.

The parameters are selected by the grid search method, among which λ is determined in the sequence $[2^0, 2^2, 2^4, \ldots, 2^{20}]$ and τ and θ in $[0.2, 0.4, 0.6, 0.8]$.

3.4. Evaluation Criteria for Fault Detection Performance. To evaluate the fault detection performance of the model, a confusion matrix [29] is introduced, as defined in Table 1.

The following five evaluation indicators are obtained via the confusion matrix:

$$\begin{aligned} \text{accuracy} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}, \\ \text{precision} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \\ F1 - \text{score} &= \frac{2}{(1/\text{precision}) + (1/\text{recall})}, \\ \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}, \\ \mathrm{FNR} &= \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}. \end{aligned} \tag{8}$$

Table 1: Confusion matrix.

	Number of predicted faulty samples	Number of predicted normal samples
Number of actual faulty samples (P)	TP	FN
Number of actual normal samples (N)	FP	TN

Note. TP: the number of the samples classified as faulty samples and predicted to be faulty in the sample set; FP: the number of the samples classified as faulty samples and predicted to be normal in the sample set; TN: the number of the samples classified as normal samples and predicted to be normal in the sample set; FN: the number of the samples classified as normal samples and predicted to be faulty in the sample set.

Table 2: Some data of faulty fan on July 23, 2016.

Status parameter	Time									
	0:38:02	0:40:14	0:46:18	0:55:46	1:01:42	1:05:16	1:16:22	1:21:08	1:32:34	1:40:22
Rotor speed (r/m)	17.38	16.97	9.36	1.23	0	0	0	0	0.21	0.14
Generator speed (r/m)	1763.6	1702.8	1214.6	27.3	8.9	7.3	9.0	6.1	953.4	1651.6
Temperature of bearing A (°C)	43.8	42.6	44.0	47.7	48.5	48.4	47.1	46.7	454	45.2
Temperature of bearing B (°C)	45.1	45.8	46.2	44.5	48.8	47.3	46.4	45.5	46.8	47.1
Current of pitch motor 1 (A)	120	80	0	0	0	0	0	0	0	40
Current of pitch motor 2 (A)	70	50	20	0	0	0	0	0	30	20
Current of pitch motor 3 (A)	50	40	40	0	0	0	0	0	10	50
Brake pressure (N)	0	27.53	110.14	124.47	160.29	143.21	120.45	150.41	140.12	134.47
1 min average wind speed (m/s)	8.32	9.95	10.39	8.87	9.81	10.32	11.14	9.48	8.22	9.13
U1 phase winding current (A)	890	832	32	6	6	6	6	6	4	4
U2 phase winding current (A)	883	838	37	6	6	6	6	6	4	4
U3 phase winding current (A)	880	828	35	6	6	6	6	6	4	4
Lubricant filter inlet pressure (N)	−3.96	−3.78	−3.67	−3.54	−3.66	−3.92	−3.96	−3.88	−3.68	−3.86
Lubricant filter outlet pressure (N)	5.77	6.04	3.03	2.78	4.66	3.45	6.20	2.97	3.42	2.64
Pressure of main hydraulic system (N)	143.12	151.24	160.14	152.13	130.24	163.48	153.46	143.34	132.04	130.19
Wind angle (°)	4.21	3.24	5.25	3.54	2.14	1.67	3.45	2.87	2.21	4.15
Grid voltage (kV)	407.2	406.4	405.7	407.2	403.1	401.8	406.7	405.3	404.1	405.9
Pitch angle 1 (°)	0.25	0.33	89.04	89.04	89.04	89.04	89.04	89.04	89.04	89.04
Pitch angle 2 (°)	0.27	0.34	89.02	89.02	89.02	89.02	89.02	89.02	89.02	89.02
Pitch angle 3 (°)	0.27	0.33	89.04	89.04	89.04	89.04	89.04	89.04	89.04	89.04
Pitch controller location (°)	56.21	52.47	0	0	0	0	0	0	0	0
Voltage of pitch capacitor (V)	59.23	59.17	59.10	59.11	59.14	59.18	59.17	59.21	59.33	59.12
Yaw speed (°/s)	0	0.12	0	0	0	0	0	0	0	0
Nacelle vibration (mm)	0.06	0.04	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.03
Gearbox inlet oil temperature (°C)	46.5	44.1	45.6	43.9	43.2	44.5	44.7	43.5	42.3	44.9
Generator torque deviation (N·m)	56.63	44.12	23.14	17.19	0.12	0.11	0	0	0	0

4. Experimental Analysis

4.1. Data Description. In order to verify the effectiveness of the proposed fault detection method, the actual operating data of one wind farm in Shandong in one year was used in the experiment. This wind farm includes 33 variable-speed and variable-pitch wind turbine generator systems in total, which are separately connected to the monitoring center through sensors. The data were sampled at intervals of 2 s and stored in the database.

Among them, the main power supply of the pitch system of the #11 wind turbine generator system failed on March 14, 2016. The failure lasted from 0 : 43 to 1 : 29. The data from 30 min before the fault to 30 min after the fault were selected as the experimental data, to effectively classify samples and fully reflect fault features. Accordingly, the status parameters were selected from 0 : 13 to 1 : 59 on March 14. Part of the original data is given in Table 2.

4.2. Selection of Sample Features. According to the operation mechanism of the wind turbine generator system, when the pitch system fails, the status parameter affected directly is the power output of the wind turbine generator system. The correlation between the power output and each variable was analyzed based on the Pearson correlation coefficient, to select effective variables.

The raw data of the aforesaid status parameters were first subject to data cleaning, to eliminate "no data" and the data corresponding to the value "0" of all status variables. After the data were normalized, the correlation with the output power was calculated. Some calculation results are given in Table 3.

As can be seen from the correlation results in Table 3, some variables of the status parameters have a low correlation with the output power. Based on the nature of the Pearson correlation coefficient, the variables with the absolute value of the correlation coefficient less than 0.55 were deleted, and those with the absolute value of the correlation coefficient greater than 0.55 were taken as the main influencing factors of the fault, as indicated by the bold part in Table 3. In order to prevent model overfitting due to the interference of redundant variables in model training, these status variables were subject to a secondary calculation with the Pearson correlation coefficient to identify the redundant parameters that have a high correlation and simplify the sample size. Some secondary Pearson calculation results are given in Table 4.

Table 3: Some correlation calculation results 1.

Status parameter	Pearson correlation coefficient
Rotor speed (r/m)	**0.97753**
Generator speed (r/m)	**−0.99738**
Temperature of bearing A (°C)	0.33852
Temperature of bearing B (°C)	−0.26418
Current of pitch motor 1 (A)	**0.79989**
1 min average wind speed (m/s)	−0.20292
U1 phase winding current (A)	**0.99992**
U2 phase winding current (A)	**0.99988**
Pitch angle 2 (°)	**−0.94072**
Pitch angle 3 (°)	**−0.95908**
Yaw speed (°/s)	0.03103
Voltage of pitch capacitor (V)	**0.56784**
Pressure of main hydraulic system (N)	0.02426
U3 phase winding current (A)	**0.99985**
Current of pitch motor 2 (A)	**0.80794**
Lubricant filter inlet pressure (N)	−0.02909
Lubricant filter outlet pressure (N)	**−0.81853**
Current of pitch motor 3 (A)	**0.80471**
Wind angle (°)	0.21073
Grid voltage (kV)	**−0.88644**
Pitch angle 1 (°)	**−0.94581**
Nacelle vibration (mm)	**0.58414**
Brake pressure (N)	0.12353
Gearbox inlet oil temperature (°C)	−0.25683
Generator torque deviation (N·m)	**−0.85781**
Pitch controller location (°)	**−0.85769**

Table 4: Some correlation calculation results 2.

Pearson correlation coefficient	Pitch angle 1	U1 phase winding current (A)	Rotor speed	Current of pitch motor 1
Yaw angle 1	**0.99998**	−0.60992	**−0.97406**	−0.83843
U2 phase winding current (A)	−0.58954	**0.99863**	0.64357	0.51048
Pitch angle 2	**0.99981**	−0.60992	**−0.97406**	−0.83843
Current of pitch motor 2	−0.85505	0.53974	0.83172	**0.91606**

As can be seen from some calculation results in Table 4, the correlation coefficient of the yaw angle 1 and pitch angle 1 of the blade was close to 1, and that of the pitch angle 2 and rotor speed was also close to 1. The same status parameters of different parts also had a high correlation. They essentially had the same effect during the operation of the pitch system. If these status parameters are considered simultaneously in a model, redundant variables will be introduced, which will increase the complexity and calculation of the model and may lead to overfitting and other problems. Therefore, the redundant parameters were eliminated in conjunction with the correlation results in Tables 3 and 4. The sample feature set was built with the remaining status parameters.

4.3. Experimental Results. The sample set corresponding to the normal operation of the wind turbine generator systems was classified as a normal category and that corresponding to the failure of the main power supply of the pitch system as a fault category. The entire sample set was divided into two parts: training set and testing set, including normal and fault data, respectively. The training set was used to train the mcODM model, while the testing set to test the model. The one-versus-rest SVM (ovrSVM) and one-versus-one SVM (ovoSVM) were compared in the experiment on the Matlab platform.

According to the performance evaluation indicators of the model, five indicators were compared, i.e., the accuracy, precision, $F1$-score, false-negative rate, and false-positive rate. The test set was subject to tenfold cross-validation, using the average of results.

The comparison results of accuracy and precision are given in Table 5, the box chart of accuracy is presented in Figure 2, and the box chart of precision is presented in Figure 3. The comparison results of $F1$-score and FPR and FNR are given in Table 6.

As shown above, the accuracy, precision, and $F1$-score of the mcODM model were higher than those of the other two models, while its false-negative rate and false-positive rate were the lowest.

In order to verify the universality of the method proposed in this paper, the operating data of multiple wind turbine generator systems with failure in their pitch systems in this wind farm were used in the experiment. The number of 23 wind turbine generator system occurred the overtemperature of the servo drive of the pitch blade 1 on July 23, 2016, in this wind farm, the comparison results of accuracy and precision are given in Table 7, the box chart of accuracy

Table 5: Comparison of fault detection performance for the main power supply of pitch system 1.

Model	Accuracy	Precision
mcODM	**96.11% (±0.0345)**	**95.09% (±0.0185)**
ovrSVM	93.84% (±0.0237)	92.62% (±0.0095)
ovoSVM	91.32% (±0.0514)	90.77% (±0.0126)

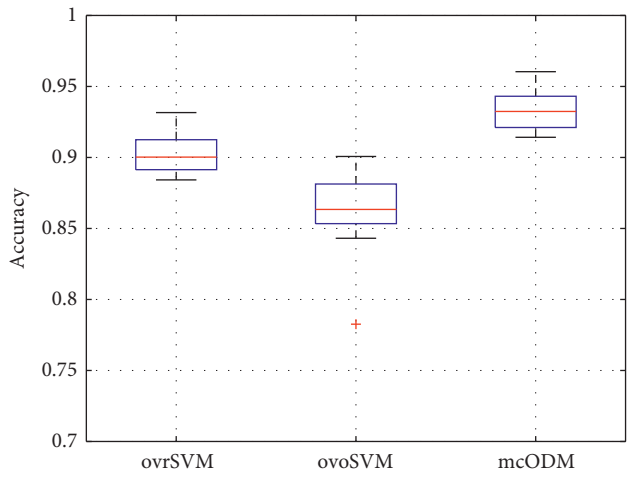


Figure 2: Box chart of fault detection accuracy for the main power supply of the pitch system.

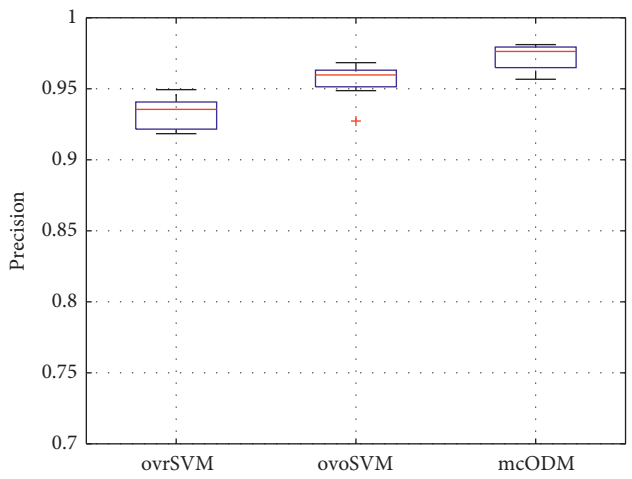


Figure 3: Box chart of fault detection precision for the main power supply of the pitch system.

Table 6: Comparison of fault detection performance for the main power supply of pitch system 2.

Model	$F1$-score	FPR	FNR
mcODM	**96.03% (±0.0027)**	**5.01% (±0.0127)**	**3.01% (±0.0116)**
ovrSVM	93.75% (±0.0013)	8.26% (±0.0084)	6.43% (±0.0237)
ovoSVM	90.14% (±0.0019)	10.64% (±0.0092)	8.18% (±0.0134)

is presented in Figure 4, and the box chart of precision is presented in Figure 5. The comparison results of $F1$-score and FPR and FNR are given in Table 8. The number of 28 wind turbine generator system occurred the emergency stop of the pitch system on June 8, 2016, in this wind farm, the comparison results of accuracy and precision are given in Table 9, the box chart of accuracy is presented in Figure 6, and the box chart of precision is presented in Figure 7. The comparison results of $F1$-score and FPR and FNR are given in Table 10.

In the fault detection at the overtemperature of the servo drive of the pitch blade 1 and the emergency stop of the pitch system, the mcODM model has the highest accuracy, precision, and $F1$-score and lowest false-negative rate and false-positive rate.

It can be seen from the aforesaid comparison results that, in terms of the faults of the pitch systems of different wind turbine generator systems, the mcODM model has high efficiency in sample classification and capabilities in generalization, since the distribution model is built based on the features of the sample distribution. In conjunction with the aforesaid method of feature selection, the status parameters of low correlation can be eliminated, thus reducing the sample size and model training burden and avoiding overfitting. When the mcODM algorithm is combined with the proposed feature selection method in fault detection of pitch systems of wind turbine generator systems, higher capabilities can be achieved in fault detection.

5. Conclusions

This paper proposes the mcODM-based method for fault detection of pitch systems of wind turbine generator systems. The features are extracted according to the operating features of the pitch system and the Pearson correlation coefficient of the wind turbine generator system. The correlation of status parameters is fully considered, and the model complexity is subject to secondary Pearson analysis, which can eliminate the redundant parameters and avoid model overfitting while ensuring the detection rate. This also solves the problem of selecting the feature parameters reflecting the faults of pitch systems from a large amount of SCADA data. Considering the detection model, the mcODM model has been successfully applied in fault detection of the pitch systems of wind turbine generator systems. Due to the combination of the margin mean and variance and full consideration to the sample distribution features, this model solves the problem of inefficient classification arising from the sample category unbalance and complex distribution of pitch system fault samples.

In order to verify the universality of this method, the SCADA data of wind turbine generator systems with different pitch system faults were used in the fault detection experiment. At the same time, the ovrSVM and ovoSVM models were introduced for comparison. The experimental results show that the proposed method has good performance in generalization and high accuracy and precision as well as low false-negative rate and false-positive rate in fault detection of the pitch systems of wind turbine generator systems.

Since wind turbine generator systems are affected by multiple factors (e.g., operating environment and load) and

Table 7: Comparison of fault detection performance at overtemperature of servo drive of pitch blade 1.

Model	Accuracy	Precision
mcODM	**92.07% (±0.0214)**	**90.26% (±0.0426)**
ovrSVM	89.94% (±0.0186)	88.75% (±0.0329)
ovoSVM	86.44% (±0.0621)	86.16% (±0.0084)

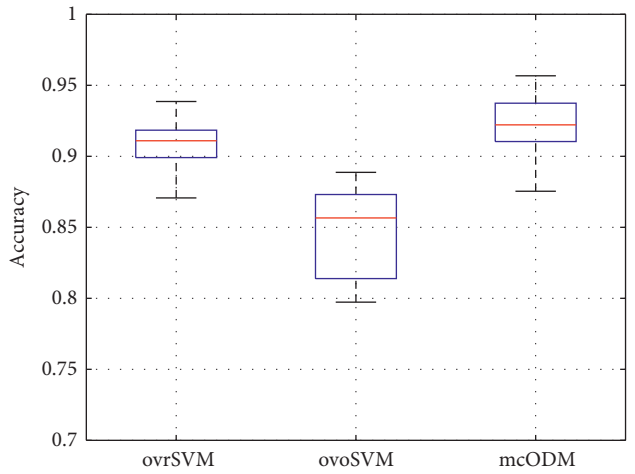


Figure 4: Box chart of fault detection accuracy at overtemperature of servo drive of pitch blade 1.

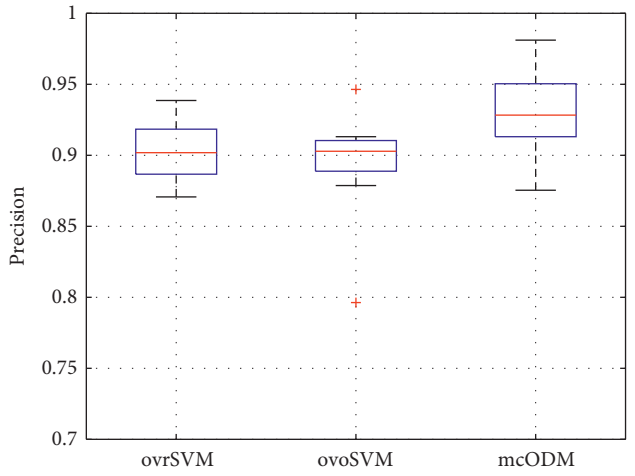


Figure 5: Box chart of fault detection precision at overtemperature of servo drive of pitch blade 1.

Table 8: Comparison of fault detection performance at overtemperature of servo drive of pitch blade 2.

Model	$F1$-score	FPR	FNR
mcODM	**91.93% (±0.0023)**	**8.24% (±0.0134)**	**6.07% (±0.0112)**
ovrSVM	89.81% (±0.0018)	11.07% (±0.0121)	8.82% (±0.0082)
ovoSVM	85.78% (±0.0012)	13.39% (±0.0219)	10.48% (±0.0148)

their operating conditions are changing during fault detection, it is difficult to meet the fault detection requirements for the entire wind turbine generator systems in most cases.

Table 9: Comparison of fault detection performance at the emergency stop of pitch system 1.

Model	Accuracy	Precision
mcODM	**94.73% (±0.0219)**	**94.41% (±0.0427)**
ovrSVM	89.14% (±0.0271)	88.83% (±0.0316)
ovoSVM	90.81% (±0.0184)	89.76% (±0.0251)

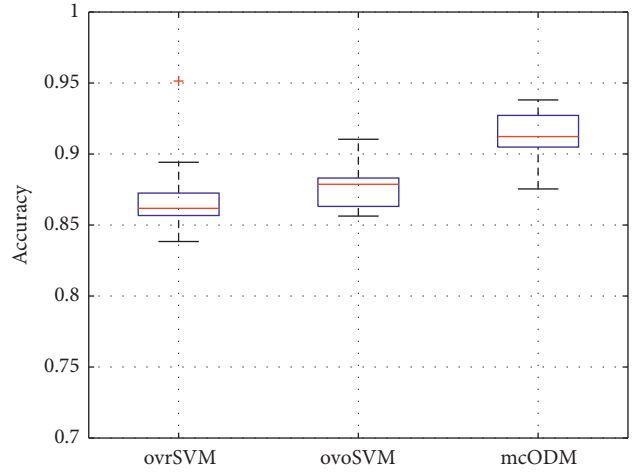


Figure 6: Box chart of fault detection accuracy at the emergency stop of the pitch system.

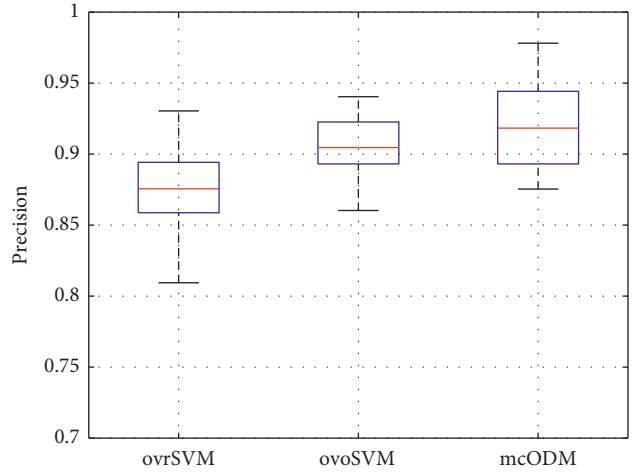


Figure 7: Box chart of fault detection precision at the emergency stop of the pitch system.

Table 10: Comparison of fault detection performance at the emergency stop of pitch system 2.

Model	$F1$-score	FPR	FNR
mcODM	**93.16% (±0.0024)**	**6.25% (±0.0121)**	**3.52% (±0.0057)**
ovrSVM	88.15% (±0.0019)	11.70% (±0.0079)	8.43% (±0.0081)
ovoSVM	90.08% (±0.0021)	9.07% (±0.0237)	7.51% (±0.0048)

Therefore, the research on status monitoring and fault detection of the entire wind turbine generator systems under changing conditions can help effectively reduce the fault rate and improve operating stability.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors contributed equally to this work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 61403046), the Natural Science Foundation of Hunan Province, China (Grant no. 2019JJ40304), Changsha University of Science and Technology "The Double First Class University Plan" International Cooperation and Development Project in Scientific Research in 2018 (Grant no. 2018IC14), the Research Foundation of Education Bureau of Hunan Province (Grant no. 19K007), Hunan Provincial Department of Transportation 2018 Science and Technology Progress and Innovation Plan Project (Grant no. 201843), the Key Laboratory of Renewable Energy Electric-Technology of Hunan Province, the Key Laboratory of Efficient and Clean Energy Utilization of Hunan Province, Innovative Team of Key Technologies of Energy Conservation, Emission Reduction and Intelligent Control for Power-Generating Equipment and System, CSUST, Hubei Superior and Distinctive Discipline Group of Mechatronics and Automobiles (Grant no. XKQ2020009), and Major Fund Project of Technical Innovation in Hubei (Grant no. 2017AAA133).

References

[1] J. Chen, F. Wang, and K. A. Stelson, "A mathematical approach to minimizing the cost of energy for large utility wind turbines," *Applied Energy*, vol. 228, pp. 1413–1422, 2018.

[2] J. Tautz-Weinert and S. J. Watson, "Using SCADA data for wind turbine condition monitoring-a review," *IET Renewable Power Generation*, vol. 11, no. 4, pp. 382–394, 2017.

[3] M. D. Hossain, A. Abu-Siada, and S. Muyeen, "Methods for advanced wind turbine condition monitoring and early diagnosis: a literature review," *Energies*, vol. 11, no. 5, p. 1309, 2018.

[4] M. Tang, S. X. Ding, C. Yang et al., "Cost-sensitive large margin distribution machine for fault detection of wind turbines," *Cluster Computing*, vol. 22, no. 3, pp. 7525–7537, 2019.

[5] L. Li, H. Luo, S. X. Ding, Y. Yang, and K. Peng, "Performance-based fault detection and fault-tolerant control for automatic control systems," *Automatica*, vol. 99, pp. 308–316, 2019.

[6] L. Li and S. X. Ding, "Optimal detection schemes for multiplicative faults in uncertain systems with application to rolling mill processes," *IEEE Transactions on Control Systems Technology*, 2020.

[7] W. Long, T. Wu, J. Jiao, M. Tang, and M. Xu, "Refraction-learning-based whale optimization algorithm for high-dimensional problems and parameter estimation of PV model," *Engineering Applications of Artificial Intelligence*, vol. 89, Article ID 103457, 14 pages, 2020.

[8] R. V. Pandit and D. Infield, "SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes," *IET Renewable Power Generation*, vol. 12, no. 11, pp. 1249–1255, 2018.

[9] J. Liu, F. Qu, X. Hong, and H. Zhang, "A small-sample wind turbine fault detection method with synthetic fault data using generative adversarial nets," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3877–3888, 2018.

[10] F. Ruiming, W. Minling, G. Xinhua, S. Rongyan, and S. Pengfei, "Identifying early defects of wind turbine based on SCADA data and dynamical network marker," *Renewable Energy*, vol. 154, pp. 625–635, 2020.

[11] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems & Signal Processing*, vol. 107, pp. 241–265, 2018.

[12] J. Dai, W. Yang, J. Cao, D. Liu, and L. Xing, "Ageing assessment of a wind turbine over time by interpreting wind farm SCADA data," *Renewable Energy*, vol. 116, no. 2, pp. 199–208, 2018.

[13] A. T. Eseye, J. Zhang, and D. Zheng, "Short-term photovoltaic solar power forecasting using a hybrid wavelet-PSO-SVM model based on SCADA and meteorological information," *Renewable Energy*, vol. 118, no. 4, pp. 357–367, 2018.

[14] H. Zheng, Y. Zhang, J. Liu, H. Wei, J. Zhao, and R. Liao, "A novel model based on wavelet LS-SVM integrated improved PSO algorithm for forecasting of dissolved gas contents in power transformers," *Electric Power Systems Research*, vol. 155, pp. 196–205, 2018.

[15] Z. Wang, L. Yao, Y. Cai, and J. Zhang, "Mahalanobis semi-supervised mapping and beetle antennae search based support vector machine for wind turbine rolling bearings fault diagnosis," *Renewable Energy*, vol. 155, pp. 1312–1327, 2020.

[16] Y. Li, S. Liu, and L. Shu, "Wind turbine fault diagnosis based on Gaussian process classifiers applied to operational data," *Renewable Energy*, vol. 134, pp. 357–366, 2019.

[17] A. Soussa, M. D. Mouss, S. Aitouche, H. Melakhessou, and M. Titah, "The MAED and SVM for fault diagnosis of wind turbine system," *International Journal of Renewable Energy Research*, vol. 7, no. 2, pp. 758–769, 2017.

[18] Q. W. Gao, W. Y. Liu, B. P. Tang, and G. J. Li, "A novel wind turbine fault diagnosis method based on intergral extension load mean decomposition multiscale entropy and least squares support vector machine," *Renewable Energy*, vol. 116, pp. 169–175, 2018.

[19] W. Liu, Z. Wang, J. Han, and G. Wang, "Wind turbine fault diagnosis method based on diagonal spectrum and clustering binary tree SVM," *Renewable Energy*, vol. 50, no. 2, pp. 1–6, 2013.

[20] J. Hang, J. Zhang, and M. Cheng, "Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine," *Fuzzy Sets and Systems*, vol. 297, pp. 128–140, 2016.

[21] J. Saari, D. Strömbergsson, J. Lundberg, and A. Thomson, "Detection and identification of windmill bearing faults using a one-class support vector machine (SVM)," *Measurement*, vol. 137, pp. 287–301, 2019.

[22] M. Tanveer, A. Sharma, and P. N. Suganthan, "General twin support vector machine with pinball loss function," *Information Sciences*, vol. 494, pp. 311–327, 2019.

[23] T. Zhang and Z. Zhou, "Multi-class optimal margin distribution machine," in *Proceedings of the International*

Conference on Machine Learning, pp. 4063–4071, Sydney, Australia, 2017.

[24] E. Muljadi and C. P. Butterfield, "Pitch-controlled variable-speed wind turbine generation," *IEEE Transactions on Industry Applications*, vol. 37, no. 1, pp. 240–246, 2001.

[25] X. Yin, W. Zhang, Z. Jiang, and L. Pan, "Adaptive robust integral sliding mode pitch angle control of an electro-hydraulic servo pitch system for wind turbine," *Mechanical Systems and Signal Processing*, vol. 133, Article ID 105704, 2019.

[26] H. Li, C. Yang, Y. Hu, X. Liao, Z. Zeng, and C. Zhe, "An improved reduced-order model of an electric pitch drive system for wind turbine control system design and simulation," *Renewable Energy*, vol. 93, no. 8, pp. 188–200, 2016.

[27] Q. Jiang, S. Yan, H. Cheng, and X. Yan, "Local-global modeling and distributed computing framework for nonlinear plant-wide process monitoring with industrial big data," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[28] Q. Jiang, S. Yan, X. Yan, H. Yi, and F. Gao, "Data-driven two-dimensional deep correlated representation learning for nonlinear batch process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2839–2848, 2020.

[29] M. Tang, Q. Zhao, S. X. Ding et al., "An improved lightGBM algorithm for online fault detection of wind turbine gearboxes," *Energies*, vol. 13, no. 4, Article ID 807, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 1247079, 11 pages
https://doi.org/10.1155/2020/1247079

Research Article

Composite Compensation Control of Robotic System Subject to External Disturbance and Various Actuator Faults

Hao Sheng and Xia Liu

School of Electrical Engineering and Electronic Information, Xihua University, Chengdu 610039, China

Correspondence should be addressed to Xia Liu; xliu_uestc@yahoo.com

Received 2 May 2020; Revised 24 June 2020; Accepted 27 June 2020; Published 26 July 2020

Guest Editor: Esam Hafez Abdelhameed

This paper studies the problems of external disturbance and various actuator faults in a nonlinear robotic system. A composite compensation control scheme consisting of adaptive sliding mode controller and observer-based fault-tolerant controller is proposed. First, a sliding mode controller is designed to suppress the external disturbance, and an adaptive law is employed to estimate the bound of the disturbance. Next, a nonlinear observer is designed to estimate the actuator faults, and a fault-tolerant controller is obtained based on the observer. Finally, the composite compensation control scheme is obtained to simultaneously compensate the external disturbance and various actuator faults. It is proved by Lyapunov function that the disturbance compensation error and fault compensation error can converge to zero in finite time. The theoretical results are verified by simulations. Compared to the conventional fault reconstruction scheme, the proposed control scheme can compensate the disturbance while dealing with various actuator faults. The fault compensation accuracy is higher, and the fault error convergence rate is faster. Moreover, the robot can track the desired position trajectory more accurately and quickly.

1. Introduction

Robotic system is a complex nonlinear system with the characteristics of multiple variables, high nonlinearity, and strong coupling. In robotic system, there are a variety of problems, such as external disturbance and actuator fault. The position tracking performance of the robot will decrease due to disturbance. Meanwhile, the controller needs to tolerate actuator fault to keep the robotic system stable [1–3]. Therefore, disturbance and actuator fault are two of the main issues to be solved in robot control.

For robotic system with disturbance, sliding mode control has been widely applied due its robustness to disturbance and uncertainty [4]. However, there are some drawbacks in conventional sliding mode control. For example, the error cannot converge in finite time, and there exits chattering phenomenon. In addition, the upper bound of the disturbance needs to be known. In order to avoid the drawbacks in conventional sliding mode control, observer is one of the effective approaches. In [5], a composite controller based on a nonlinear controller and a nonlinear disturbance observer was proposed for nonlinear systems, where the observer was employed to estimate the disturbance generated by an exogenous system. In [6], the external disturbance in a nonlinear system was viewed as an unknown input. An adaptive extended state observer was designed to estimate the unknown input, and then, a controller was designed to compensate the external disturbance using the estimated value. In [7], for the unknown matched and mismatched time-varying disturbances in a robotic system, a continuous sliding mode control based on generalized proportional integral observer was proposed. The observer was to estimate the matched disturbance and mismatched disturbance, respectively. The continuous sliding mode manifold was to remove the offset caused by the mismatched disturbance. In [8], the uncertain hydrodynamics and unknown external disturbance in an underwater robotic system were regarded as a lumped disturbance. An integral sliding mode controller based on extended state observer was presented. The extended state observer was to

estimate the lumped disturbance and unmeasurable states, and the adaptive gain update algorithm was to estimate the bound of the lumped disturbance. In [9], the model errors, uncertainties, friction, and unknown external disturbances in automobile electrocoating conveying mechanism were all regarded as a lumped disturbance. A nonlinear disturbance observer was to estimate the lumped disturbance, and a sliding mode controller was designed for the hybrid series-parallel mechanism. Although the approaches in [5–9] can effectively deal with the disturbance in the system, they all potentially assume that all the actuators in the system are working normally without any fault.

In fact, in addition to external disturbance, many mechanical systems and electronic devices, such as sensors, actuators, and amplifiers, may undergo fault due to aging, affecting the performance and even safety of the system [10–12]. In order to ensure the performance and safety of the system when actuator fault occurs, different fault-tolerant control schemes have been proposed. In [13], a fault reconstruction scheme based on terminal sliding mode observer and fault-tolerant control was proposed for robotic manipulators. The fault reconstruction error can converge to zero in finite time. Nevertheless, only actuator fault was considered. In [14], for external disturbance and actuator fault in manipulator, a fault-tolerant control based on adaptive dynamic sliding mode was proposed. However, only loss of effectiveness fault was considered. In [15], actuator faults and friction in a robotic system were regarded as total uncertain dynamics. A sliding mode observer was designed to estimate the total uncertain dynamics. A nonlinear observer was used to reconfigure the uncertainty. However, since the fault and friction were regarded as total uncertain dynamics, their respective characteristic cannot be reflected. In [16], actuator faults and external collision in robot manipulator were regarded as centralized disturbance. A sliding mode observer was used to estimate the velocity and centralized disturbance. A protective control framework based on disturbance reconstruction was proposed. Nevertheless, the characteristic of fault was not formally described in [16]. In [17], for robots subject to unmatched disturbance and actuator fault, a fault-tolerant adaptive control based on disturbance observer and backstepping control was proposed. Nevertheless, the disturbance error cannot converge to zero in finite time, and the error convergence rate was slow. In [18, 19], for actuator fault, matched or unmatched disturbance in a class of uncertain nonlinear systems, an active fault-tolerant control was designed based on integral-type sliding mode control. However, since active fault-tolerant control was based on fault information, delay of the fault information feedback will result in delay of the fault compensation time. Consequently, the system may become unstable. In [20], actuator fault, external disturbance, and input saturation were regarded as total uncertainty for the robotic system, and a finite-time fault-tolerant adaptive robust control strategy was proposed. The total uncertainty was estimated by the adaptive law, and then, a fault-tolerant adaptive robust controller was obtained by the integral backstepping control. However, as actuator fault, external disturbance, and input saturation in the system were treated as total uncertainty, and their respective characteristic could not be reflected well. Moreover, only time-varying fault was considered in [20].

In this paper, a composite compensation control approach is proposed for a nonlinear robotic system with external disturbance and various actuator faults. The proposed composite compensation controller consists of an adaptive sliding mode controller and an observer-based fault-tolerant controller. Compared to the conventional fault reconstruction scheme, the proposed control can compensate disturbance while dealing with various actuator faults, including no fault, loss of effectiveness fault, and floating around trim fault. The fault compensation accuracy is higher, and the fault error convergence rate is faster. Moreover, the robot can track the desired position trajectory more accurately and quickly.

The remainder of this paper is organized as follows: in Section 2, the model of robotic system subject to external disturbance and actuator faults is formally described; in Section 3, the composite compensation control is designed based on adaptive sliding mode control and observer-based fault-tolerant control, and the convergence of the disturbance compensation error and fault compensation error is proved; simulations are provided in Section 4; and the paper is concluded in Section 5.

2. Model of Robotic System Subject to External Disturbance and Actuator Faults

A nonlinear robotic system with external disturbance and actuator faults is considered in this paper, as shown in Figure 1.

2.1. Model of Robotic System Subject to External Disturbance. The dynamic model of a n-DOF nonlinear robot subject to external disturbance can be described as follows [21]:

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) + \tau_d = u, \tag{1}$$

where $q \in R^{n\times 1}$, $\dot{q} \in R^{n\times 1}$, and $\ddot{q} \in R^{n\times 1}$ represent the joint position, joint velocity, and joint acceleration of the robot, respectively; $M(q) \in R^{n\times n}$, $C(q,\dot{q}) \in R^{n\times n}$, and $G(q) \in R^{n\times 1}$ represent the inertia matrix, Coriolis and centrifugal term, and gravity term. $u \in R^{n\times 1}$ is the control torque, and $\tau_d \in R^{n\times 1}$ denotes the external disturbance.

In practical applications, the external disturbance of a system is usually bounded [22], i.e.,

$$\|\tau_d\| \le F, \tag{2}$$

where F is an unknown constant.

For the dynamic model of the robot (1), there are two important properties.

Property 1. The inertia matrix $M(q)$ is symmetric and positive definite which satisfies

$$\lambda_0\|\xi\|^2 \le \xi^T M(q)\xi \le \lambda_1\|\xi\|^2, \tag{3}$$

where λ_0 and λ_1 are positive constants and $\forall \xi \in R^{n\times 1}$.

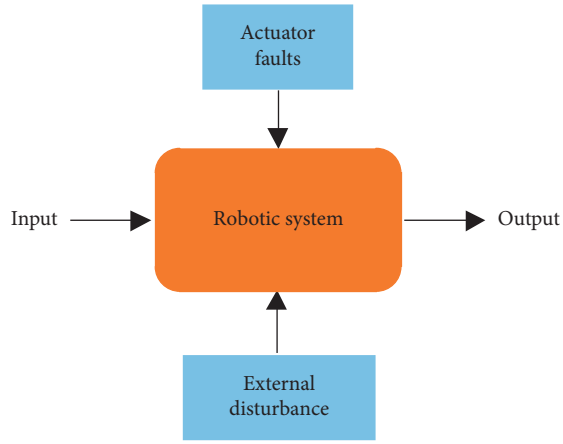


Figure 1: Robotic system with external disturbance and actuator faults.

Property 2. The matrix $\dot{M}(q) - 2C(q,\dot{q})$ is skew symmetric, i.e., $\xi^T(\dot{M}(q) - 2C(q,\dot{q}))\xi = 0$, $\forall \xi \in R^{n\times 1}$.

2.2. Model of Actuator Faults. In a practical robotic system, the actuators may undergo fault due to aging, affecting the performance and even safety of the system. The mathematical model of actuator faults can be described as follows [13]:

$$u_f(T) = u - u_{\text{nom}}, \tag{4}$$

where $u_f(T) \in R^{n\times 1}$ represents the actuator fault and $u_{\text{nom}} \in R^{n\times 1}$ represents the control torque from the nominal controller. Besides, $T = [\,T_1 \;\; T_2 \;\; \ldots \;\; T_n\,]^T \in R^{n\times 1}$ is the fault time-profile, where $T_i (i = 1, 2, \ldots, n)$ denotes the time at which the i^{th} actuator undergoes fault. Generally, there are four types of actuator faults [23]:

(i) No fault: the controller is the nominal controller, i.e., $u = u_{\text{nom}}$ and $u_f(T) = 0$.

(ii) Locked-in-place fault: the actuator fault is a constant, and the nominal controller is zero, i.e., $u = u_f(T)$, $u_{\text{nom}} = 0$, and $u_f(T)$ is a constant.

(iii) Loss of effectiveness fault: it means $u = D(t)u_{\text{nom}} + u_f(T)$. $u \in R^{n\times 1}$ is the actual control generated by the actuator. $D(t) = \text{diag}[l_1(t), l_2(t), \ldots, l_n(t)]$ denotes the effectiveness of the actuator, where $0 < l_i(t) \le 1$ means that the i^{th} actuator experiences a partial loss of effectiveness, and $u_f(T) = 0$, $i = 1, 2, \ldots, n$.

(iv) Floating around trim fault: it can be accounted as $u = u_{\text{nom}} + u_f(T)$ and $u_f(T) \neq 0$.

3. Composite Compensation Control of Robotic System

For robotic system subject to external disturbance (1) and actuator faults (4), the structure of the proposed composite compensation control scheme is shown in Figure 2. First, a sliding mode controller is designed to suppress the external disturbance τ_d. An adaptive law is employed to estimate the bound of the disturbance and obtain its estimation $\hat{F}$. Then, a nonlinear observer is designed to directly estimate the state vector of the nonlinear function and obtain its estimation $\hat{\alpha}(t)$ such that the actuator faults $u_f(T)$ can be indirectly estimated. A fault-tolerant controller is obtained based on the observer to compensate the actuator faults. Finally, the composite compensation controller u_{com} is composed of the adaptive sliding mode controller τ_{dcom} and observer-based fault-tolerant controller τ_{fcom}. Furthermore, the actual controller u is obtained by combining the composite compensation controller u_{com} and the nominal controller u_{nom}. In this way, the external disturbance and various actuator faults can be accurately compensated, and the real position q of the robot can accurately track the desired position q_d.

3.1. Design of the Adaptive Sliding Mode Controller. Take $x_1 = q \in R^{n\times 1}$ and $x_2 = \dot{q} \in R^{n\times 1}$ as the state variables of the system, and (1) can be directly rewritten into the state-space form as

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = M(x_1)^{-1}(u - \tau_d - C(x_1, x_2)x_2 - G(x_1)). \end{cases} \tag{5}$$

Define the sliding manifold as

$$s = x_2 - \phi, \tag{6}$$

where ϕ is the state of the following nonlinear system (7):

$$\begin{aligned} \dot{\phi} = M(x_1)^{-1}\left(k_1 s + \hat{F}\,\text{sign}(s) + k_2|s|^{n_1/n_2} \cdot \text{sign}\left(|s|^{n_1/n_2}\right)\right) \\ + u - G(x_1) - C(x_1, x_2)x_2 + C(x_1, x_2)s), \end{aligned} \tag{7}$$

where $k_1 > 0$ and $k_2 > 0$ are positive constants and $n_1 > 0$ and $n_2 > 0$ are two odd integers satisfying $n_1 < n_2$.

In order to suppress the external disturbance τ_d in (5), the sliding mode controller τ_{dcom} can be designed as

$$\tau_{\text{dcom}} = -k_1 s - \hat{F}\,\text{sign}(s) - k_2|s|^{n_1/n_2} \cdot \text{sign}\left(|s|^{n_1/n_2}\right). \tag{8}$$

As the bound of the external disturbance is usually unknown, an adaptive law is designed to estimate the bound F:

$$\dot{\hat{F}} = \beta s^T \text{sign}(s), \tag{9}$$

where $\hat{F}$ is the estimation of F, $\beta > 0$ is a positive constant, and sign is signum function.

3.2. Design of the Observer-Based Fault-Tolerant Controller. Let us introduce a new vector $M_a(q) = M(q)\dot{q} - \int_0^t e_d(l)\text{d}l$, where $e_d = \tau_{\text{dcom}} - \tau_d$ is the disturbance compensation error. Then, from (1) and (4), we can get

$$\dot{M}_a(q) = u_{\text{nom}} + u_f(T) - \omega(q, \dot{q}) - \tau_{\text{dcom}}, \tag{10}$$

where $\omega(q, \dot{q}) = C(q, \dot{q})\dot{q} + G(q) - \dot{M}(q)\dot{q}$.

Now, define a new nominal controller $u_{\text{anom}} = u_{\text{nom}} + \tau_{\text{dcom}}$ and a new variable

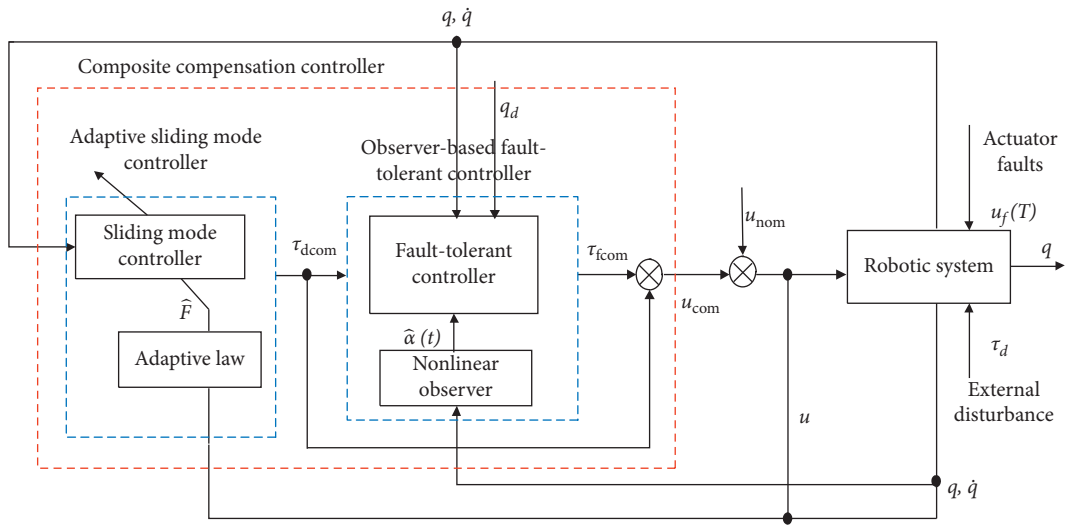


Figure 2: Structure of composite compensation control.

$$\alpha(t) = k_3 \int_0^t [u_{\text{anom}} - \omega(q,\dot{q}) - 2\tau_{\text{dcom}} - \alpha(x)]\mathrm{d}x - k_3 M_a(q), \tag{11}$$

where $k_3 > 0$ is a constant and $\alpha(t)$ is the state vector of the nonlinear function (11).

Differentiating (11) with respect to time and substituting (1) and (10) into it, we have

$$\dot{\alpha}(t) = -k_3\alpha(t) - k_3 u_f(T), \tag{12}$$

where $u_f(T)$ can be regarded as the unknown input of the system (12).

As for the output of the system (12), we can take it as

$$y = k_4\alpha(t), \tag{13}$$

where $k_4 > 0$ is a positive constant.

Now, the fault $u_f(T)$ can be indirectly estimated by directly estimating the state $\alpha(t)$ of the system (12) through the following nonlinear observer

$$\dot{\hat{\alpha}}(t) = -k_3\hat{\alpha}(t) + \frac{1}{k_4}\dot{y} + k_5 y + k_6|e|^{n_1/n_2}, \tag{14}$$

where $\hat{\alpha}(t)$ is the estimation of $\alpha(t)$, $e = \alpha - \hat{\alpha}$ denotes the observation error of $\alpha(t)$, and $k_5 = k_3/k_4$ and $k_6 > 0$ are observation gains.

Since $\alpha(t)$ can be estimated by the nonlinear observer (14), the fault-tolerant controller can be designed as

$$\tau_{\text{fcom}} = -\hat{\alpha}(t) - \frac{1}{k_3 k_4}\dot{y}. \tag{15}$$

3.3. Design of the Composite Compensation Controller. With the adaptive sliding mode controller (8)-(9) and the observer-based fault-tolerant controller (14)-(15), the composite compensation controller u_{com} can be designed as

$$u_{\text{com}} = \tau_{\text{dcom}} + \tau_{\text{fcom}} = -k_1 s - \hat{F}\,\text{sign}(s) - k_2|s|^{n_1/n_2} \cdot \text{sign}\left(|s|^{n_1/n_2}\right) - \hat{\alpha}(t) - \frac{1}{k_3 k_4}\dot{y}. \tag{16}$$

The composite compensation controller (16) can simultaneously compensate external disturbance and various types of actuator faults.

Theorem 1. *Consider the nonlinear robotic system subject to external disturbance (1) and actuator faults (4). If it is controlled by the composite compensation controller (16), which is composed of the adaptive sliding mode controller (8)-(9) and the observer-based fault-tolerant controller (14)-(15), then the disturbance compensation error and fault compensation error of the robotic system can converge to zero in finite time, i.e.,* $\lim_{t\to t_c} e_d = \lim_{t\to t_c}(\tau_{dcom} - \tau_d) = 0$ *and* $\lim_{t\to t_c} e_f = \lim_{t\to t_c}(\tau_{fcom} - u_f(T)) = 0$.

Proof. Differentiating the observation error $e = \alpha - \hat{\alpha}$ with respect to time and substituting (12)–(14) into it, we can obtain

$$\begin{aligned}\dot{e} &= \dot{\alpha} - \dot{\hat{\alpha}} \\ &= -k_3\alpha - k_3 u_f + k_3\hat{\alpha} - \frac{1}{k_4}\dot{y} - k_5 y - k_6|e|^{n_1/n_2} \\ &= -k_3 e - k_6|e|^{n_1/n_2}.\end{aligned} \tag{17}$$

Define a Lyapunov function as

$$V_1 = \frac{1}{2}e^T e + \frac{1}{2}s^T M(x_1)s. \tag{18}$$

Differentiating (18) with respect to time and substituting (5)–(7) into it, we get

$$\begin{aligned}\dot{V}_1 &= e^T\dot{e} + \frac{1}{2}s^T\dot{M}(x_1)s + s^TM(x_1)\dot{s}\\ &= e^T\dot{e} + \frac{1}{2}s^T\dot{M}(x_1)s + s^TM(x_1)\left[M(x_1)^{-1}\right.\\ &\quad\cdot\left(u - \tau_d - C(x_1,x_2)x_2\right.\\ &\quad\left.-G(x_1)\right) - M(x_1)^{-1}\left(k_1s + F\mathrm{sign}(s) + k_2|s|^{n_1/n_2}\right.\\ &\quad\left.\cdot\,\mathrm{sign}\left(|s|^{n_1/n_2}\right) + u - G(x_1) - C(x_1,x_2)x_2 + C(x_1,x_2)s\right)]\\ &= e^T\dot{e} + \frac{1}{2}s^T\dot{M}(x_1)s - s^T\tau_d - k_1s^Ts - Fs^T\mathrm{sign}(s)\\ &\quad - s^TC(x_1,x_2)s - k_2s^T\cdot\left(|s|^{n_1/n_2}\cdot\mathrm{sign}\left(|s|^{n_1/n_2}\right)\right).\end{aligned} \tag{19}$$

Substituting (17) into (19) and using Property 2, we can get

$$\begin{aligned}\dot{V}_1 &= e^T\left(-k_3e - k_6|e|^{n_1/n_2}\right) + \frac{1}{2}s^T\dot{M}(x_1)s - s^T\tau_d - k_1s^Ts\\ &\quad - Fs^T\mathrm{sign}(s)\\ &\quad - s^TC(x_1,x_2)s - k_2s^T\cdot\left(|s|^{n_1/n_2}\cdot\mathrm{sign}\left(|s|^{n_1/n_2}\right)\right)\\ &= -k_3e^Te - k_6e^T|e|^{n_1/n_2} - s^T\tau_d - k_1s^Ts - Fs^T\mathrm{sign}(s)\\ &\quad - k_2s^T\cdot\left(|s|^{n_1/n_2}\cdot\mathrm{sign}\left(|s|^{n_1/n_2}\right)\right).\end{aligned} \tag{20}$$

According to Property 1 and Property 2, (20) becomes

$$\begin{aligned}\dot{V}_1 &\le -2k_3\left(\frac{1}{2}e^Te\right) - 2^{(n_1+n_2)/2n_2}k_6\left(\frac{1}{2}e^Te\right)^{(n_1+n_2)/2n_2}\\ &\quad + \|s\|\left\|\tau_d\right\| - \frac{2k_1}{\lambda_1}\left(\frac{1}{2}s^TM(x_1)s\right)\\ &\quad - F\|s\| - k_2\left(\frac{2}{\lambda_1}\right)^{(n_1+n_2)/2n_2}\left(\frac{1}{2}s^TM(x_1)s\right)^{(n_1+n_2)/2n_2}\\ &\le -2k_3\left(\frac{1}{2}e^Te\right) - 2^{(n_1+n_2)/2n_2}k_6\left(\frac{1}{2}e^Te\right)^{(n_1+n_2)/2n_2}\\ &\quad - \frac{2k_1}{\lambda_1}\left(\frac{1}{2}s^TM(x_1)s\right)\\ &\quad - k_2\left(\frac{2}{\lambda_1}\right)^{(n_1+n_2)/2n_2}\left(\frac{1}{2}s^TM(x_1)s\right)^{(n_1+n_2)/2n_2}.\end{aligned} \tag{21}$$

Now, let $k_7 = 2k_3$, $k_8 = 2k_1/\lambda_1$, $k_9 = 2^{(n_1+n_2)/2n_2}k_6$, and $k_{10} = (2/\lambda_2)^{(n_1+n_2)/2n_2}k_2$, and we can further obtain

$$\begin{aligned}\dot{V}_1 &\le -k_7\left(\frac{1}{2}e^Te\right) - k_8\left(\frac{1}{2}s^TM(x_1)s\right) - k_9\left(\frac{1}{2}e^Te\right)^{(n_1+n_2)/2n_2}\\ &\quad - k_{10}\left(\frac{1}{2}s^TM(x_1)s\right)^{(n_1+n_2)/2n_2}\\ &\le -c_1\left(\frac{1}{2}e^Te + \frac{1}{2}s^TM(x_1)s\right)\\ &\quad - c_2\left(\frac{1}{2}e^Te + \frac{1}{2}s^TM(x_1)s\right)^{(n_1+n_2)/2n_2}\\ &= -c_1V_1 - c_2V_1^{(n_1+n_2)/2n_2},\end{aligned} \tag{22}$$

where $c_1 = \min\{k_7, k_8\}$ and $c_2 = \min\{k_9, k_{10}\}$ and $0 < ((n_1+n_2)/2n_2) < 1$. Solving (22) leads to $V_1(t) \equiv 0$ for all $t \ge t_c$. Therefore, from (22), it is easy to show that $\dot{V}_1(t) \le 0$ and the finite time t_c can be obtained as

$$t_c \le \frac{2n_2}{c_1(n_2-n_1)}\ln\frac{c_1V_1^{(n_2-n_1)/2n_2}(0) + c_2}{c_2}. \tag{23}$$

Now, define another Lyapunov function as

$$V_2 = V_1 + \frac{1}{2\beta}\widetilde{F}^2, \tag{24}$$

where $\widetilde{F} = F - \widehat{F}$ is the estimation error of F.

Differentiating (24) with respect to time and substituting (5)–(7) into it yield

$$\begin{aligned}\dot{V}_2 &= \dot{V}_1 - \frac{1}{\beta}(F-\widehat{F})\dot{\widehat{F}}\\ &= e^T\dot{e} + \frac{1}{2}s^T\dot{M}(x_1)s + s^TM(x_1)\dot{s} - \frac{1}{\beta}(F-\widehat{F})\dot{\widehat{F}}\\ &= e^T\dot{e} + \frac{1}{2}s^T\dot{M}(x_1)s + s^TM(x_1)\left[M(x_1)^{-1}\right.\\ &\quad\cdot\left(u - \tau_d - C(x_1,x_2)x_2\right.\\ &\quad\left.-G(x_1)\right) - M(x_1)^{-1}\left(k_1s + \widehat{F}\mathrm{sign}(s) + k_2|s|^{n_1/n_2}\right.\\ &\quad\cdot\,\mathrm{sign}\left(|s|^{n_1/n_2}\right)\\ &\quad\left.+u - G(x_1) - C(x_1,x_2)x_2 + C(x_1,x_2)s\right)] - \frac{1}{\beta}(F-\widehat{F})\dot{\widehat{F}}\\ &= e^T\dot{e} + \frac{1}{2}s^T\dot{M}(x_1)s - s^T\tau_d - k_1s^Ts - \widehat{F}s^T\mathrm{sign}(s)\\ &\quad - s^TC(x_1,x_2)s\\ &\quad - k_2s^T\cdot\left(|s|^{n_1/n_2}\cdot\mathrm{sign}\left(|s|^{n_1/n_2}\right)\right) - \frac{1}{\beta}(F-\widehat{F})\dot{\widehat{F}}\Big).\end{aligned} \tag{25}$$

Substituting (9) and (17) into (25) and using Property 2 give us

$$\begin{aligned}\dot{V}_2 &= e^T\left(-k_3 e - k_6|e|^{n_1/n_2}\right) + \frac{1}{2}s^T\dot{M}(x_1)s - s^T\tau_d - k_1 s^T s \\ &\quad - \widehat{F}s^T\operatorname{sign}(s) - k_2 s^T\cdot\left(|s|^{n_1/n_2}\cdot\operatorname{sign}\left(|s|^{n_1/n_2}\right)\right) \\ &\quad - s^T C(x_1,x_2)s - \frac{1}{\beta}(F-\widehat{F})\dot{\widehat{F}} \\ &= -k_3 e^T e - k_6 e^T|e|^{n_1/n_2} - s^T\tau_d - k_1 s^T s - \widehat{F}s^T\operatorname{sign}(s) \\ &\quad - k_2 s^T\cdot\left(|s|^{n_1/n_2}\cdot\operatorname{sign}\left(|s|^{n_1/n_2}\right)\right) - Fs^T\operatorname{sign}(s) \\ &\quad + \widehat{F}s^T\operatorname{sign}(s).\end{aligned} \tag{26}$$

Using Property 1 and Property 2, we have

$$\begin{aligned}\dot{V}_2 &\le -k_3\|e\|^2 - k_6 e^T|e|^{n_1/n_2} + \|s\|\,\|\tau_d\| - k_1\|s\|^2 \\ &\quad - k_2 s^T|s|^{n_1/n_2} - F\|s\| \\ &\le -2k_3\left(\frac{1}{2}e^T e\right) - 2^{(n_1+n_2)/2n_2}k_6\left(\frac{1}{2}e^T e\right)^{(n_1+n_2)/2n_2} \\ &\quad - \frac{2k_1}{\lambda_2}\left(\frac{1}{2}s^T M(x_1)s\right) \\ &\quad - k_2\left(\frac{2}{\lambda_2}\right)^{(n_1+n_2)/2n_2}\left(\frac{1}{2}s^T M(x_1)s\right)^{(n_1+n_2)/2n_2} \\ &= -k_7\left(\frac{1}{2}e^T e\right) - k_8\left(\frac{1}{2}s^T M(x_1)s\right) - k_9\left(\frac{1}{2}e^T e\right)^{(n_1+n_2)/2n_2} \\ &\quad - k_{10}\left(\frac{1}{2}s^T M(x_1)s\right)^{(n_1+n_2)/2n_2} \\ &\le -c_1\left(\frac{1}{2}e^T e + \frac{1}{2}s^T M(x_1)s\right) - c_2\left(\frac{1}{2}e^T e + \frac{1}{2}s^T M(x_1)s\right)^{(n_1+n_2)/2n_2} \\ &= -c_1 V_1 - c_2 V_1^{(n_1+n_2)/2n_2}.\end{aligned} \tag{27}$$

Solving (27) leads to $V_1(t) \equiv 0$ for all $t \ge t_c$. Therefore, from (27), we can obtain $\dot{V}_2(t) \le 0$.

From (5) and (8), the disturbance compensation error can be derived as

$$\begin{aligned}e_d &= \tau_{\text{dcom}} - \tau_d \\ &= -k_1 s - \widehat{F}\operatorname{sign}(s) - k_2|s|^{n_1/n_2}\cdot\operatorname{sign}\left(|s|^{n_1/n_2}\right) - u \\ &\quad + C(x_1,x_2)x_2 + G(x_1) + M(x_1)\dot{x}_2.\end{aligned} \tag{28}$$

Substituting (6) and (7) into (28), we can get

$$\begin{aligned}e_d &= -k_1 s - \widehat{F}\operatorname{sign}(s) - k_2|s|^{n_1/n_2}\cdot\operatorname{sign}\left(|s|^{n_1/n_2}\right) \\ &\quad - u + M(x_1)\dot{s} + M(x_1)\dot{\phi} + C(x_1,x_2)x_2 + G(x_1) \\ &= M(x_1)\dot{s} + C(x_1,x_2)s.\end{aligned} \tag{29}$$

From (12), (13), and (15), the fault compensation error can be derived as

$$\begin{aligned}e_f &= \tau_{\text{fcom}} - u_f \\ &= -\widehat{\alpha}(t) - \frac{1}{k_3 k_4}\dot{y} - \left(-\alpha(t) - \frac{1}{k_3}\dot{\alpha}(t)\right) \\ &= -\widehat{\alpha}(t) + \alpha(t) \\ &= -e.\end{aligned} \tag{30}$$

Solving (27) leads to $V_2(t) \equiv 0$ for all $t \ge t_c$. Then, according to (18), we have $e(t) = 0$ and $s(t) = 0$ for all $t \ge t_c$. Thus, we can further have $\dot{s}(t) = 0$. Therefore, from (29) and (30), we can get e_d=0 and $e_f = 0$ for all $t \ge t_c$. This indicates that e_d and e_f can converge to zero in finite time t_c, i.e., $\lim_{t\longrightarrow t_c} e_d = \lim_{t\longrightarrow t_c}(\tau_{\text{dcom}} - \tau_d) = 0$ and $\lim_{t\longrightarrow t_c} e_f = \lim_{t\longrightarrow t_c}(\tau_{\text{fcom}} - u_f(T)) = 0$. This concludes the proof of Theorem 1.

Remark 1. It can be seen from the nonlinear observer (14) and the proof of Theorem 1 that the nominal controller u_{nom} can be cancelled in the composite compensation controller (16). This indicates that the proposed composite compensation control scheme does not depend on the specific nominal control law.

Remark 2. In the literature [18–20], disturbance and fault are treated as centralized uncertainty. Different from them, the designed composite compensation controller (16) consists the terms regarding disturbance as well as actuator faults. Thus, the respective characteristic of disturbance and actuator faults can better be reflected.

4. Simulations

Simulations are conducted on a 2-DOF robot manipulator, as shown in Figure 3. The dynamics of the robot is

$$\begin{aligned}M(q) &= \begin{bmatrix} p_1 + p_2 + 2p_3\cos q_2 & p_2 + p_3\cos q_2 \\ p_2 + p_3\cos q_2 & p_2 \end{bmatrix}, \\ C(q,\dot{q}) &= \begin{bmatrix} -p_3\dot{q}_2\sin q_2 & -p_3(\dot{q}_1+\dot{q}_2)\sin q_2 \\ p_3\dot{q}_1\sin q_2 & 0 \end{bmatrix}, \\ G(q) &= \begin{bmatrix} p_4 g\cos q_1 + p_5 g\cos(q_1+q_2) \\ p_5 g\cos(q_1+q_2) \end{bmatrix},\end{aligned} \tag{31}$$

where $q = [\,q_1 \;\; q_2\,]^T$ and q_1 and q_2 represent the position of the first joint and the second joint, respectively. Besides,

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} m_1 h_1^2 + m_2 l_1^2 + J_1 \\ m_2 h_2^2 + J_2 \\ m_2 l_1 h_2 \\ m_1 h_1 + m_2 l_1 \\ m_2 h_2 \end{bmatrix}, \tag{32}$$

where m_1 and m_2 are the mass of the link, l_1 and l_2 are the length of the link, h_1 and h_2 are the distance to the center of the mass, J_1 and J_2 are the moment of inertia, and g is the

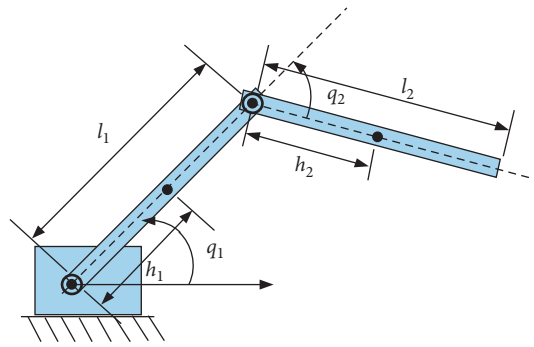


Figure 3: 2-DOF robot manipulator.

gravity coefficient. In the simulations, $m_1 = m_2 = 1$, $l_1 = l_2 = 1$, $h_1 = h_2 = 0.5$, and $J_1 = J_2 = 0.08$.

The conventional PD controller [24] which is widely applied in practice is taken as the nominal controller u_{nom}:

$$u_{\text{nom}} = M(q)\left[\ddot{q}_d + k_d\dot{e}_p + k_p e_p\right] + C(q,\dot{q}) + G(q), \quad (33)$$

where $e_p = q_d - q$ represents the position tracking error of the robot.

The initial value of the robot is $q(0) = \left[\begin{array}{cc} 0.05 & 0.1 \end{array}\right]^T$. The desired position of the robot is $q_d = \left[\begin{array}{cc} q_{d1} & q_{d2} \end{array}\right]^T$, where

$$\begin{cases} q_{d1} = 0.05\sin(4t - 0.5\pi), \\ q_{d2} = 0.06\sin(4t - 0.5\pi). \end{cases} \quad (34)$$

The external disturbance in the robotic system is

$$\tau_d = \left[\begin{array}{cc} 0.2\cos(2t) & 0.4\cos(2t) \end{array}\right]^T. \quad (35)$$

For joint 1 of the robot, during 2 sec-3 sec and 8 sec-9 sec, the actuator fault is constant deviation fault. During 4 sec–7 sec, the actuator fault is time-varying fault. For the rest of the time, the actuator is no fault.

For joint 2 of the robot, during 0 sec–5 sec, the actuator is no fault. After 5 sec, the actuator fault is loss of effectiveness fault; i.e., the actuator losses 20% of the effectiveness.

Specifically, the actuator fault for joint 1 and joint 2 is as follows:

$$u_{f_1}(T_1) = \begin{cases} -0.8, & 2 \le t \le 3, \\ -0.6\sin(4t), & 4 \le t \le 7, \\ -0.5, & 8 \le t \le 9, \\ 0, & \text{elsewhere} \end{cases}, \quad (36)$$

$$u_{f_2}(T_2) = \begin{cases} -0.2u_{\text{nom}}, & t \ge 5, \\ 0, & \text{elsewhere.} \end{cases}$$

In the simulations, the performances of the conventional fault reconstruction scheme [13] and the proposed composite compensation control scheme are compared. The parameters of the conventional fault reconstruction scheme [13] are chosen as $k_d = 70$, $k_p = 50$, $k_1 = 0.001$, $k_2 = 25$, $k_3 = 75$, $k_4 = 145$, $n_1 = 101$, and $n_2 = 103$. The parameters of the proposed composite compensation controller are chosen as $k_p = 800$, $k_d = 500$, $k_1 = 1165$, $k_2 = 680$, $k_3 = 800$, $k_4 = 50$, $k_5 = k_3/k_4 = 16$, $k_6 = 0.5$, $n_1 = 87$, $n_2 = 103$, and $\beta = 0.5$. The simulation results are shown in Figures 4–9.

The effect of the external disturbance compensation with the composite compensation controller is shown in Figures 4 and 5. It can be seen that the proposed composite compensation controller can successfully compensate the disturbance, and the disturbance compensation error can quickly converge within a short time. Since the conventional fault reconstruction scheme cannot compensate the external disturbance, the effect of the external disturbance compensation with the conventional fault reconstruction scheme is not shown.

Figures 6(a) and 6(b) show that both the conventional fault reconstruction scheme and the proposed composite compensation controller can compensate various types of actuator faults. However, it can be seen from Figures 7(a) and 7(b) that when the proposed controller is employed, the fault compensation accuracy is higher and the fault error convergence rate is faster.

Figure 8(a) shows that, with the conventional fault reconstruction scheme, the real position trajectory of the robot cannot track the desired position trajectory well. Comparatively, Figure 8(b) shows that, with the proposed composite compensation controller, the robot can track the desired position in a satisfactory way within a short time. As shown in Figure 9(a), when the fault reconstruction scheme is used, there exists obvious position tracking error, and the error convergence rate is slow. Comparatively, when the proposed controller is employed, the position tracking error is ideal, and the error convergence rate is faster in Figure 9(b). The reason is that the proposed composite compensation controller can not only deal with actuator faults but also external disturbance in the system.

To further demonstrate the superiority of the proposed composite compensation control scheme, several performance indicators are compared in quantitative in Tables 1–3. The indicator $t_{d_i s}$ denotes the adjustment time of disturbance compensation, and $|e_{d_i}|$ represents the disturbance compensation error. $t_{f_i s}$ denotes the adjustment time of fault compensation, and $|e_{f_i}|$ represents the fault compensation error. $t_{p_i s}$ denotes the adjustment time of position tracking,

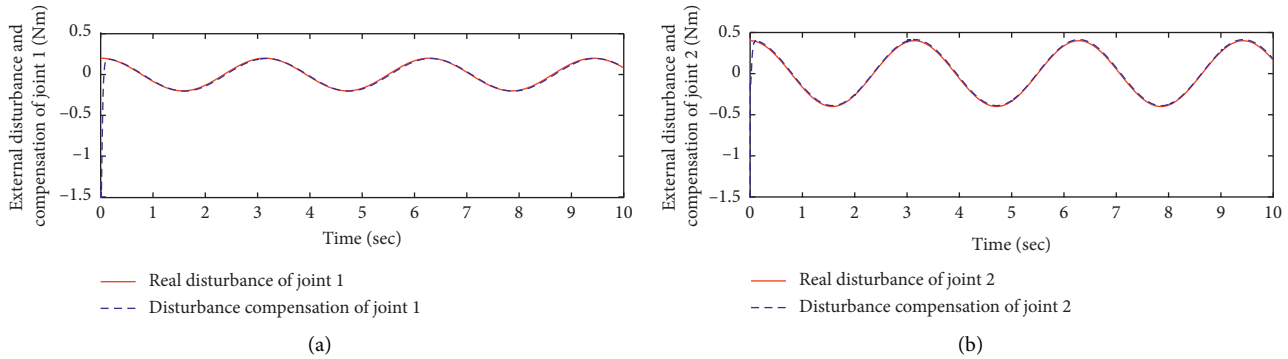


Figure 4: External disturbance and compensation of joint 1 and joint 2 (composite compensation controller).

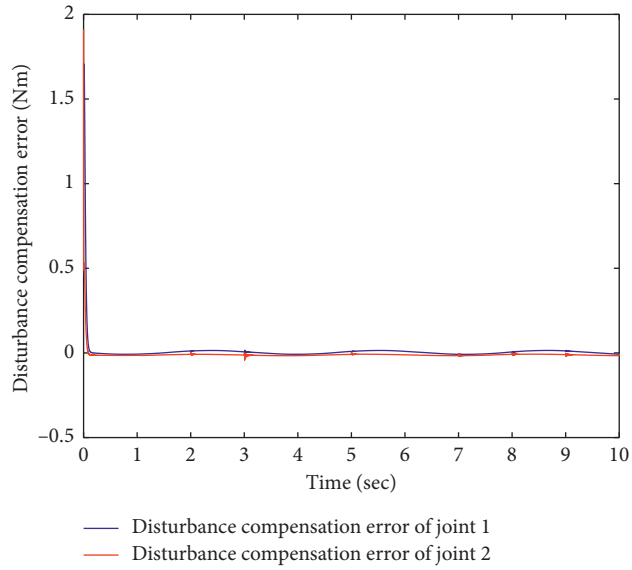


Figure 5: Disturbance compensation error of joint 1 and joint 2 (composite compensation controller).

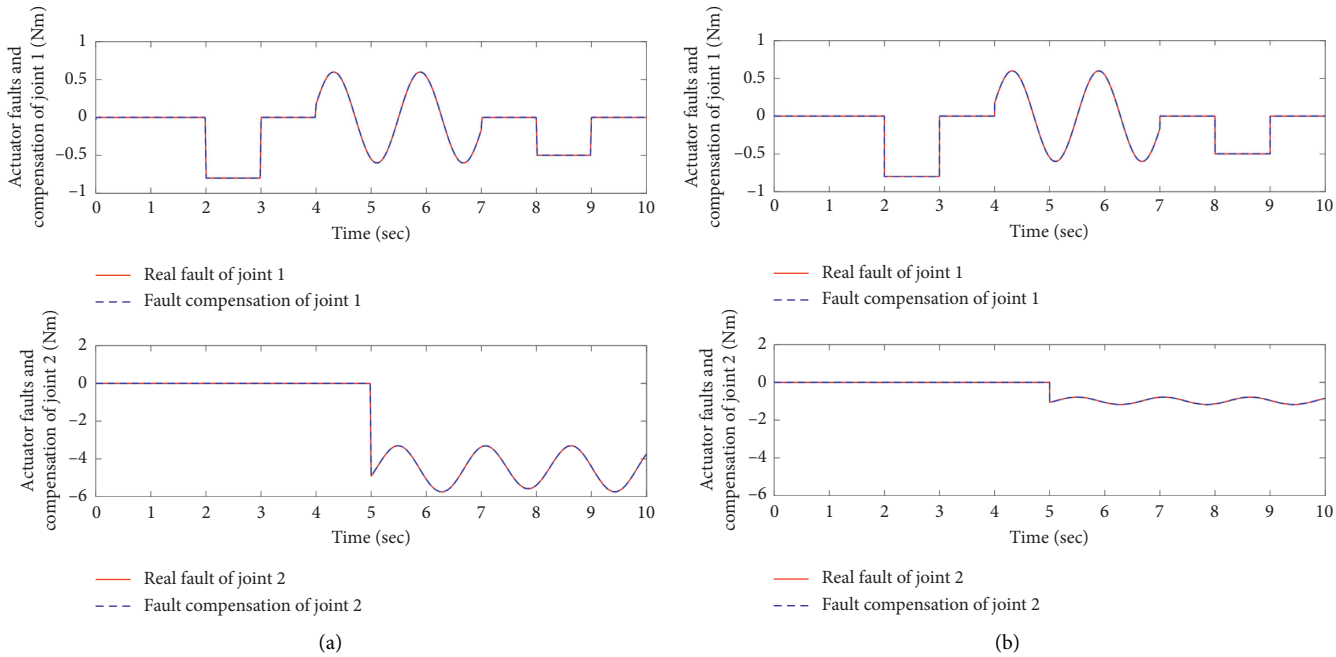


Figure 6: Actuator faults and compensation of joint 1 and joint 2: (a) fault reconstruction scheme [13]; (b) composite compensation controller.

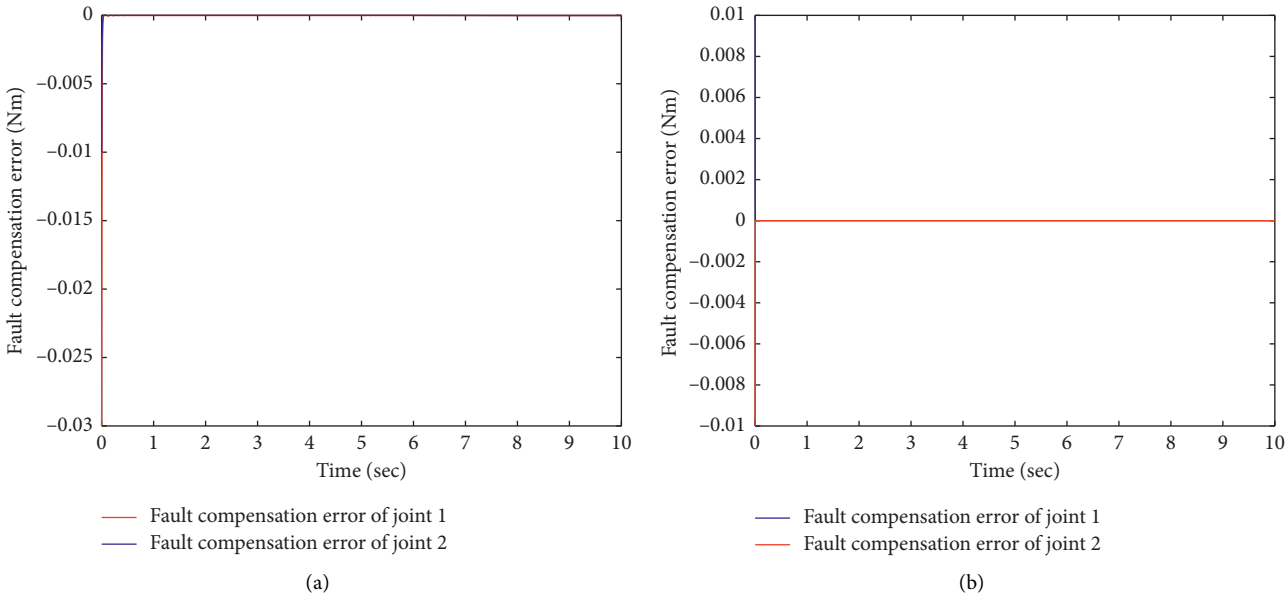


Figure 7: Fault compensation error of joint 1 and joint 2: (a) fault reconstruction scheme [13]; (b) composite compensation controller.

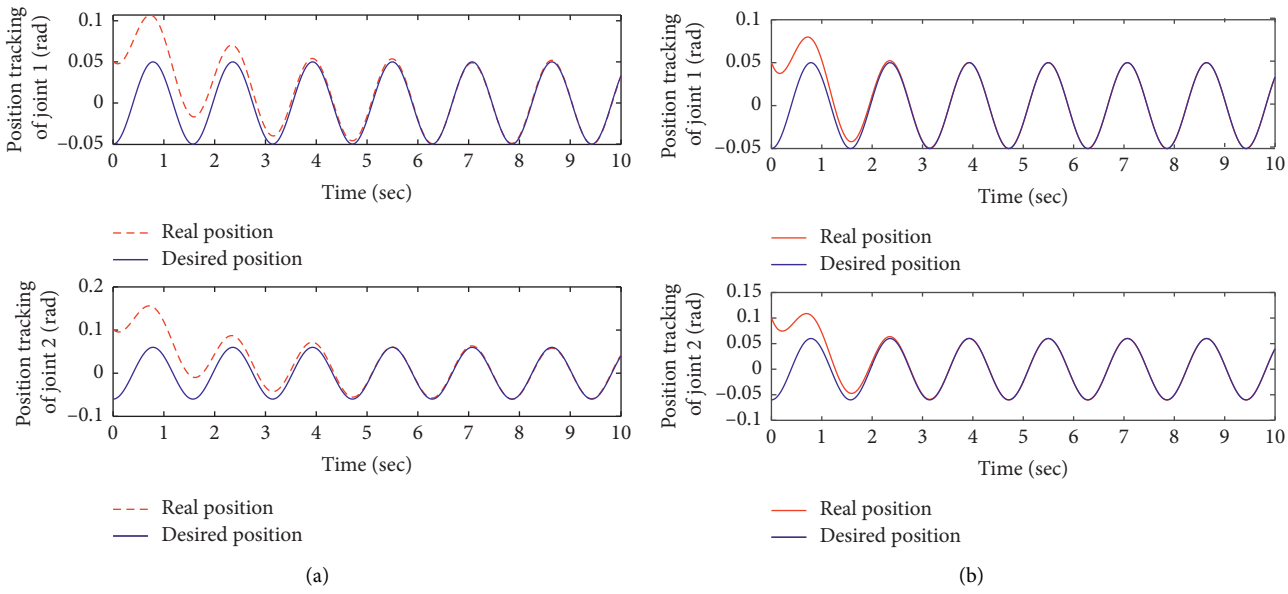


Figure 8: Position tracking of joint 1 and joint 2: (a) fault reconstruction scheme [13]; (b) composite compensation controller.

and $|e_{p_i}|_{\max}$ represents the position tracking error, where $i = 1, 2$ represent joint 1 and joint 2 of the robot, respectively.

Table 1 indicates that, with the proposed composite compensation controller, the disturbance compensation error of joint 1 and joint 2 can rapidly converge in 0.1191 sec and 0.0833 sec, respectively. Nevertheless, the conventional fault reconstruction scheme cannot compensate disturbance.

Table 2 shows that, with the proposed composite compensation controller, the adjustment time of fault compensation is shorter and the absolute value of the fault compensation error is smaller. In other words, when the proposed controller is employed, the fault compensation accuracy is higher and the fault error convergence rate is faster.

Table 3 shows that, with the proposed composite compensation controller, the adjustment time of position tracking for both joint 1 and joint 2 is shorter, and the absolute value of the position tracking error is smaller. In other words, when the proposed controller is employed, the robot can track the desired position trajectory more accurately and quickly.

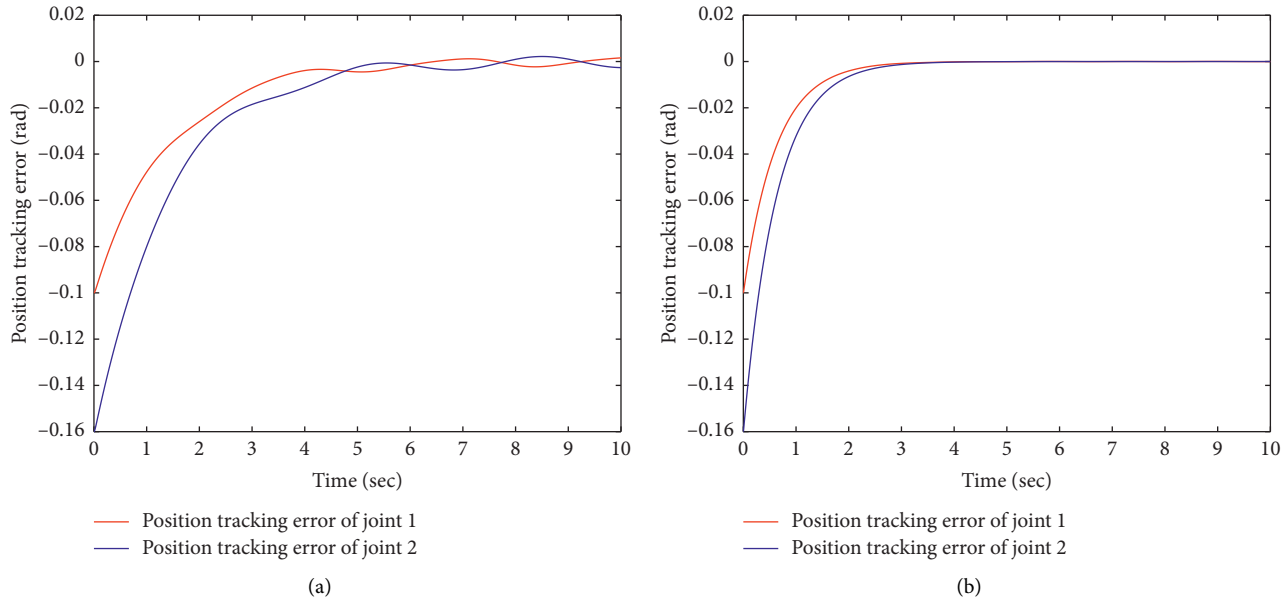


Figure 9: Position tracking error of joint 1 and joint 2: (a) fault reconstruction scheme [13]; (b) composite compensation controller.

Table 1: Quantitative comparison of external disturbance compensation.

	Composite compensation controller		Fault reconstruction scheme	
	$t_{d_i s}(\pm 2\%)$ (sec)	$\lvert e_{d_i}\rvert_{\max}$ (Nm)	$t_{d_i s}(\pm 2\%)$ (sec)	$\lvert e_{d_i}\rvert_{\max}(Nm)$
Joint 1	0.1191	1.147×10^{-2}	None	None
Joint 2	0.0833	1.53×10^{-2}	None	None

Table 2: Quantitative comparison of actuator fault compensation.

	Composite compensation controller		Fault reconstruction scheme	
	$t_{f_i s}(\pm 5\%)$ (sec)	$\lvert e_{f_i}\rvert_{\max}$ (Nm)	$t_{f_i s}(\pm 5\%)$ (sec)	$\lvert e_{f_i}\rvert_{\max}(Nm)$
Joint 1	0.006	1.11×10^{-16}	0.044	1.89×10^{-5}
Joint 2	0.006	8.88×10^{-15}	0.044	1.89×10^{-5}

Table 3: Quantitative comparison of position tracking.

	Composite compensation controller		Fault reconstruction scheme	
	$t_{p_i s}(\pm 3\%)$ (sec)	$\lvert e_{p_i}\rvert_{\max}$ (rad)	$t_{p_i s}(\pm 3\%)$ (sec)	$\lvert e_{p_i}\rvert_{\max}$ (rad)
Joint 1	2.435	3.32×10^{-5}	5.777	2.49×10^{-3}
Joint 2	2.734	3.65×10^{-6}	4.834	3.57×10^{-3}

5. Conclusions

For a robotic system subject to simultaneous external disturbance and various actuator faults, a composite compensation control scheme based on adaptive sliding mode controller and observer-based fault-tolerant controller is proposed. Compared to the conventional fault reconstruction scheme, the proposed scheme can compensate not only external disturbance but also various actuator faults. The fault compensation accuracy is higher, and the fault error convergence rate is faster. Moreover, the robot can track the desired position trajectory more accurately and quickly. Experimental verification of the proposed control in this paper is quite necessary and remains as our work in the next step. Besides, the extension of the proposed control to online estimates the fault information for a nonlinear robotic system using a fault diagnosis approach remains as our future research.

Data Availability

The data that support our manuscript conclusions are some open access articles that have been properly cited, and the

readers can easily obtain these articles to verify the conclusions.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant nos. 61973257, 61973331, and 61875166) and Sichuan Youth Science and Technology Foundation (Grant 2017JQ0022).

References

[1] Y. Guo, "Globally robust stability analysis for stochastic cohen-grossberg neural networks with impulse control and time-varying delays," *Ukrainian Mathematical Journal*, vol. 69, no. 8, pp. 1220–1233, 2017.

[2] W. Liu, J. Cui, and J. Xin, "A block-centered finite difference method for an unsteady asymptotic coupled model in fractured media aquifer system," *Journal of Computational and Applied Mathematics*, vol. 337, pp. 319–340, 2018.

[3] L. Gao, D. Wang, and G. Wang, "Further results on exponential stability for impulsive switched nonlinear time-delay systems with delayed impulse effects," *Applied Mathematics and Computation*, vol. 268, pp. 186–200, 2015.

[4] V. Utkin, *Sliding Modes on Control and Optimization*, Springer-Verlag, Berlin, Germany, 1992.

[5] W.-H. Chen, "Disturbance observer based control for nonlinear systems," *IEEE/ASME Transactions on Mechatronics*, vol. 9, no. 4, pp. 706–710, 2004.

[6] Z. Pu, R. Yuan, J. Yi, and X. Tan, "A class of adaptive extended state observers for nonlinear disturbed systems," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 9, pp. 5858–5869, 2015.

[7] H. Wang, Y. Pan, S. Li, and H. Yu, "Robust sliding mode control for robots driven by compliant actuators," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 3, pp. 1259–1266, 2019.

[8] R. Cui, L. Chen, C. Yang, and M. Chen, "Extended state observer-based integral sliding mode control for an underwater robot with unknown disturbances and uncertain nonlinearities," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6785–6795, 2017.

[9] W. Yuan and G. Gao, "Sliding mode control of the automobile electro-coating conveying mechanism with a nonlinear disturbance observer," *Advances in Mechanical Engineering*, vol. 10, no. 9, pp. 1–9, 2018.

[10] M. Li and Y. Chen, "Robust adaptive sliding mode control for switched networked control systems with disturbance and faults," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 193–204, 2019.

[11] Q. Meng, T. Zhang, X. Gao, and J.-y. Song, "Adaptive sliding mode fault-tolerant control of the uncertain stewart platform based on offline multibody dynamics," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 3, pp. 882–894, 2014.

[12] M. S. Mahmoud, A. M. Memon, and P. Shi, "Observer-based fault-tolerant control for a class of nonlinear networked control systems," *International Journal of Control*, vol. 87, no. 8, pp. 1707–1715, 2014.

[13] B. Xiao and S. Yin, "An intelligent actuator fault reconstruction scheme for robotic manipulators," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 639–647, 2018.

[14] J. Zhao, S. Jiang, F. Xie, Z. He, and J. Fu, "A novel nonlinear fault tolerant control for manipulator under actuator fault," *Mathematical Problems in Engineering*, vol. 2018, Article ID 5198615, 11 pages, 2018.

[15] B. Xiao, S. Yin, and H. Gao, "Reconfigurable tolerant control of uncertain mechanical systems with actuator faults: a sliding mode observer-based approach," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 4, pp. 1249–1258, 2018.

[16] Y. Sun, Z. Zhang, M. Leibold et al., "Protective control for robot manipulator by sliding mode based disturbance reconstruction approach," in *Proceedings of the 2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 1015–1022, Munich, Germany, 2017.

[17] W. Liu and P. Li, "Disturbance observer-based fault-tolerant adaptive control for nonlinearly parameterized systems," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8681–8691, 2019.

[18] C.-C. Chen, S. S.-D. Xu, and Y.-W. Liang, "Study of nonlinear integral sliding mode fault-tolerant control," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 1160–1168, 2016.

[19] Y.-W. Liang, C.-C. Chen, D.-C. Liaw, and Y.-T. Wei, "Nonlinear reliable control with application to a vehicle antilock brake system," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2114–2123, 2013.

[20] S. M. Smaeilzadeh and M. Golestani, "Finite-time fault-tolerant adaptive robust control for a class of uncertain nonlinear systems with saturation constraints using integral backstepping approach," *IET Control Theory & Applications*, vol. 12, no. 15, pp. 2109–2117, 2018.

[21] A. Mohammadi, M. Tavakoli, H. J. Marquez, and F. Hashemzadeh, "Nonlinear disturbance observer design for robotic manipulators," *Control Engineering Practice*, vol. 21, no. 3, pp. 253–267, 2013.

[22] W. Liang, S. Huang, S. Chen, and K. K. Tan, "Force estimation and failure detection based on disturbance observer for an ear surgical device," *ISA Transactions*, vol. 66, pp. 476–484, 2017.

[23] B. Xiao and S. Yin, "Velocity-free fault-tolerant and uncertainty attenuation control for a class of nonlinear systems," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 7, pp. 4400–4411, 2016.

[24] M. W. Spong, S. Hutchinson, and M. Vidyasagar, *Robot Modeling and Control*, Wiley, New York, NY, USA, 2006.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8382734, 10 pages
https://doi.org/10.1155/2020/8382734

Research Article

Adaptive Cruise Control Strategy Design with Optimized Active Braking Control Algorithm

Wenguang Wu,[1,2] **Debiao Zou**,[3] **Jian Ou**,[1] **and Lin Hu**[1,2]

[1]*School of Automotive & Mechanical Engineering, Changsha University of Science & Technology, Changsha 410114, China*
[2]*Hunan Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems, Changsha 410114, China*
[3]*Mechanical and Vehicle Engineering, Hunan University, Changsha 410006, China*

Correspondence should be addressed to Lin Hu; hulin@csust.edu.cn

Received 8 June 2020; Accepted 26 June 2020; Published 21 July 2020

Guest Editor: Yong Chen

The braking quality is considered as the most important performance of the adaptive control system that influences the vehicle safety and ride comfort remarkably. This research is aimed at designing an adaptive cruise control (ACC) system based on active braking algorithm using hierarchical control. Taking into account the vehicle with safety and comfort, the upper decision-making controller is designed based on model predictive control algorithm. Throttle controller and braking controller are designed with feedforward and feedback algorithms as the bottom controller, where the braking controller is designed based on the hydraulic braking model. The whole model is simulated collaboratively with Amesim, Carsim, and Simulink. By comparison with the full deceleration model, the results show that the proposed algorithm can not only make the vehicle maintain a safe distance under the premise of following the target vehicle ahead effectively but also provide favorable driving comfort.

1. Introduction

In recent years, one of the most important goals in the automotive industry has been to offer passengers the highest level of safety, comfort, and efficiency by partially or completely removing driving duties from humans. Advanced Driver Assistant System (ADAS) has become a research hotspot in the field of intelligent transportation; it not only improves the road capacity [1], but also ensures the safety of drivers and vulnerable road users to some extent [2, 3]. Studies have shown that the active safety systems, such as adaptive cruise control, electronic stability control, or lane keeping assistant, which are already on the automotive market, can improve safety by decreasing the number of traffic accidents, among which the ACC helps a lot to reduce the driver's work intensity; an ACC equipped vehicle uses radar or other sensors that detect the distance and speed to other preceding vehicles (downstream vehicles) on the highway. In the absence of preceding vehicles, the ACC vehicle travels at a driver-set speed. If a preceding vehicle is detected on the highway by the vehicle's radar, the ACC system determines to control the throttle and braking system so as to maintain an expected distance and acceleration from the preceding vehicle [4].

The planning and decision-making modules are the "brain" of the vehicle and have a high degree of intelligence. All response actions of the vehicle are performed according to the instructions issued by the module. By processing and calculating the real-time state information and environmental information of the vehicle, this module can plan the most reasonable vehicle movement state and send it to the execution control module [5]. The most critical parts for the ACC, the planning and decision-making module, need to decide the optimal control target according to the relative motion state between the host vehicle and the target vehicle: expected longitudinal acceleration or distance [6]. So far, the decision algorithms of the ACC mainly have the following forms: PID feedback control, model predictive control, fuzzy logic control, and optimal control [7–10].

Longitudinal control is the basic function of ACC system where the control technology is used to achieve constant

speed driving of the vehicle, maintaining the distance between vehicles or the time between vehicles to follow the leading vehicle, identifying and tracking the curve of the vehicle ahead, automatic braking, and other functions. The quality of the longitudinal control effect has a direct impact on the safety and comfort of ACC system. The executive control module mainly achieves rapid response to the instructions issued by the planning and decision module and precise tracking of the expected goal through the precise control of the driving system and the braking system. ACC system in accordance with the working conditions can be divided into cruise mode, following mode, and overtaking mode [11]; the research scope of this article is car-following model, whose function is to keep an appropriate distance and speed with the leading vehicle. In order to further improve the effect of vehicle longitudinal control, dynamic model has become one of the key links in the field of vehicle longitudinal control. Among them, Zhan established a longitudinal dynamic model and braking system model for ACC system [12]. The researchers adopted longitudinal control method based on vehicle longitudinal inverse model and used vehicle inverse model to control electronic throttle and braking pressure [13, 14].

With the development of ACC system, more and more working conditions involve speeds of 30 km/h and below, so the vehicle longitudinal control has experienced the development process from single throttle control to combined throttle-braking control [15]. Due to strong robustness, low accuracy requirements for controlled objects, and no need for accurate modeling, classical control methods represented by PID control and numerical look-up tables are widely used. In addition, many researchers use the modified form of PID controller to study longitudinal control of vehicles, and try to improve longitudinal control effect by improving PID controller [16, 17].

Adaptive Neural Network scheme has been used in a platoon, in order to solve the traffic stability problem [18]. PID algorithm is used to directly control the accelerator pedal and the brake pedal to control the acceleration and deceleration of the vehicle to maintain the distance from the preceding vehicle [19, 20]. The fuzzy logic-based ACC controller is used to make one vehicle follow another vehicle stably, having no shock during the process of the accelerator and brake switching [21, 22]. The fuzzy ACC system with speed sign detection capability and synovial control is used for adaptive control system [23, 24]. The change of signal light is also considered to control the driving of vehicles at intersections [25]. The prospective velocity of the preceding vehicle is estimated by a prediction model based on the measured intervehicle distance and the I2V communication to enable an anticipatory driving behavior for the controlled vehicle [26].

One can conclude from the research that the previous active braking functions of adaptive cruise-following system also did not fully consider the ride comfort and hydraulic hysteresis problem. This research is aimed at designing an ACC considering the vehicle ride and proposing an analysis model based on active braking algorithm using hierarchical control.

In this paper, considering the safety, comfort, and the physical characteristics of hydraulic braking system, by switching on and off the valve and motor start-stop, adjusting the hydraulic cylinder pressure, a new ACC control strategy based on active braking is proposed. By comparison with the full deceleration model, the proposed method can improve the braking ride comfort obviously. The remainder of this paper is structured as follows: Section 2, modeling; Section 3, control algorithm research; Section 4, simulation and discussion; Section 5, conclusions.

2. Modeling

This paper is aimed at designing a control scheme that could guarantee safety considering the vehicle characteristic and ensure braking comfort at all times. As shown in Figure 1, the vehicle longitudinal dynamics model, the hydraulic braking system, and the control mechanism are included in the proposed research model. The main idea of the controller model is as follows:

(1) The real-time safety distance according to the vehicle speed and the actual distance and relative speed between leader vehicle and follower vehicle are obtained as the controller input.

(2) The limitation of the acceleration and relative distance of the follower vehicle is calculated by the longitudinal dynamics model.

(3) The expected acceleration of the follower vehicle is calculated and the optimized brake pressure is transmitted to the executive agency including active brake controller and active throttle controller.

(4) The braking pressure is produced by the hydraulic braking system, and the vehicle speed slows down. In this process, the brake pressure information is also transmitted to the longitudinal dynamics model.

2.1. Vehicle Dynamics Model. In this paper, Carsim software is used to build vehicle dynamics model for collaborative simulation. The vehicles are four-wheel drive B-class hatchback with the engine power of 125 kW, and with the hydraulic ABS braking system. The vehicle model includes 7 subsystems: the body, aerodynamics input, transmission system, braking system, steering system, suspension system, and the tire. The parameters of the vehicles are shown in Table 1. The output of the model includes the longitudinal velocity v, acceleration a, engine speed ω_e, and position S.

2.2. Vehicle Reverse Longitudinal Dynamics Model. In the ACC system, the control command from the host controller is a desired vehicle acceleration that needs to be shifted to the desired throttle opening and brake pressure by the vehicle reverse longitudinal dynamic model, which then transmitted to the vehicle longitudinal dynamics model to control the vehicle acceleration, deceleration, or uniform motion in order to achieve the function of the car adaptive cruise system [27, 28].

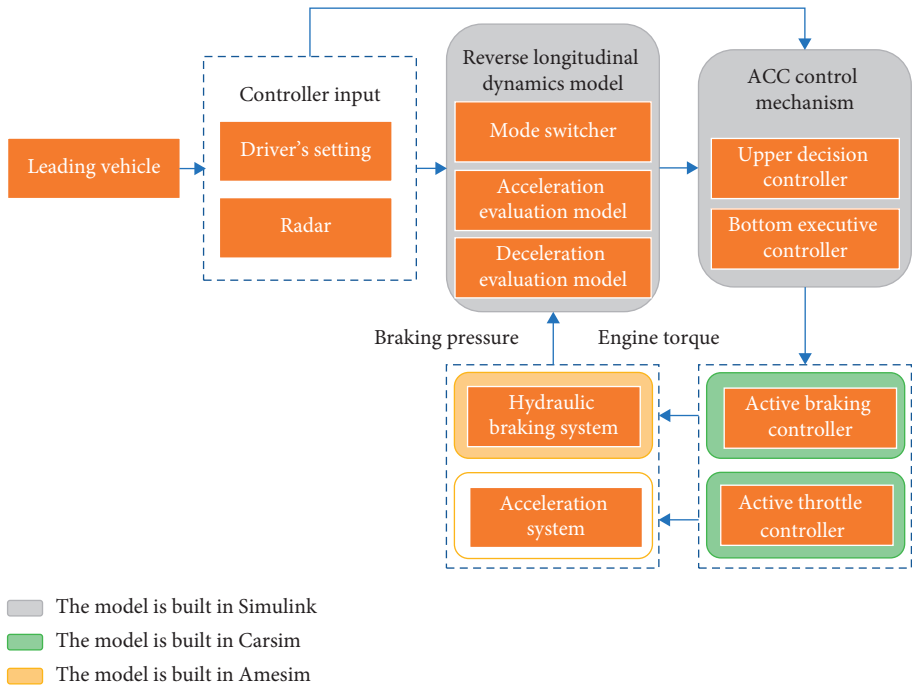


Figure 1: Scheme of the proposed ACC control model.

Table 1: Parameters of the vehicles.

Parameters	Symbol	Value
Sprung mass (kg)	M	1111
Distance between CM and front axle (m)	a	1.04
Distance between CM and rear axle (m)	b	1.56
Air density (kg/m^3)	ρ	1.206
Rolling resistance coefficient	f	0.02
Track (m)	d	1.695
Centroid height (m)	H	0.54
Gear ratio of main gear	i_0	4.1
Transmission gear	N	6
Gear ratio of transmission	i_g	1
Tire rolling radius (m)	r	0.311
Air resistance coefficient	C_D	0.342
Frontal area (m^2)	A	1.6
The efficiency of the drive system	η	0.9

2.2.1. Mode Switch. To the vehicle dynamics system, acceleration and braking are separate movements. When braking, the car should first release the accelerator pedal, using engine drag, wind resistance, and rolling resistance and other ways to brake. If the above action still cannot meet the needs of vehicle deceleration, then depress the brake pedal, applying brake force to increase vehicle deceleration. Besides, taking into account the driving comfort and the reliability of the corresponding parts of the vehicle, the designing process should avoid frequent switching between acceleration control and braking control.

It is easy to directly measure the maximum deceleration value a_{max} at different speeds in Carsim software, as shown in Figure 2. In order to improve the driving comfort of the vehicle, the width of the transition area is set on the upper and lower sides of the switching curve, which is generally taken from experience.

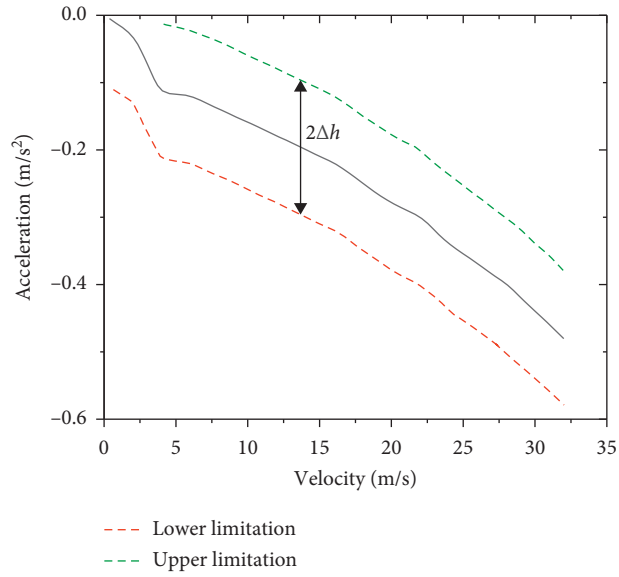


Figure 2: Acceleration control/braking control switching curve.

The expected acceleration of the vehicle is defined as a_{fdes}. According to the switching curve, when $a_{fdes} \geq a_{max}$, the car switches to acceleration control. On the contrary, when $a_{fdes} \leq a_{max}$, the car switches to braking control.

2.2.2. Acceleration Control. If the vehicle switches to acceleration control mode, it is necessary to do as the expected acceleration requires. The expected torque is calculated from

the expected acceleration, and then the desired throttle opening can be checked through the engine mapping.

Without considering the conversion quality of rotating parts, the longitudinal dynamic analysis of the vehicle is analyzed and the vehicle longitudinal dynamics model is as follows:

$$ma_{\text{fdes}} = F_{\text{t}} - F_{\text{xb}} - \sum F(v),$$
$$\sum F(v) = \frac{1}{2}C_{\text{D}}A\rho v^2 + mgf, \tag{1}$$

where a_{fdes} is the expected acceleration, m is the vehicle mass, F_{t} is the driving force, F_{xb} is the braking force, $\sum F(v)$ is the sum of the resistances, C_{D} is the air resistance coefficient, A is the frontal area, ρ is the air density, v is the car speed, g is the gravitational acceleration, and f is the rolling resistance coefficient.

Regardless of the elastic deformation of the transmission system, the driving force can be calculated as follows:

$$F_{\text{t}} = \frac{\eta\tau(\omega_{\text{t}}/\omega_{\text{e}})i_{\text{g}}i_0}{r}T_{\text{e}} = K_{\text{d}}T_{\text{e}}, \tag{2}$$

where η is the mechanical efficiency, T_{e} is the engine torque, ω_{t} is the torque converter turbine speed, ω_{e} is the engine speed, i_{g} is the transmission gear ratio, i_0 is the main gear ratio, $\tau(\omega_{\text{t}}/\omega_{\text{e}})$ is a torque converter characteristic function, r is the wheel rolling radius, and K_{d} is a variable that can be observed in real time:

$$K_{\text{d}} = \frac{\eta\tau(\omega_{\text{t}}/\omega_{\text{e}})R_{\text{g}}R_{\text{m}}}{r} = \frac{\eta\tau\left(\left(vR_{\text{g}}R_{\text{m}}\right)/\left(r\omega_{\text{e}}\right)\right)R_{\text{g}}R_{\text{m}}}{r}. \tag{3}$$

When the vehicle is accelerating, $F_{\text{xb}} = 0$. And the expected engine output torque can be obtained according to the transmission gear ratio and speed ratio:

$$T_{\text{des}} = \frac{ma + \sum F(v)}{K_{\text{d}}}. \tag{4}$$

It is easy to get the throttle opening of the engine from the mapping by taking the throttle opening required to output different torques at different speeds. The values are expressed as follows:

$$\alpha_{\text{des}} = f\left(T_{\text{des}}, \omega_{\text{e}}\right). \tag{5}$$

2.2.3. Braking Control. If the car switches to braking control mode, it is necessary to do as the expected deceleration requires. The desired braking force can be calculated according to the desired acceleration, and the braking pressure can be obtained through the braking reverse model [29].

In this case, the engine output torque is terminated, $T_{\text{e}} = 0$; according to equation (2), it can be seen that $F_{\text{t}} = 0$; the vehicle longitudinal force can be shown as

$$ma_{\text{fdes}} = -F_{\text{xb}} - \sum F(v). \tag{6}$$

The braking force and braking pressure can be approximated as a linear relationship as follows:

$$F_{\text{bdes}} = K_{\text{b}}P_{\text{des}}, \tag{7}$$

where K_{b} is a constant.

It is not hard to calculate the braking pressure from equations (6) and (7):

$$P_{\text{des}} = \frac{\left|-ma_{\text{fdes}} - 0.5C_{\text{D}}A\rho v^2 - mgf\right|}{K_{\text{b}}}. \tag{8}$$

2.3. Active Braking Hydraulic System Model. The expected acceleration got from upper-level decision controller is transformed by the inverse vertical dynamic model into the desired braking pressure or desired throttle opening to the underlying accelerator and brake actuator. Active braking objective is archived by controlling the plunger pump and valves to start or stop to achieve the object hydraulic oil pressure, thereby controlling the brake calipers.

2.3.1. Designing of the Active Braking Principle. The simplified hydraulic structure of active braking system is shown in Figure 3. The working principle is as follows. If the system switches into the active braking mode, there are three active modes: booster, packing, and decompression. When pressure increases, high-pressure directional valve 6 and directional valve 5 are opened and the pump motor is started. Brake fluid flows through high-pressure valve 6 and motor pump and then through the inlet valve 12 into the wheel cylinder, then pushing the piston of wheel cylinder to slow down the wheel rotate speed. When braking force reaches a certain intensity, active braking system switches into the pressure hold-on mode, directional valve 5 is opened, pump motor and high-pressure valve 6 are closed, and wheel cylinder pressure keeps constant at this state. When pressure decreases, the high-pressure valve 6 is opened, directional valve 5 and the motor are closed, and the braking fluid flows into the low-pressure accumulator 9, increasing the braking fluid storage of the accumulator. In the process of the new pressure increase case, plunger pump 8 works, and the braking fluid flows out of the low-pressure accumulator 9 and then through inlet valve 12 to the wheel cylinder.

2.3.2. Modeling of the Hydraulic Braking System

(1) Accumulator model

The pressure and volume of the accumulator follow the idea gas law. The mathematical model is as follows:

$$P_A V_A^n = P_1 V_1^n = P_2 V_2^n, \tag{9}$$

where P_A and V_A are the inflation pressure and accumulator capacity, respectively, P_1 and P_2 are the highest and the lowest pressure values of the accumulator, and V_1 and V_2 are the highest and the lowest volume values of the accumulator. Considering that the braking process could be seen as

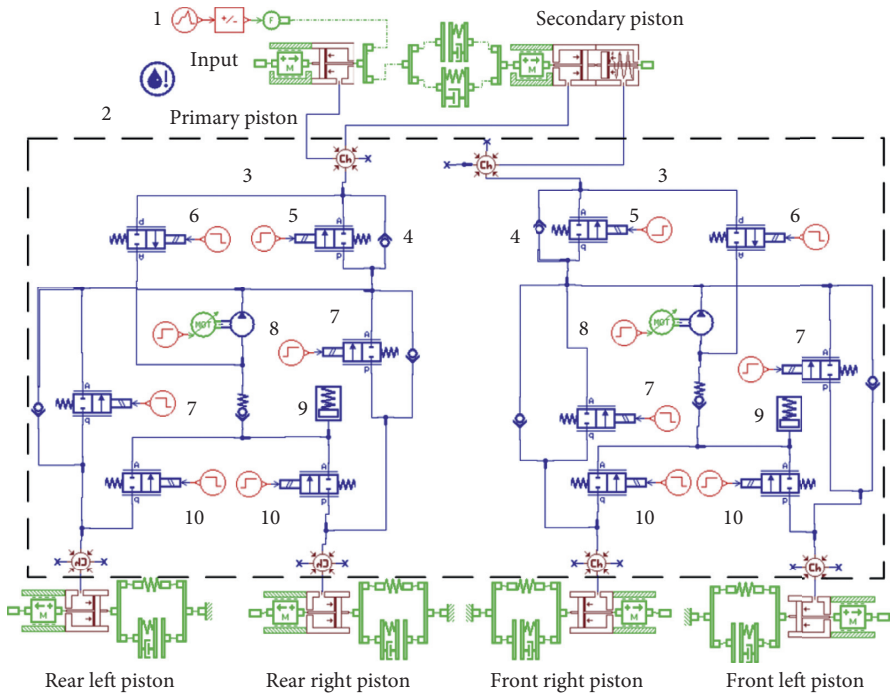


FIGURE 3: Schematic of the active hydraulic braking system. (1) Master cylinder; (2) hydraulic unit; (3) hydraulic circuit; (4) check valve; (5) directional valve; (6) high pressure directional valve; (7) pressure-increasing valve; (8) oil returning pump; (9) low pressure accumulator; (10) pressure-reducing valve.

adiabatic, $n = 1.4$. Apart from these, P_A should meet the requirement that $0.25P_1 < P_A < 0.9P_2$.

(2) Motor pump model

The motor starts to work when the accumulator pressure is below the lower limit and stops when the accumulator pressure reaches the upper limit. The mathematical model is as follows [30]:

$$Q_b = V_c \omega \frac{E}{E[\alpha P_{in} + (1-\alpha)P_{out}]}, \tag{10}$$

(i) where Q_b is the oil pump flow rate, V_c is the pump displacement, ω is the motor speed, P_{out} and P_{in} are the output and input of pump pressure, respectively, E is the bulk modulus of braking fluid, and α is the pump pressure factor.

(3) High-speed switch solenoid switch model

For the on-off action of the solenoid switch that is controlled by the input voltage, there will be a certain delay phenomenon. In addition, inertia of the spool can also cause delay. The mathematical model of the high-speed on-off valve with the second-order delay is as follows:

$$G(s) = \frac{K_1 \omega}{s_2 + 2\xi\omega s + \omega^2}, \tag{11}$$

where K_1 is the current gain, ω is the valve frequency, and ξ is the equivalent damping ratio of the valve.

(4) Restrictor model

The restrictor controls the flow rate by the order of system pressure, and the mathematical model is as follows:

$$q = a \tanh\left(\frac{2\chi\sqrt{((2\Delta p)/\rho)}}{\nu \mathrm{Re}}\right)_{q\max}, \tag{12}$$

where q is the hydraulic medium flow, A is the effective circulation area of valves, χ is the hydraulic diameter, ρ is the fluid density, Δp is the valve's pressure difference, ν is the sports viscosity, and Re is the critical Reynolds number.

(5) Braking model

The braking model is as follows:

$$m\frac{d^2b}{dt^2} = -PS + C_{eq}\frac{db}{dt} + K_m(b_0 + b), \tag{13}$$

$$bS = \int_0^t Q dt,$$

where m is the brake caliper mass, b is the brake caliper displacement, P is the hydraulic cylinder braking pressure, C_{eq} is the equivalent damping, K_m is the spring stiffness, x_0 is the spring initial position, and S is the area of hydraulic cylinder cross section.

3. Control Algorithms

Due to the complex conditions of the vehicle following, the former researches have shown that the ACC system should

both control the vehicle speed and adapt to external interference such as the leading vehicle's velocity [31, 32]. Independent hierarchical control method is used in the proposed ACC model. And the control method is divided into the upper controller (decision-making controller) module and the bottom controller (underlying executive module controller).

The upper controller determines the expected acceleration a_{fdes} based on the driving information provided by the sensors and the driver's settings at this time. Based on the output from the upper controller, the bottom controller makes the vehicle dynamics system to achieve the desired acceleration.

3.1. Upper Controller Design

3.1.1. Establishment of the Follower Model. Car-following model is built based on the driver desired distance and the vehicle dynamic characteristic. Equation (14) describes the driver desired distance [33], and equation (15) shows the relationship of the vehicle dynamic:

$$d_{\text{des}} = T_{\text{h}} v_{\text{f}} + d_0, \tag{14}$$

where d_{des} is the expect distance, T_{h} is the time headway, v_{f} is the velocity of the following vehicle, and d_0 is the minimum safe distance when the two vehicles stop.

It is clear that the distance error and the velocity difference can be as follows:

$$\begin{cases} \Delta d = d_{\text{des}} - d, \\ \Delta v = v_{\text{l}} - v_{\text{f}}, \end{cases} \tag{15}$$

where d is the factual distance, Δd is the distance error, v_{l} is the velocity of the leading vehicle, and Δv is the velocity difference.

A simulation system is built to analyze the vehicle dynamic relationship. The frequency response method is adopted to identify the system input and output characteristics, and finally the transfer function is obtained as equation (18):

$$a_{\text{f}} = \frac{K}{Ts + 1} a_{\text{fdes}}, \tag{16}$$

where K is the gain and T is the time delay.

Combining equations (16)–(19), the car-following model can be as follows:

$$\dot{x} = A' x + B' u + G' v, \tag{17}$$

where $x = [\Delta d \, \Delta v \, a_{\text{f}}]^T$, $u = a_{\text{fdes}}$, $\lambda = a_{\text{p}}$, $A' = \begin{bmatrix} 0 & 1 & -T_{\text{h}} \\ 0 & 0 & -1 \\ 0 & 0 & -(1/T) \end{bmatrix}$, $B' = \begin{bmatrix} 0 \\ 0 \\ K/T \end{bmatrix}$, and $G' = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$.

x is the system status variable; u is the system input; λ is the disturbance of the input, which is the preceding vehicle's acceleration a_{p} here; A', B', and G' are the coefficient matrix of the input.

3.1.2. Performance Index Design

(1) Following performance index

The ACC system needs to control the vehicle following the leading vehicle steadily, and the following performance is manifested in the performance index of speed and the safety index [34–36].

The square sum of the speed error $\Delta v(k)$ and the distance error $\Delta d(k)$ is taken as the following performance index:

$$l_{\text{t}}(k) = w_{\Delta d} (\Delta d(k))^2 + w_{\Delta v} (\Delta v(k))^2, \tag{18}$$

where $\Delta d(k) = s_{\text{f}}(k) - s_{\text{l}}(k) - d_{\text{des}}(v_{\text{l}}(k))$, $\Delta v(k) = v_{\text{h}}(k) - v_{\text{f}}(k)$,$w_{\Delta d}$ and $w_{\Delta v}$ are the distance error weight and speed error weight, s_f is the displacement of the front car, and s_{h} is the distance of the car traveled.

The following performance index in the forecast time domain is as follows:

$$L_{\text{t}}(k) = \sum_{P}^{k=1} l_{\text{t}}(k) . \tag{19}$$

(2) Safety index

The vehicle should keep a safe distance to avoid collision. Meanwhile, the distance between the two vehicles should avoid being too large to avoid accidental vehicle insertion. The vehicle should also keep an appropriate speed difference to ensure safety and to increase traffic efficiency. And the velocity error between two vehicles should not be too large. The optimization problem is solved subject to desired intervehicle distance and acceleration limitation, which are incorporated as constraints. Therefore, the constraints of the vehicle distance error and speed error are as follows:

$$\begin{cases} 0 \le \Delta d(k) \le \Delta d_{\max}, \\ \Delta v_{\min} \le \Delta v(k) \le \Delta v_{\max}. \end{cases} \tag{20}$$

(3) System prediction optimization

Based on the index functions and constraints established before, the integrated index for the optimization problem is established as follows:

$$L = \sum_{i=1}^{P} \| \Delta d(k+i+1 \mid k) \|^2_{w_{\Delta d}} + \sum_{i=1}^{P} \| \Delta v(k+i+1 \mid k) \|^2_{w_{\Delta v}}, \tag{21}$$

where P is the length of the predictive sample time.

System constraints are as follows:

$$\begin{cases} x_{\min} \le x(k+i \mid k) \le x_{\max}, \\ y_{\min} \le y(k+i \mid k) \le y_{\max}, \end{cases} \quad i = 0 : P-1. \tag{22}$$

From the above analysis, the objective optimization problem of the system can be described as follows:

$$\min_{i=0:\ P-1} L, \tag{23}$$

subject to

$$\begin{cases} 0 \le \Delta d(k) \le \Delta d_{\max}, \\ \Delta v_{\min} \le \Delta v(k) \le \Delta v_{\max}, \\ x_{\min} \le x(k+i \mid k) \le x_{\max}, \\ y_{\min} \le y(k+i \mid k) \le y_{\max}. \end{cases} \tag{24}$$

3.2. Bottom Controller Design. The bottom controller is the system that ensures the vehicle response is constant with the expected value calculated by the upper controller as much as possible. The desired acceleration from the upper controller is translated into the desired braking pressure or throttle opening to the braking controller and the throttle controller via the inverse longitudinal braking model.

3.2.1. Throttle Controller. The PID algorithm is adopted in the throttle controller to ensure the system is working in robust and reliable condition. The algorithm takes the linear combination of the error's proportion (P), integral (I), and differential (D) as control variables and controlling object.

The PID control law is as follows:

$$\Delta y = K_{\mathrm{p}}\left[\varepsilon + \frac{1}{T_{\mathrm{I}}}\int_0^t \varepsilon \mathrm{d}t + T_{\mathrm{D}}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right], \tag{25}$$

where ε is the difference between the expected acceleration a_{des} and the actual car acceleration a, K_{p} is the proportional gain, T_{I} is the integration time constant, and T_{D} is the derivative time constant.

The conversion to transfer function is as follows:

$$(s) = \frac{U(s)}{E(s)} = K_{\mathrm{p}}\left(1 + \frac{1}{T_{\mathrm{I}}s} + T_{\mathrm{D}}s\right). \tag{26}$$

3.2.2. Braking Controller. The purpose of the braking pressure controller is to make the real braking pressure and the expected as close as possible so as to follow up the desired acceleration. Due to inertial links (mechanical system inertia, electrical system inertia, and control system inertia) of the active braking control system, real-time control value cannot act on the control system timely. Even if the parameters of the classical discrete PID algorithm are optimized, the control result still has serious lag and overshoot, which cannot meet the control requirements. The ideas for solving these problems will be given in the next paragraph.

By using proportional feedback control, it is easy to double the interference noise in feedback acceleration, which is not conducive to the stability. Feedback control structure is used to make the actual pressure follow the target pressure. And the feedforward control structure is used to improve the controller execution response to recuperate the time lag of the feedback controller. The cooperation of the feedback and feedforward controller is used to eliminate the static error and improve the accuracy of acceleration control.

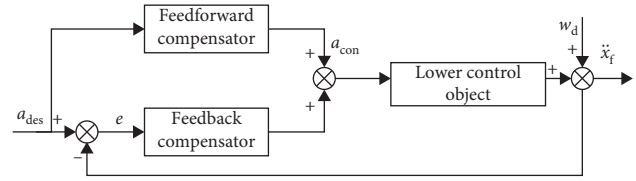


Figure 4: Structure of the braking controller.

Table 2: Duty ratio of high-pressure valve and directional valve under different pressure differences.

X (MPa)	−6	−4	−3	−2	−1	−0.5	−0.3	−0.1	
Y1 (%)	100	40	35	30	25	20	15	10	
Y2 (%)	0	0	0	0	0	0	0	0	
X (MPa)	0	0.1	0.3	0.5	1	2	3	4	6
Y1 (%)	0	0	0	0	0	0	0	0	0
Y2 (%)	0	10	20	20	25	30	40	40	100

The overall structure of the braking controller is shown in Figure 4.

Feedforward compensator uses the look-up table method. According to the pressure difference between the actual pressure and the ideal pressure of the hydraulic cylinder, the system controls the duty cycle signal of the increasing and reducing valve so as to control the pressure change rate of the hydraulic cylinder, precisely controlling the hydraulic cylinder pressure.

When the pressure difference is active, the system will select a larger duty cycle in order to quickly increase or reduce hydraulic cylinder pressure. When the pressure difference is positive, the system will choose a smaller duty cycle to accurately track the ideal pressure, and to improve the wheel pressure accuracy and robustness [37, 38], specifically as listed in Table 2, where X is the pressure difference, Y1 is the valve duty signal of the booster valve, and Y2 is the duty cycle of the valve control signal of the pressure reducing valve.

Similar to the throttle controller, the PID algorithm is used in the braking controller. The difference between the expected braking pressure and the actual braking pressure is taken as the target control variable. From the design process, one can conclude that the designs of the feedforward compensator and the feedback compensator do not affect each other and can be performed independently.

4. Simulation and Discussion

A collaborative simulation model is built by Matlab/Simulink, Carsim, and Amesim to validate the proposed algorithm. The simulation parameters and restrictions are defined, as shown in Table 3.

Simulation conditions are as follows: at 0–15 s, the leading vehicle drives at 20 m/s; at 15–25 s, the leading vehicle accelerates to 30 m/s with acceleration 1 m/s^2; at 25–35 s, the leading vehicle drives at 30 m/s; at 35–42 s, the leading vehicle slows down to 18 m/s with deceleration −1.7 m/s^2. The initial distance between two vehicles is 50 m, and the initial speed of the follower vehicle is 25 m/s. The simulation results are shown in Figures 5–8. The follow process includes 3 stages, shown as follows.

Table 3: The simulation parameters.

Items	t_s (s)	t_h (s)	τ (s)	d_0 (m)	d_c (m)	$v_{\min}$ (m/s)	$v_{\max}$ (m/s)
Value	0.2	1.5	0.7	7	5	0	36
Items	$a_{u\min}$ (m/s^2)	$a_{u\max}$ (m/s^2)	R	N	P	R	
Value	−3	2	1	5	10	diag{2, 10, 0}	

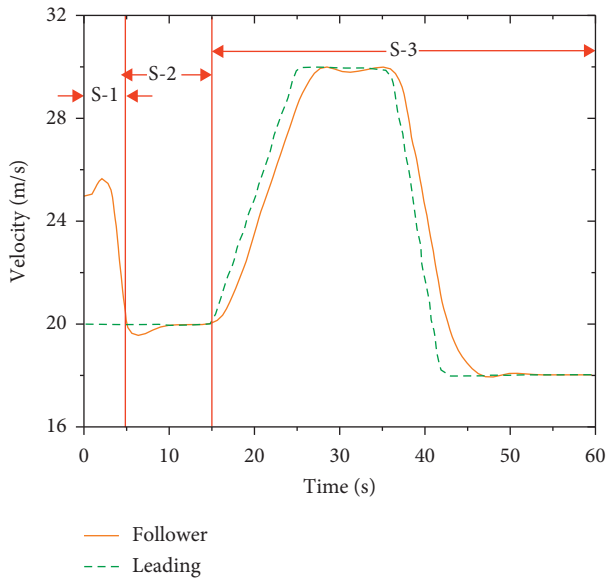


Figure 5: The vehicle velocity of the two vehicles.

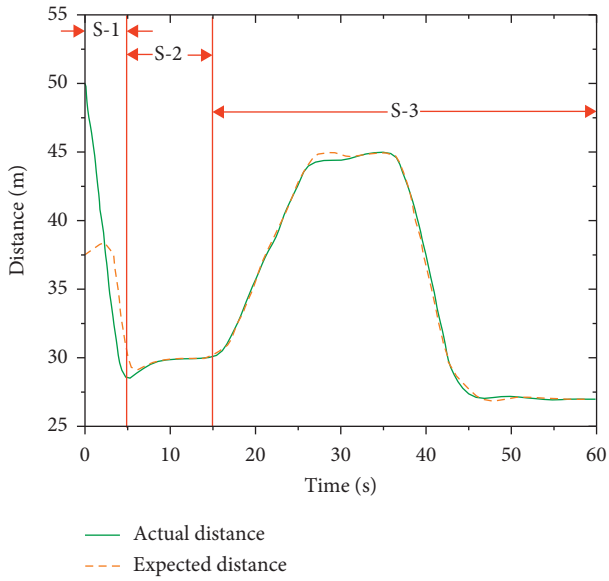


Figure 6: Distance between the two vehicles.

4.1. S-1 Follow Distance Adjustment. During the time 0 – 5 s, the actual distance between the two vehicles is greater than the desired distance; the system judges the condition is safe. The upper controller instructs the bottom controller to accelerate to shorten the distance to improve the traffic efficiency. The braking controller is on the standby mode, and the throttle is in a small opening. As the velocity increases, the expected distance also increases.

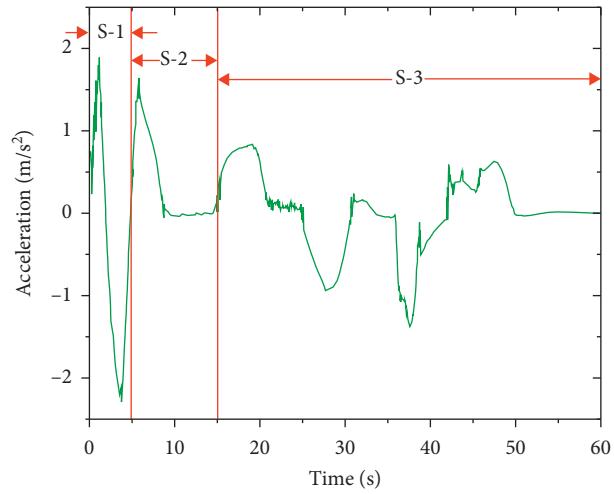


Figure 7: Acceleration of the follower vehicle.

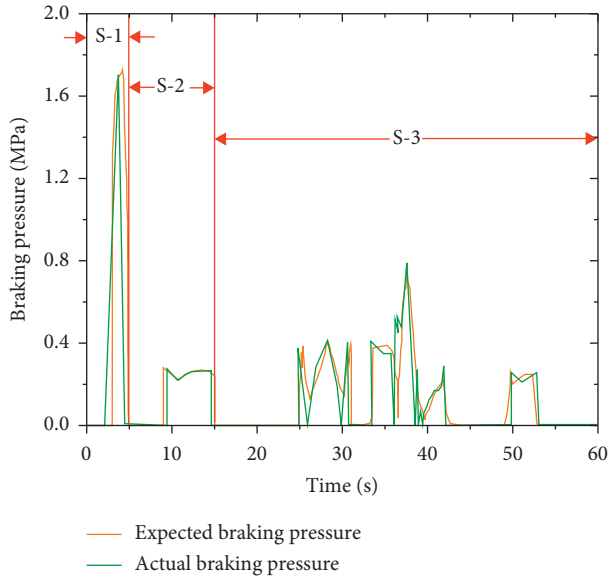


Figure 8: Braking pressure of the follower vehicle.

4.2. S-2 Follow Velocity Adjustment. As the velocity of the follower vehicle increases, the expected distance increases also. The actual distance is shortened as the velocity difference increases; the system judges the condition is in danger. The upper controller instructs the bottom controller to decelerate to lengthen the distance to improve the safety. The throttle controller is on the standby mode; the braking controller opens valves 5 and 6 and starts the brake pump (Figure 3); and the braking pressure is increased. And then, the vehicle deceleration increases, and the velocity of the follower vehicle decreases near the velocity of the leading vehicle.

4.3. S-3 Follow with the Leading Vehicle. As the leading vehicle accelerates during 15–25 s and decelerates during 35–42 s, the follower vehicle changes the throttle opening and braking pressure, keeping the desired distance and speed. As can be seen from Figure 7, although the acceleration of the vehicle has slight fluctuation, the acceleration falls in a narrow range of −3 to 2 m/s^2, which ensures the ride comfort. As can be seen in Figure 8, the target pressure follows the change of the desired acceleration response quickly and steadily with less hysteresis.

To illustrate purposes and evaluate results conveniently, a comparison with a state-of-the-art ACC used in the automotive industry research is analyzed [39]. In this paper, the safety and comfort are evaluated and provided directly. The full declaration method is used in the mode which presents detailed simulation results for one considered scenario. The results show that the velocity of the leading vehicle accelerates from 10 m/s–15 m/s and the acceleration of follower vehicle falls in a width range of −10 − 5 m/s^2. The jerk caused by application of full braking results in uncomfortable driving.

By comparison, the proposed strategy results show that the acceleration of the vehicle has slight fluctuation and fast response, which implies comfortable driving without jerky maneuvers. The ability of keeping intervehicle distance as close as possible to safe distance shows good tracking performance. In this way, both safety and comfort are achieved by utilizing the proposed model based on optimization of active braking strategy.

5. Conclusions

This research is aimed at proposing an ACC strategy considering the safety and comfort based on the active braking where the system hysteresis problem is included. For this purpose, vehicle dynamics model, vehicle reverse longitudinal dynamics model, and active hydraulic braking system model are proposed. And the models are simulated in Carsim, MATLAB/Simulink, and Amesim collaboratively. The control algorithm is proposed and optimized to improve the ride comfort. From the results, it can be seen that the velocity and distance values are preserved in the specified comfortable range although the vehicle velocity changes obviously:

(1) The control algorithm based on the model predictive control algorithm can be optimized by considering the multivariable constraints simultaneously; that is to say, the cruise-following control safety can be ensured and the ride comfort can be satisfied.

(2) The proposed algorithm is evaluated by comparison with using full deceleration simulation, and it shows active performance on position and velocity tracking. Thus, we can conclude that the proposed approach guarantees safety and comfort for ACC-equipped vehicles in low velocity conditions.

This study only focuses on the occupant kinematics during the pre-crash period; the occupant kinematics and injury indexes within the in-crash phase of such typical scenario require subsequent study.

Data Availability

All data included in this study are available upon request to the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China under Grant nos. 51775178, 51875049, and 51705035 and Hunan Science Foundation for Distinguished Young Scholars of China under Grant no. 2019JJ20017.

References

[1] L. Hu, J. Ou, J. Huang, Y. Chen, and D. Cao, "A review of research on traffic conflicts based on intelligent vehicles," *IEEE Access*, vol. 8, pp. 24471–24483, 2020.

[2] L. Hu, X. Hu, J. Wan, M. Lin, and J. Huang, "The injury epidemiology of adult riders in vehicle-two-wheeler crashes in China, Ningbo, 2011-2015," *Journal of Safety Research*, vol. 72, pp. 21–28, 2020.

[3] Y. Peng, C. Fan, L. Hu et al., "Tunnel driving occupational environment and hearing loss in train drivers in China," *Occupational and Environmental Medicine*, vol. 76, no. 2, pp. 97–104, 2019.

[4] R. Rajamani, *Adaptive Cruise Control. Vehicle Dynamics and Control*, pp. 141–170, Springer US, Boston, MA, USA, 2012.

[5] A. Weißmann, D. Görges, and X. Lin, "Energy-optimal adaptive cruise control combining model predictive control and dynamic programming," *Control Engineering Practice*, vol. 72, pp. 125–137, 2018.

[6] F. Schrödel, P. Herrmann, and N. Schwarz, "An improved multi-object adaptive cruise control approach," *IFAC-PapersOnLine*, vol. 52, no. 8, pp. 176–181, 2019.

[7] Y. He, B. Ciuffo, Q. Zhou et al., "Adaptive cruise control strategies implemented on experimental vehicles: a review," *IFAC-PapersOnLine*, vol. 52, no. 5, pp. 21–27, 2019.

[8] M. Li, Y. Chen, and C.-C. Lim, "Stability analysis of complex Network control system with dynamical topology and delays," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–10, 2020.

[9] L. Yang, Y. Chen, Z. Liu, K. Chen, and Z. Zhang, "Adaptive fuzzy control for teleoperation system with uncertain kinematics and dynamics," *International Journal of Control, Automation and Systems*, vol. 17, no. 5, pp. 1158–1166, 2019.

[10] B. Guo and Y. Chen, "Robust adaptive fault-tolerant control of four-wheel independently actuated electric vehicles," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 2882–2894, 2018.

[11] B. Goñi-Ros, W. J. Schakel, A. E. Papacharalampous et al., "Using advanced adaptive cruise control systems to reduce congestion at sags: an evaluation based on microscopic traffic simulation," *Transportation Research Part C: Emerging Technologies*, vol. 102, pp. 411–426, 2019.

[12] J. Zhan, "Setup of vehicle longitudinal dynamic model for adaptive cruise control," *Journal of Jilin University (Engineering)*, vol. 36, no. 2, pp. 157–160, 2006.

[13] S. Huang and W. Ren, "Autonomous intelligent cruise control with actuator delays," *Journal of Intelligent and Robotic Systems*, vol. 23, no. 1, pp. 27–43, 1998.

[14] A. S. A. Rachman, A. F. Idriz, S. Li, and S. Baldi, "Real-time performance and safety validation of an integrated vehicle dynamic control strategy," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 13854–13859, 2017.

[15] J. E. Naranjo, C. Gonzalez, R. Garcia, and T. DePedro, "ACC+Stop&Go maneuvers with throttle and brake fuzzy control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 213–225, 2006.

[16] P. Shakouri, A. Ordys, D. S. Laila, and M. Askari, "Adaptive cruise control system: comparing gain-scheduling PI and LQ controllers," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 12964–12969, 2011.

[17] P. Ioannou, Z. Xu, S. Eckert, D. Clemons, and T. Sieja, "Intelligent cruise control: theory and experiment," in *Proceedings of 32nd IEEE Conference on Decision and Control*, pp. 1885–1890, San Antonio, TX, USA, December 1993.

[18] S. Kitazono and H. Ohmori, "Semi-Autonomous adaptive cruise control in mixed traffic," Proceedings of 2006 SICE-CASE International Joint Conferencepp. 3240–3245, Busan, South Korea, October 2006.

[19] M. H. Lee, H. G. Park, S. H. Lee, K. S. Yoon, and K. S. Lee, "An adaptive cruise control system for autonomous vehicles," *International Journal of Precision Engineering and Manufacturing*, vol. 14, no. 3, pp. 373–380, 2013.

[20] X. Wu, G. Qin, H. Yu, S. Gao, L. Liu, and Y. Xue, "Using improved chaotic ant swarm to tune PID controller on cooperative adaptive cruise control," *Optik*, vol. 127, no. 6, pp. 3445–3450, 2016.

[21] N. C. Basjaruddin, K. Kuspriyanto, D. Saefudin et al., "Developing adaptive cruise control based on fuzzy logic using hardware simulation," *International Journal of Electrical & Computer Engineering*, vol. 4, no. 6, 2014.

[22] G. Prabhakar, S. Selvaperumal, and P. N. Pugazhenthi, "Fuzzy PD plus I control-based adaptive cruise control system in simulation and real-time environment," *IETE Journal of Research*, vol. 65, no. 1, pp. 69–79, 2019.

[23] R. Rizvi, S. Kalra, C. Gosalia et al., "Fuzzy Adaptive Cruise Control system with speed sign detection capability," in *Proceedings of 2014 IEEE International Conference on Fuzzy Systems*, pp. 968–976, IEEE, Beijing, China, July 2014.

[24] B. Ganji, A. Z. Kouzani, S. Y. Khoo, and M. Shams-Zahraei, "Adaptive cruise control of a HEV using sliding mode control," *Expert Systems with Applications*, vol. 41, no. 2, pp. 607–615, 2014.

[25] L. Hu, Y. Zhong, W. Hao et al., "Optimal route algorithm considering traffic light and energy consumption," *IEEE Access*, vol. 6, pp. 59695–59704, 2018.

[26] K. Gao, F. Han, P. Dong, N. Xiong, and R. Du, "Connected vehicle as a mobile sensor for real time queue length at signalized intersections," *Sensors*, vol. 19, no. 9, p. 2059, 2019.

[27] D. Hou, *Study on Vehicle Forward Collision Avoidance System*, Tsinghua University, Beijing, China, 2004.

[28] Y. Liu, J. Che, and C. Cao, "Advanced autonomous underwater vehicles attitude control with L 1 backstepping adaptive control strategy," *Sensors*, vol. 19, no. 22, p. 4848, 2019.

[29] H. Zheng and M. Zhao, "Development a HIL test bench for electrically controlled steering system," Proceedings of SAE Technical Paper Series, SAE International, April 2016.

[30] X. Qi J. Song et al., "Modeling and analysis of vehicle ESP hydraulic control device using AMESim," *Machine Tools and Hydraulic*, vol. 8, pp. 115-116, 2005.

[31] S. Moon, I. Moon, and K. Yi, "Design, tuning, and evaluation of a full-range adaptive cruise control system with collision avoidance," *Control Engineering Practice*, vol. 17, no. 4, pp. 442–455, 2009.

[32] D. H. Han, K. S. Yi, J. K Lee, B. S. Kim, and S. Yi, "Design and evaluation of inteligent vehicle cruise control systems using a vehicle simulator," *International Journal of Automotive Technology*, vol. 7, no. 3, pp. 377–383, 2006.

[33] D. Yanakiev and I. Kanellakopoulos, "Longitudinal control of heavy-duty vehicles for automated highway systems," Proceedings of the 1995 American Control Conference, Seattle, WA, USA, June 1995.

[34] M. Guocheng, *Research on the Adaptive Cruise Control Tracking System Applied for Motor Vehicle*, Beijing Institute of Technology, Beijing, China, 2014.

[35] Z. Zhang, D. Luo, Y. Rasim et al., "A vehicle active safety model: vehicle speed control based on driver vigilance detection using wearable EEG and sparse representation," *Sensors*, vol. 16, no. 2, p. 242, 2016.

[36] R. Du, G. Qiu, K. Gao, L. Hu, and L. Liu, "Abnormal road surface recognition based on smartphone acceleration sensor," *Sensors*, vol. 20, no. 2, p. 451, 2020.

[37] L. Hu, X. Hu, Y. Che et al., "Reliable state of charge estimation of battery packs using fuzzy adaptive federated filtering," *Applied Energy*, vol. 262, 2020.

[38] Z. Zhang, L. Zhang, L. Hu, C. Huang et al., "Active cell balancing of lithium-ion battery pack based on average state of charge," *International Journal of Energy Research*, vol. 44, no. 4, pp. 2535–2548, 2020.

[39] S. Magdici and M. Althoff, "Adaptive cruise control with safety guarantees for autonomous vehicles," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 5774–5781, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9656434, 18 pages
https://doi.org/10.1155/2020/9656434

Research Article

Probability of Roadside Accidents for Curved Sections on Highways

Guozhu Cheng,[1] **Rui Cheng**,[1] **Yulong Pei**,[1] **and Liang Xu**[2]

[1]*School of Traffic and Transportation, Northeast Forestry University, 150040 Harbin, China*
[2]*School of Civil Engineering, Changchun Institute of Technology, 130012 Changchun, China*

Correspondence should be addressed to Guozhu Cheng; guozhucheng@126.com

Received 1 April 2020; Revised 9 May 2020; Accepted 16 May 2020; Published 30 May 2020

Guest Editor: Yong Chen

To predict the probability of roadside accidents for curved sections on highways, we chose eight risk factors that may contribute to the probability of roadside accidents to conduct simulation tests and collected a total of 12,800 data obtained from the PC-crash software. The chi-squared automatic interaction detection (CHAID) decision tree technique was employed to identify significant risk factors and explore the influence of different combinations of significant risk factors on roadside accidents according to the generated decision rules, so as to propose specific improved countermeasures as the reference for the revision of the Design Specification for Highway Alignment (JTG D20-2017) of China. Considering the effects of related interactions among different risk factors on roadside accidents, path analysis was applied to investigate the importance of the significant risk factors. The results showed that the significant risk factors were in decreasing order of importance, vehicle speed, horizontal curve radius, vehicle type, adhesion coefficient, hard shoulder width, and longitudinal slope. The first five important factors were chosen as predictors of the probability of roadside accidents in the Bayesian network analysis to establish the probability prediction model of roadside accidents. Eventually, the thresholds of the various factors for roadside accident blackspot identification were given according to probabilistic prediction results.

1. Introduction

Roadside accidents occur when a vehicle leaves the travel line, crosses an edge line or a centre line, collides with trees, guardrails, utility poles, and other natural or man-made objects located on roadsides, or overturns or falls into deep ditches or rivers. According to the Fatal Accident Reporting System (FARS), these accident types account for more than 39% of fatal accidents in the United States [1]. In China, roadside accidents account for approximately 50% of the collisions in which more than three people perish [2]. A European study also shows that 20% of all traffic accidents every year are roadside accidents; however, the fatality rate is over 35% in these accidents, and approximately one-third of run-off-road (ROR) collision fatalities occurred on curved road sections [3], the road type upon which this study focuses.

There are several complex reasons a vehicle departs from the travelled path, such as an inappropriate avoidance manoeuvre or inattention of a driver, crossing a curve segment with a high speed, or understeering. A variety of contributing factors to roadside accidents have been identified based on various collected data and data analysis methods. Numerous studies have confirmed that highway geometric design indexes (i.e., roadway characteristics and roadside characteristics) play a significant role in whether a crash occurs resulting from driver error [4], especially for curve sections on highways. In terms of roadway characteristics, a wider shoulder has been found to decrease the occurrence of ROR accidents on horizontal curves [3, 5], but the increase of the shoulder width is also associated with an increasing vehicle operating speed [6]. The frequency of ROR accidents will increase if vehicles travel in a narrower lane because the requirement for sharing the roadway with other vehicles increases the chance of conflicts, whereas driveway density has little impact on ROR accidents [3, 5]. Moreover, some research has confirmed that pavement edge

drop-off and low friction of pavement surfaces tend to cause a high frequency of single-vehicle accidents [7]. Sharp curves are also a key factor contributing to roadside accident occurrence and approximately 30% of the ROR events occur on road curves [8–10]. In an attempt to identify roadside design risk factors, a large number of studies have been implemented, involving analysis of the relationship between the frequency of roadside accidents and critical slope, fences, bridges, guardrail, ditches, utility pole density, distance to pole and distance to tree, and so on [3, 11–17].

Among the environmental factors, most ROR accidents tend to occur on weekends [3, 5]. Area type and lighting conditions are found to be significant factors contributing to the probability of roadside accidents [9]. A study investigated whether road type and local amenities are associated with single-vehicle accident frequency [18]. Additionally, local population density is also related to accident occurrence [19].

In terms of human factors, the National Highway Traffic Safety Administration (NHTSA) suggested that driver distraction, fatigue, driver's degree of familiarity with the roadway, blood alcohol presence, age, and gender were the most significant factors contributing to roadside accidents [18], and 30% of these accidents occurred due to driver inattention [8–10]. All of these factors have a direct or indirect effect on changes in vehicle speeds, and the risk of accidents increases, followed by an increase in vehicle speeds.

From a methodological perspective, different methods have been employed to determine these factors. Originally, Zegeer and Deacon [20] developed a lognormal regression model to investigate the relationship between ROR accident frequencies and various variables, such as average annual daily traffic (AADT), shoulder width, lane width, terrain type, and clear roadside recovery distance (CRRD). In a further study, they added some variables (i.e., density and lateral offset of the roadside object) to the previous model [21]. However, this conventional linear regression model has been demonstrated as inappropriate and to be often erroneous [22–24]. More appropriate prediction models for accident frequency (i.e., Poisson and negative binomial (NB) regression models) have been widely used in recent decades [22, 25–27]. To address the problem of zero-inflated counting processes in accident frequency analysis, the zero-inflected Poisson (ZIP) and zero-inflected negative binomial (ZINB) regression models have gained considerable acceptance [28–32].

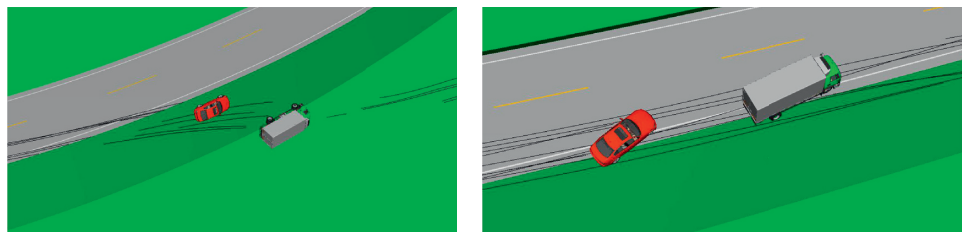
Although there have been a considerable number of roadside accident frequency studies, few studies have focused on the quantitative analysis of roadside accident probability. Various approaches (i.e., Poisson model, NB model, ZIP model, and ZINB model) are capable of predicting the number or frequency of roadside accidents based on mass accident data but cannot precisely calculate probability values under the effects of various variables. Moreover, the research results based on the prediction of accident frequency or number are often influenced by different traffic characteristics in various regions, which is not universal. Considering that accident probability is more able to represent the degree of frequent accidents, it is better to carry out the prediction of accident probability than to carry out the prediction of accident frequency or number. To identify the roadside accidents blackspot and reduce the accidents probability, we therefore used a data mining technique (i.e., CHAID decision tree technique) to identify significant risk factors contributing to roadside accidents and another data mining technique (i.e., Bayesian network analysis) to establish the probability prediction model of roadside accidents. Additionally, we investigate the importance of various variables under the interactions of accident occurrence by developing a path analysis based on a logistic regression model. To the best of our knowledge, no research has used these three methods together in the study of the probability of roadside accidents.

2. Data and Methodology

2.1. Data. Substantial statistical analysis generally relies on historical accident data. However, the constantly changing traffic environment, the high cost of maintaining or collecting roadside accident data, and the long-term lack of detailed data have formed a barrier to developing a study of the relationship between road design and the probability and severity of accidents [25, 33]. Automobile dynamics simulation technology, regarded as an alternative approach, has been popularly applied to obtain accident data in recent years. Compared to collected accident data, the data from simulation software has the following advantages: (1) comprehensive accident information, (2) no consideration of the impact of time and traffic condition on accident data, (3) universal applicability in the absence of regional characteristics, (4) low cost of research, and (5) free choice of variables according to your interest. In the present study, we used accident data obtained from PC-crash simulation software. This software is primarily developed to take accident reconstruction and has been used for collisions between vehicles [34] and accidents involving vehicles and pedestrians [35], as well as single-vehicle accidents [36, 37]. It has been demonstrated that PC-crash has good performance in simulating single-vehicle (rollover) accidents [36–40].

We chose highway geometric design indexes (horizontal curve radius, hard shoulder width, longitudinal slope, superelevation slope, and width value of the curve), pavement condition (adhesion coefficient), and traffic characteristics (vehicle speed and vehicle type) as input variables, and vehicle final states as the output variable. In the present study, the final states of vehicles include departing from the roadway and not departing from the roadway. The former state refers to the circumstances of vehicle rollover or any of the vehicle wheels entering the slope represents the occurrence of a roadside accident (see Figure 1), and the latter state involves a vehicle running normally and represents no occurrence of a roadside accident.

Consider that the values of slope gradient and slope height mainly affect the severity of the roadside accident when the vehicle enters the slope and have little effect on the occurrence of the roadside accident. In addition, in



(a) (b)

Figure 1: Occurrence of roadside accidents. (a) Vehicle rollover. (b) Wheel of vehicle entering slope.

combination with the provisions of carriageway width and crown slope in the Design Specification for Highway Alignment (DSHA) (JTG D20-2017) of China [41], in the PC-crash simulation software, we built a two-way two-lane road model with a carriageway width of 3.75 m, a crown slope of 2%, a slope gradient of 1 : 1 and a slope height of 5 m as a typical representative, and two rigid models for the car and truck. "BMW-116d autom" and "ASCHERSLEBEN KAROSS" were chosen as the represented car and truck model, respectively, and the initial position of the vehicles were set on the centre of the one-way lane; their parameters are shown in Table 1.

Notably, in the vehicle parameter setting, the steering of the vehicle was set ahead to match with different horizontal curve radii because we are unable to involve the driving behaviour factors considering the characteristics of the simulation software. For instance, when the horizontal curve radius is 200 m, the steering degree of the car is automatically updated to 1.57° and 1.54° to match the above radius by setting the turning radius of the vehicle as 200 m in the simulation software (see Figure 2(a)). In terms of the width value of the curve setting, according to the DSHA [41], the widened value was set only when the horizontal curve radius was no more than 250 m (see Figure 2(b)), and in case the horizontal curve radius was 200 m~250 m, the widened value was 0.4 m for car and 0.6 m for truck. Therefore, the corresponding widened values were set for different vehicle types in the simulation test.

Each variable value is shown in Table 2. Among these variables, horizontal curve radius, hard shoulder width, width value of the curve, adhesion coefficient, vehicle speed, and vehicle type can be set directly in the simulation software; however, longitudinal slope and superelevation slope need some complex operations to be set. For instance, the setting of the longitudinal slope can be achieved by adjusting the difference in height from the beginning to the end of the test section, and the difference in height h_1 is calculated as follows:

$$h_1 = l_1 \times \sin\left(\arctan\left(i_1\right)\right), \tag{1}$$

where l_1 represents the length of the test section (m) and i_1 denotes the longitudinal slope (%) (the value of the downhill slope is positive).

Similarly, the setting of superelevation slope can be achieved by adjusting the difference in height from the outside to the inside of the test section, and the difference in height h_2 is shown as follows:

$$h_2 = l_2 \times \sin\left(\arctan\left(i_2\right)\right), \tag{2}$$

where l_2 is the width of the test section (including hard shoulder width) (m) and i_2 denotes the superelevation slope (%), which is set in the middle of the test section and its value is positive when the outside height is greater than the inside height of the test section.

According to the value of each variable (excluding the width value of the curve) from the highway geometric design indexes and pavement condition (see Table 2), $5 \times 4 \times 4 \times 4 \times 4 = 1280$ combinations were established, and then two kinds of the flat curve and curved slope combination sections were constructed corresponding to different hard shoulder widths, adhesion coefficients, and superelevation slopes. By applying 5 initial speeds to the vehicle and setting the width value of curve according to different vehicle types, simulation experiments were carried out for truck and car. Eventually, $1280 \times 5 \times 2 = 12800$ simulation data were collected, in which the data of no roadside accidents occurrence was 9,973 (77.9%) and the data involving the occurrence of roadside accidents was 2,827 (22.1%).

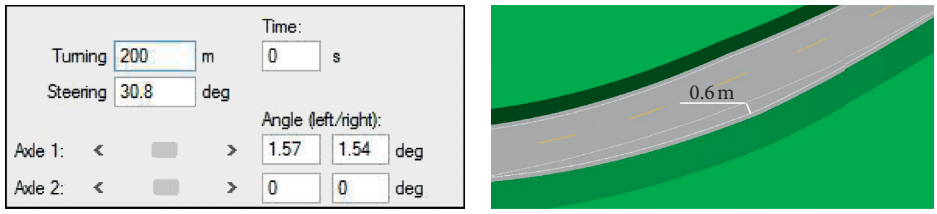
2.2. CHAID Decision Tree Technique. The CHAID decision tree, as a data mining technique, has been widely applied in various fields, such as the airline industry and public transport management. However, few studies have investigated traffic risk, especially for roadside accidents.

The CHAID decision tree is a technique of database segmentation that is capable of extracting significant information from a large quantity of data [42, 43]. After a test order is conducted, the data are split by means of a statistical algorithm in CHAID. The original node on the independent variable is split into as many subgroups as possible, which are significantly different from binary variables. The process then splits these new nodes according to the variables that distinguish each of them. This process continues until no other splits are significant.

The CHAID analysis is generally called tree analysis, similar to a trunk (i.e., original node) being split into multiple branches; then, more branches until the trunk cannot be split any further in which case overfitting occurs. To identify optimal splits, the chi-square independence test is employed to examine and test the cross tabulations between each of the input variables (i.e., predictors of the occurrence of roadside accidents) and the outcome variables (i.e., occurrence of roadside accidents). The CHAID decision

Table 1: Vehicle parameter.

Parameter	Value	
	Car	Truck
Length (m)	4.325	6.370
Width (m)	1.765	2.500
Height (m)	1.420	3.100
Wheelbase (m)	2.690	3.700
Weight (kg)	1385	7200
Height of centre of gravity (m)	0.450	1.200
Distance of height of centre of gravity from front axle (m)	1.210	1.070
Tyre pattern	215/50 R 16 (621 mm)	7.50 R 16 (719 mm)
ABS	Yes	Yes
ESP	Yes	No



(a) (b)

Figure 2: Setting (a) the steering degree for the vehicle and (b) the width value of the curve.

Table 2: Description of variables.

Variable		Value				
Highway geometric design indexes	Horizontal curve radius (m)	200	300	400	500	600
	Hard shoulder width (m)	0.75	1.5		2.25	3.00
	Longitudinal slope (%)	0	2		4	6
	Superelevation slope (%)	0	2		4	6
	Width value of curve (m)	0.4			0.6	
Pavement condition	Adhesion coefficient	0.2	0.4		0.6	0.8
Traffic characteristics	Vehicle type	"Truck" = 0			"Car" = 1	
	Vehicle speed (km/h)	40	60	80	100	120
Output variable	Vehicle final state	"No departing from road" = 0			"Departing from road" = 1	

tree is, therefore, capable of providing detail that identifies the significant factors that result in the highest or lowest risk of roadside accidents using a series of if-then-else rules.

Furthermore, to prevent the occurrence of overfitting, CHAID uses P values with a Bonferroni correction as splitting criteria; P value criteria are sensitive to the number of data involved in the split and tend to avoid splitting into too small groups [44]; the smaller the P value is, the greater the goodness of tree model fitting. The P value of the F statistic for the difference in mean values is shown as follows [45]:

$$F = \frac{\mathrm{TSS} - \mathrm{WSS}\left(\mathrm{WSS}/\left(\mathrm{g}-1\right)\right)}{\mathrm{WSS}/\left(n-\mathrm{g}\right)} \sim F_{(\mathrm{g}-1),(n-g)}, \quad (3)$$

where TSS denotes the total sum of squares before the split, WSS is the variance, g represents the nodes generated by the split, and n is the number of categories of variables.

2.3. Path Analysis. Path analysis is a form of structural equation modelling (SEM), in which all the variables are observed variables. In the present study, SEM was used because the mediated and moderated relationships of a set of variables can be tested in SEM. In other words, SEM can not only test the direct impact of independent variables on dependent variables but also analyses the indirect effect on dependent variables through other variables (mediators). In path analysis, mediation, moderation, moderated mediation, and mediated moderation can all be tested [46], and mediation is a statistical approach applied to understand how a variable x delivers its effects to another variable z. In other words, whether the effect of x on z is direct only, indirect only or both direct, and indirect can be obtained in mediation analysis [47].

A simple mediation model describes a model in which the independent variable x has an impact on the dependent

variable z through a single mediator variable y (i.e., x is assumed to have an impact on y, and this impact then transmits to z, apart from the direct relationship between x and z). Two basic mediation models are built in equations (4) and (5). In particular, equation (4) represents the combination of the paths from x to z and y to z, and equation (5) represents the path from x to y:

$$z = \alpha_0 + \alpha_1 y_i + \alpha_2 x_i + \varepsilon_1, \tag{4}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_2, \tag{5}$$

where z is the outcome variable, y_i is the mediator variable, x_i is the independent variable, ε_1 and ε_2 are the errors, α_0 and β_0 are the intercepts of the models, and α_1, β_1, and α_2 are partial regression coefficients of the models.

However, these partial regression coefficients in the above models denote the direct effect of various variables but cannot reflect the magnitude of impact from these variables on the outcome variable due to the presence of their different units and standard deviation. For this purpose, a binary logistic regression model was fitted to obtain a standard regression coefficient that can meet the demand of testing the magnitude of direct effects from input variables on the outcome variable as follows:

$$\alpha_i' = \alpha_i \left(\frac{S_i}{S_Z} \right), \tag{6}$$

where α_i' is the standard regression coefficient of x_i; α_i is the partial regression coefficient of x_i; S_i is the standard deviation of x_i; and S_Z is the standard deviation of the Z random variable in the logistic regression model, set as $\pi/\sqrt{3}$ [48]. α_i' represents the magnitude of the direct effect of x_i on the outcome variable (z).

Then, the indirect effect of x on z through all other mediator variables (y_i) can be estimated using the product-of-coefficient estimator as in [46]

$$\gamma_i = \sum_{j=1}^{n} \beta_{ij} \alpha_j', \tag{7}$$

where γ_i represents the magnitude of the indirect effect of x_i on z and β_{ij} is the correlation coefficient between x_i and y_j. Finally, the overall effects χ_i (i.e., both the direct and indirect effects) of x_i on z can be computed as follows:

$$\chi_i = \gamma_i + \alpha_i'. \tag{8}$$

2.4. Bayesian Network Analysis. The Bayesian network became popular in the late 1990s and has been increasingly used since 2000. The Bayesian network, also known as the belief network, is regarded as one of the most effective theoretical models applied for representation and reasoning of uncertain knowledge. Bayesian nets and probabilistic directed acyclic graphs are technologies for graphically representing the joint probability distribution of a set of selected variables [49, 50]. The structure of the Bayesian network is a directed acyclic graph, in which node sets represent various variables and directed edges denote the dependencies between variables. The confidence level or correlation strength between variables can be described using a conditional probability table (CPT). Tasks such as prediction, diagnosis, and classification can be realized through statistical inference functions and automatic learning of the Bayesian theorem. The structure of the Bayesian network can be regarded as the qualitative part of the model, while the added probability parameter represents a quantitative dimension to the model [51]. The Bayesian network represents various forms of uncertainty by using probability and applies probabilistic rules for achieving the training and reasoning processes, as shown in equations (9) and (10), respectively:

$$P\left(B_{ij} \mid A_j\right) = \frac{P\left(A_j \mid B_{ij}\right) P\left(B_{ij}\right)}{\sum_{j=1}^{m} P\left(A_j \mid B_{ij}\right) P\left(B_{ij}\right)}, \tag{9}$$

$$P\left(A_j \mid B_{1j} \cdots B_{nj}\right) = \frac{P\left(B_{1j} \mid A_j\right) P\left(B_{2j} \mid A_j\right) \cdots P\left(B_{nj} \mid A_j\right) P\left(A_j\right)}{\sum_{j=1}^{m} P\left(B_{1j} \mid A_j\right) P\left(B_{2j} \mid A_j\right) \cdots P\left(B_{nj} \mid A_j\right) P\left(A_j\right)}, \tag{10}$$

where $P(A_j \mid B_{ij})$ $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$ represents the prior probability of A_j (i.e., the final state of the vehicles in the accident simulation) under the effect of variable B_{ij} (i.e., the risk factor leading to roadside accident), $P(B_{ij} \mid A_j)$denotes the conditional probability of variable B_{ij} under the premise of A_j occurrence, and $P(A_j \mid B_{ij} \ldots B_{nj})$ is the posterior probability of A_j under the effects of a set of variables $(B_1 \ldots B_{nj})$. The above processes can also be achieved by efficient algorithms, such as the gradient descent (GD) algorithm in Netica software.

Compared to other theoretical models, the Bayesian network is suitable for traffic safety studies based on the following advantages: (1) combining data with expert experience and prior knowledge, (2) avoiding overfitting, (3) dealing with missing data, and (4) denoting causality by means of providing an understandable graph [52]. The Bayesian network, as an effective tool for developing an accident prediction model, has been widely used to predict accident injury [53–56] and frequency [57–59] and has demonstrated higher accuracy in predicting crash severity compared to regression models [60]. However, few studies have involved quantitative analysis of the probability of roadside accidents using the Bayesian network.

3. Results and Discussion

3.1. Identification of Risk Factor. For crossvalidation, we divided the accident data obtained from the simulation into a training dataset (70%) and a test dataset (30%). The training data were applied to fit the model and estimate the model parameters, while the test data were used to determine the model for its ability to generalize and confirm the model's applicability to independent variables. In the present study, we used exhaustive CHAID because it is superior in checking all possible splits [61]. To limit the growth of decision trees, we set the classification level to

four. Additionally, to minimize the intrinsic imbalanced nature of the data, a misclassification cost ratio of 100 : 1 was selected to promote CHAID to identify roadside accidents accurately more often.

CHAID provides the percentage of records with a particular value to the outcome variable, and the given value represents the confidence (accuracy) of the generated rules for the input variables. The overall classification accuracy of both the training set and testing set was 94% using the CHAID decision tree. Moreover, the p value in each node of both the training set and testing set was $0.001 < 0.05$ (significance level), which indicates quite accurate classification with no overfitting.

CHAID analysis took 3,783 samples from the overall dataset for testing, and the percentage of roadside accident data was 22%. All data involving roadside accidents and nonroadside accident occurrences were divided into 67 subgroups from the parent node to child nodes through different branches. The percentage of roadside accidents varied from 0% to 100%. The decision tree included horizontal curve radius, hard shoulder width, longitudinal slope, adhesion coefficient, vehicle speed, and vehicle type in the final structure, which indicates that these variables are significant risk factors in determining the occurrence of roadside accidents. Other predictors not involved in the tree structure (i.e., superelevation slope and width value of the curve) only play a slight role in improving roadside safety performance.

Figure 3 only displays major tree structures that have a higher accuracy of generated classification rules due to the limitation of scope. The split at the first classification level was according to vehicle speed, which indicates that the influence of vehicle speed on the roadside accidents is relatively significant, while other risk factors are considered as nonsignificant risk factors at this classification level. By analogy, the classification of data at the second to four level could be obtained. Through the analysis of 3 783 test data, the generated decision rules were screened and sorted, as shown in Table 3.

Each decision rule in Table 3 corresponds to different combinations of risk factors. By analyzing the influence of these combinations on the percentage of roadside accidents, some important conclusions and specific improved measure were obtained as follows:

(1) According to decision rule 1, when $V \leq 40$ km/h, other risk factors had no significant effect on the roadside accidents, and the percentage of roadside accidents was 0%. 40 km/h is, therefore, considered as the relatively safe speed to ensure the no occurrence of roadside accidents. Decision rules 2~12 presented that when $V > 40$ km/h, there was a significant influence of horizontal curve radius on roadside accidents, and roadside accidents tend to decrease with an increase in the horizontal curve radius. Decision rules 12 showed that only when 100 km/h $< V \leq$ 120 km/h and 200 m $< R \leq$ 300 m, the longitudinal slope had a certain impact on the occurrence of truck roadside accidents, and in case of longitudinal slope ≥4%, the accidents percentage increased to 100%. This finding shows that the frequency of roadside accidents increases with a larger longitudinal slope.

(2) Decision rules 6 and 9 presented that the percentage of roadside accidents for trucks was larger than that for cars under the same road condition, which can be concluded that trucks have a higher risk of roadside accidents compared to accidents involving cars because the higher centre of gravity for trucks cause them to be more likely to rollover than cars.

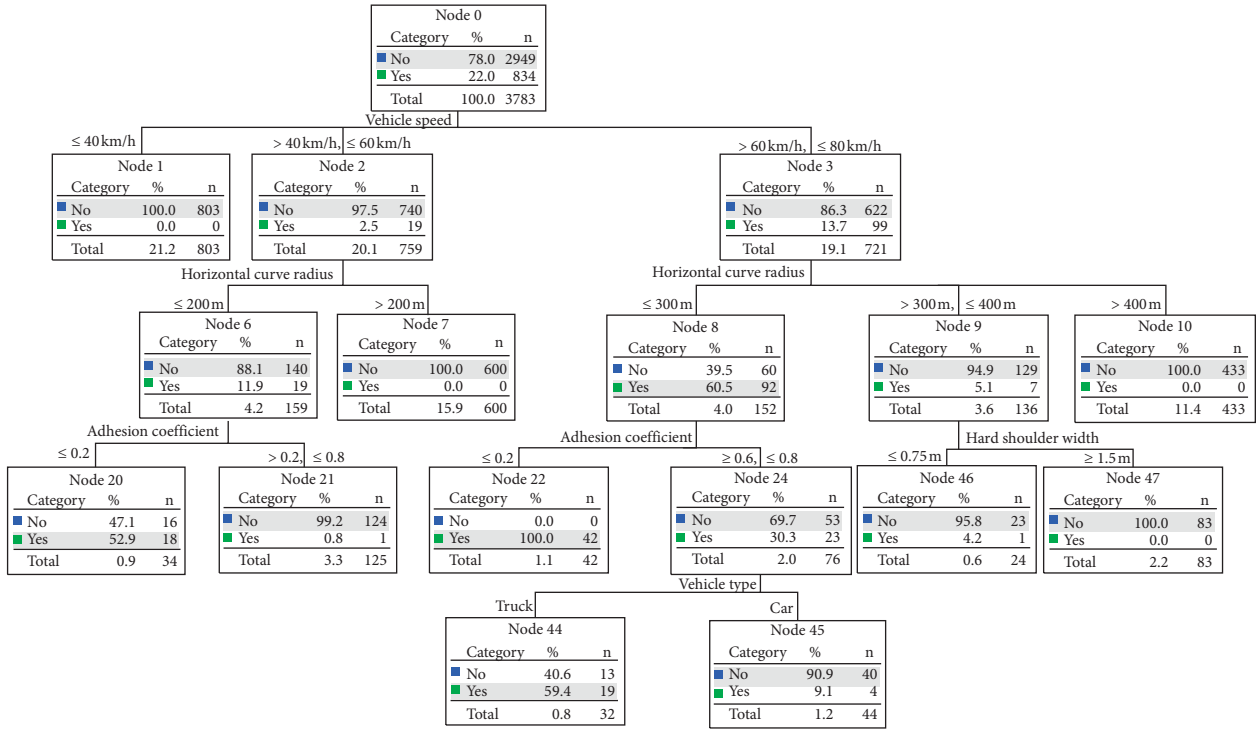
(3) It can be seen from decision rules 2 and 5 that, in case of 40 km/h $< V \leq$ 60 km/h and $R \leq 200$ m, as well as when 60 km/h $< V \leq$ 80 km/h and $R \leq 300$ m, adhesion coefficient showed a significant impact on roadside accidents, and the percentage of roadside accidents gradually decreased as adhesion coefficient increased. Therefore, antislip measures should be strengthened for the highway with the above operating speed and horizontal curve radius. The abovementioned conclusion can be used as a supplement to the revision of the DSHA.

(4) According to decision rule 7, when 60 km/h $< V \leq$ 80 km/h and 300 m $< R \leq$ 400 m, hard shoulder width played a certain role in reducing roadside accidents, but the improvement is not obvious. According to decision rules 8 and 9, in case of 80 km/h $< V \leq$ 100 km/h and 300 m $< R \leq$ 600 m, hard shoulder width had a significant impact on roadside accidents, and setting hard shoulder width ≥1.5 m could obviously reduce the percentage of roadside accidents. Therefore, for the highway with the above operating speed and horizontal curve radius, the width of hard shoulder should be set as ≥1.5 m.

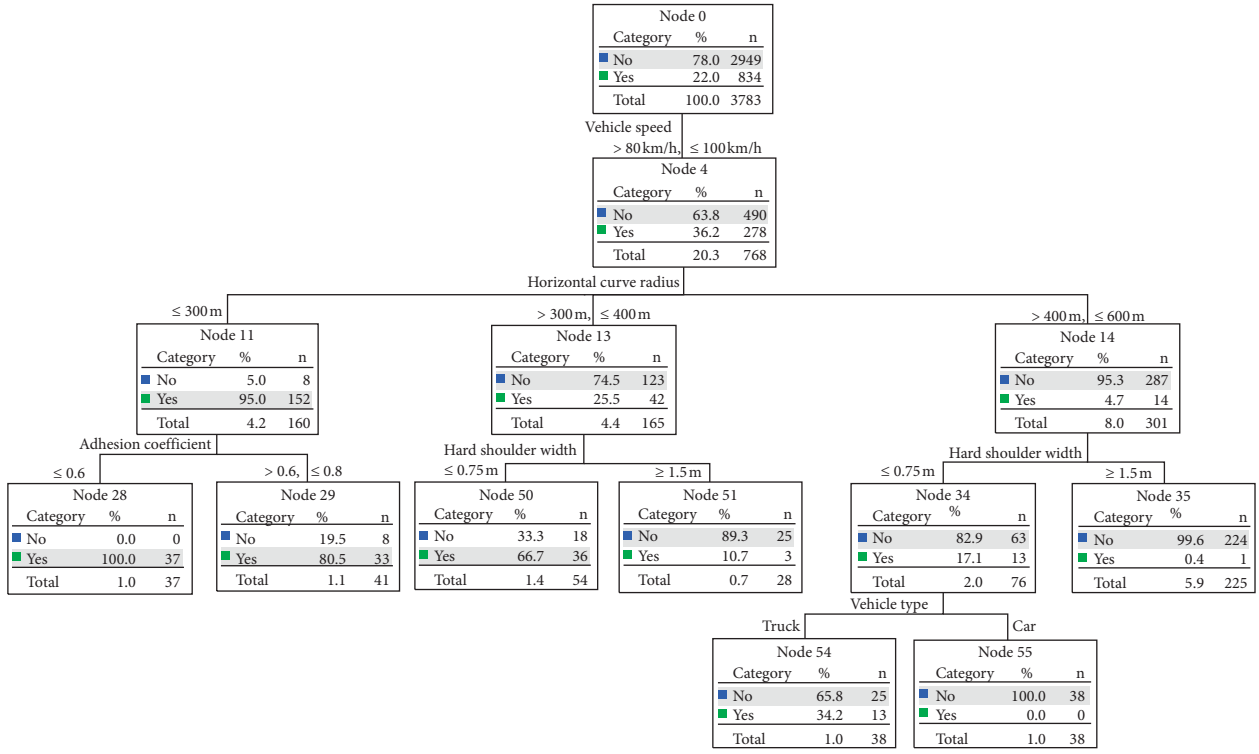
(5) Decision rule 9 showed that, in case of 80 km/h $< V \leq$ 100 km/h and 400 m $< R \leq$ 600 m, if the width of hard shoulder ≤ 0.75 m, the percentage of truck roadside accidents was 34.2% and that of car roadside accidents was 0%; if the hard shoulder width ≥1.5 m, the percentage of roadside accidents was only 0.4% for both trucks and cars. This finding adequately illustrates that the hard shoulder width has more significant impact on the frequency of roadside accidents involving trucks than cars.

(6) It can be seen from decision rules 10 and 11 that when 100 km/h $< V \leq$ 120 km/h and 300 m $< R \leq$ 600 m, a setting of hard shoulder width ≥2.25 m can effectively avoid the occurrence of truck roadside accidents. Therefore, for freeway with the above operating speed and horizontal curve radius, the width of hard shoulder should be set as ≥2.25 m to ensure driving safety of trucks.

Using decision tree analysis, we discussed the relationship between different combinations of risk predictors and the occurrence of roadside accidents and identified the significant risk factors resulting in roadside accidents.

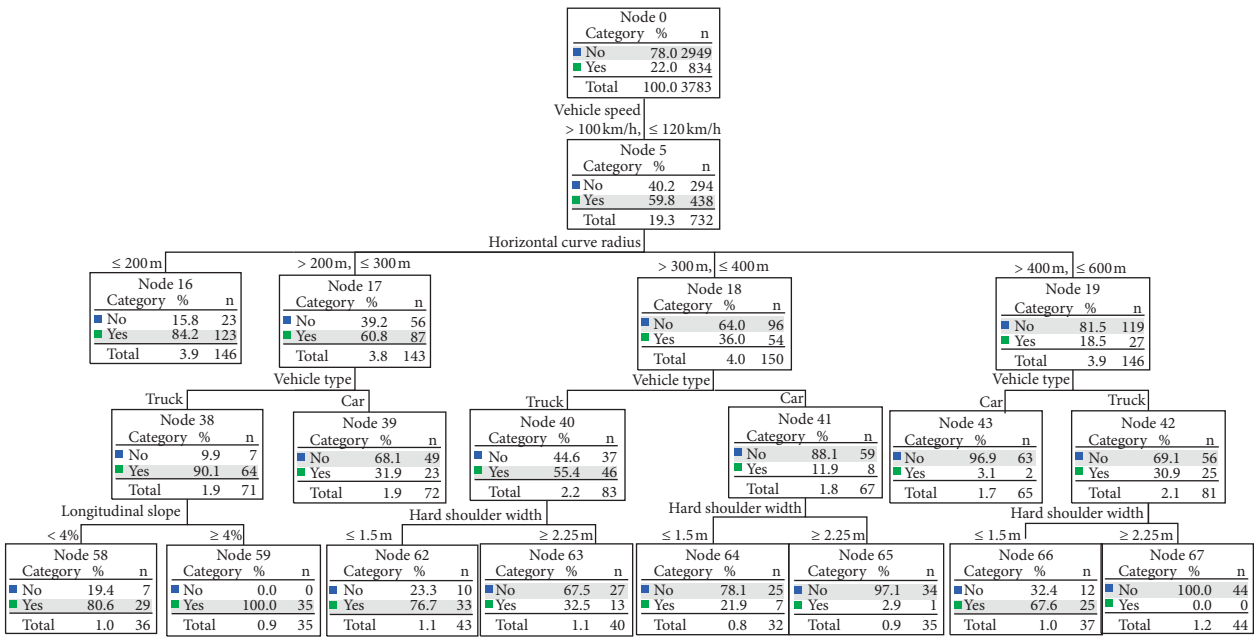


(a)



(b)

Figure 3: Continued.



(c)

Figure 3: Decision tree for the identification of risk factors. (a) Tree 1. (b) Tree 2. (c) Tree 3. *Note*. "No" represents "no roadside accident occurrence" and "Yes" represents "roadside accident occurrence".

Table 3: Decision rules.

No.	Classification level 1	2	3	4	Percentage (%)
1	≤40	—	—	—	0
2	$40 < V \le 60$	$R \le 200$	$\mu \le 0.2$	—	52.9
			$0.2 < \mu \le 0.8$		0.8
3		$200 < R$	—	—	0
		$R \le 300$	—	—	60.5
4		$300 < R \le 400$	—	—	5.1
		$400 < R$	—	—	0
5	$60 < V \le 80$	$R \le 300$	$\mu \le 0.2$	—	100
			$0.6 \le \mu \le 0.8$	—	30.3
6			$0.6 \le \mu \le 0.8$	Truck	59.4
				Car	9.1
7		$300 < R \le 400$	$w \le 0.75$	—	4.2
			$w \ge 1.5$		0
8	$80 < V \le 100$	$300 < R \le 400$	$w \le 0.75$	—	66.7
			$w \ge 1.5$		10.7
9		$400 < R \le 600$	$w \le 0.75$	Truck	34.2
				Car	0
			$w \ge 1.5$	—	0.4
10	$100 < V \le 120$	$300 < R \le 400$	Truck	$w \le 1.5$	76.7
				$w \ge 2.25$	32.5
11		$400 < R \le 600$	Truck	$w \le 1.5$	67.6
				$w \ge 2.25$	0
12		$200 < R \le 300$	Truck	$\alpha < 4$	80.6
				$\alpha \ge 4$	100

Note: V represents the vehicle speed (km/h), R is the horizontal curve radius (m), μ represents the adhesion coefficient, w denotes the hard shoulder width (m), α represents the longitudinal slope (%), and — indicates that the risk factors at the corresponding decision rules have no significant impact on roadside accidents.

However, the magnitude of the importance of these factors has not been investigated. To obtain a deeper insight into the interactions of factors and their impacts on roadside accidents, a path analysis based on a logistic regression model was built.

3.2. Importance of Risk Factors. We input the risk factors (horizontal curve radius, hard shoulder width, longitudinal slope, adhesion coefficient, vehicle speed, and vehicle type) into the path analysis model and found that these factors were also statistically significant because they were all retained by the model. The coefficient of determination $R^2 = 0.868$, illustrating the model fit, is good. Table 4 shows the outcomes of the model and represents the direct effects of different variables on roadside accident occurrence. According to the magnitude of direct effects, the most important risk factors were in decreasing order of importance, vehicle speed (3.321), horizontal curve radius (−2.572), vehicle type (−1.005), adhesion coefficient (−0.827), hard shoulder width (−0.812), and longitudinal slope (0.314). As expected, vehicle speed and longitudinal slope were found to be positively correlated with the occurrence of roadside accidents. In contrast, horizontal curve radius, vehicle type, adhesion coefficient, and hard shoulder width were inversely related to roadside accidents.

It is important to note that unlike real accident data, there seemed to be no interaction between factors in the present study because the values of all these factors were set artificially in the simulation. However, to investigate the indirect effects caused by the interaction of variables on the occurrence of roadside accidents, we assumed that the correlation coefficient between variables could be regarded as their interaction.

A structural diagram of path analysis is shown as Figure 4. This figure represents that all risk factors are correlated

Table 4: Modelling results.

Variable	Parameter estimate	S.E.[a]	S.D.[b]	P value	Standard parameter estimate
Horizontal curve radius	−0.033	0.001	141.388	<0.05**	−2.572
Hard shoulder width	−2.527	0.108	0.583	<0.05**	−0.812
Longitudinal slope	0.430	0.027	1.325	<0.05**	0.314
Adhesion coefficient	−6.699	0.279	0.224	<0.05**	−0.827
Vehicle speed	0.213	0.006	28.282	<0.05**	3.321
Vehicle type	−3.645	0.136	0.5	<0.05**	−1.005

Note. [a]is standard error. [b]is standard deviation. **indicates 95% confidence level is used (i.e., P value <0.05 is statistically significant).

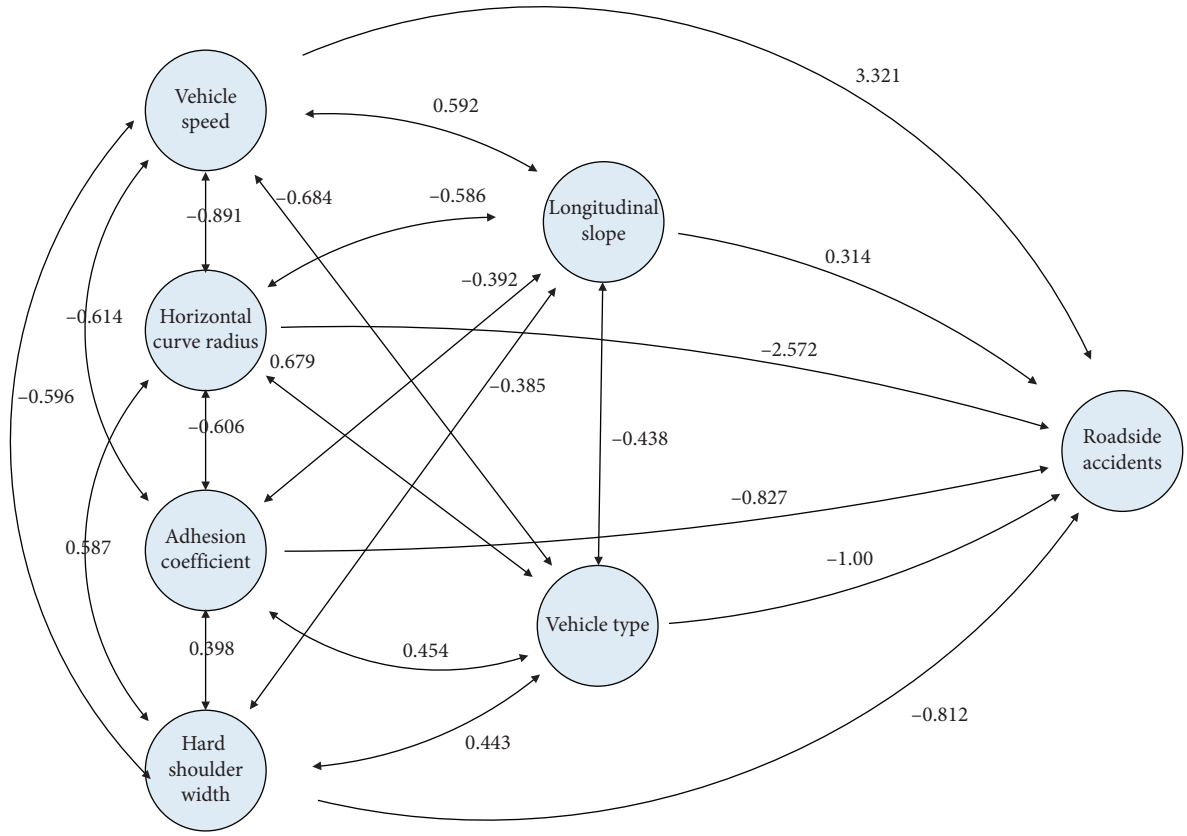


Figure 4: Structure diagram of path analysis.

and indicates that apart from direct effects, all risk factors had indirect effects on roadside accidents through other factors based on these correlations in the model. Among these interactions, the combination of vehicle speed-horizontal curve radius had the largest impact of interaction (−0.891) on roadside accident occurrence, and the negative interaction indicated that there was a mutually restricted relationship between these two factors. There were also other large interactions involved in Figure 4, including vehicle speed-vehicle type (−0.684), horizontal curve radius-vehicle type (0.679), vehicle speed-adhesion coefficient (−0.614), horizontal curve radius-adhesion coefficient (−0.606), and vehicle speed-hard shoulder width (−0.596).

Table 5 mainly shows the indirect effect of each risk factor through another mediating factor and the overall effect on roadside accident occurrence. It can be observed that vehicle speed transmitted its largest indirect effect (2.292) on roadside accidents through the horizontal curve radius than other factors, while horizontal curve radius had the largest indirect effect (−2.960) on accidents by vehicle speed. In addition, it was interesting to note that all other factors also had their largest and second indirect effects on roadside accidents by means of vehicle speed and horizontal curve radius. These results emphasize that vehicle speed and horizontal curve radius are still the most significant risk factors causing roadside accidents.

According to the overall effect of each risk factor on roadside accidents shown in Table 5, the most important risk factors were in decreasing order of importance, vehicle speed (7.749), horizontal curve radius (−7.644), vehicle type (−6.086), adhesion coefficient (−5.496), hard shoulder width (−5.373), and longitudinal slope (2.607). It is significant to

Table 5: The magnitude of the impact of factors on roadside accidents.

Variable	Direct effect	Indirect effect						
		Vehicle speed	Horizontal curve radius	Adhesion coefficient	Hard shoulder width	Longitudinal slope	Vehicle type	Overall effect
Vehicle speed	3.321	—	2.292	0.508	0.484	0.461	0.684	7.749
Horizontal curve radius	−2.572	−2.960	—	−0.501	−0.477	−0.456	−0.679	−7.644
Adhesion coefficient	−0.827	−2.040	−1.559	—	−0.323	−0.305	−0.443	−5.496
Hard shoulder width	−0.812	−1.979	−1.510	−0.329	—	−0.300	−0.443	−5.373
Longitudinal slope	0.314	1.066	1.007	0.124	0.013	—	0.083	2.607
Vehicle type	−1.005	−2.272	−1.746	−0.366	−0.360	−0.341	—	−6.086

note that the order of importance of these risk factors was not changed by the overall effects compared to the direct effects. This finding indicates that the indirect effects of different factors are not expected to play an important role in the occurrence of roadside accidents.

3.3. Probability of Roadside Accidents. Given that Bayesian network performs best with a small set of variables [62] and the least impact was longitudinal slope on roadside accidents compared to other important factors, we input the first five important factors (i.e., vehicle speed, horizontal curve radius, vehicle type, adhesion coefficient, and hard shoulder width) into a Bayesian network analysis to establish the probability prediction model for roadside accidents.

In the present study, the Bayesian network structure was developed based on the results of path analysis, and the Bayesian network parameter learning of roadside accidents was performed using the GD algorithm in Netica software, in which the prior and conditional probability distribution of each node could be obtained. In addition, according to the sensitivity analysis (see Table 6), the order of nodes (variables) based on the magnitude of mutual information (impact on roadside accidents) was consistent with the order obtained from path analysis, indicating that an accurate Bayesian network model used to predict the probability of roadside accidents was built (see Figure 5).

The probability of roadside accidents (i.e., posterior probability) under different combinations of variables can be obtained in this prediction model. For instance, assuming that a road section was a dry asphalt pavement with a speed limit of 80 km/h, a horizontal curve radius of 235 m, and a hard shoulder width of 0.75 m, then the probability of roadside accidents for truck passing through above road section need be predicted. First, the state of $60\,\text{km/h} < V \le 80\,\text{km/h}$, $200\,\text{m} < R \le 300\,\text{m}$, $0.6 \le \mu \le 0.8$, $w \le 0.75\,\text{m}$, and vehicle type of 0 were as 100%, and after automatically updating the probabilities of the whole network, the calculated probability of roadside accidents for truck driving at a speed of $60\,\text{km/h} < V \le 80\,\text{km/h}$ was 38.7% (see Figure 6).

Furthermore, the developed prediction model can also predict probabilities under the effects of any number (from 1 to 5) of factors (i.e., in the absence of some factors). For example, given that the speed limit of a road section was 80 km/h and the width of hard shoulder was 0.75 m, but lack of other indicators, and it could also be calculated that the probability of roadside accidents for car with a speed of $60\,\text{km/h} < V \le 80\,\text{km/h}$ was 18.2% (see Figure 7(a)).

Table 6: Sensitivity analysis result of the node "roadside accident."

Node	Mutual info	Percent	Variance in beliefs
Roadside accident	0.84054	100	0.1968046
Vehicle speed	0.13101	15.6	0.0358728
Horizontal curve radius	0.07927	9.43	0.0220364
Vehicle type	0.01063	1.27	0.0028862
Adhesion coefficient	0.00959	1.14	0.0026972
Hard shoulder width	0.00731	0.87	0.0020842

For another example, assume a road section was a dry asphalt pavement and horizontal curve radius and hard shoulder width were unknown. If the speed limit of this section was 80 km/h, the probabilities of roadside accident were 3.52% for car and 14.9% for truck (see Figures 7(b) and 7(c)), whereas the same probabilities were 14.3% for car and 44.5% for truck if the speed limit was 100 km/h (see Figures 7(d) and 7(e)), which further indicates that trucks have a higher risk of rollover than cars, especially when the vehicle speed was great than 80 km/h. Of course, the more factors involved, the more precise the obtained probability.

It is important to note that when various variables were in extreme states tending to avoid roadside accidents, even if vehicle speed was set as 120 km/h, whether for car or truck, and the probability of roadside accidents was, not as expected, only 1.31% (see Figure 7(f)), which adequately illustrates the importance of reasonable road design in situations where the driver's behaviour cannot be controlled. Therefore, in the purpose of further improving roadside safety and identifying the road conditions in which roadside accidents occur frequently, a variety of thresholds of horizontal curve radius, adhesion coefficient, and hard shoulder width corresponding to different vehicle speeds and vehicle types are given based on the Bayesian network prediction model, as shown in Table 7.

3.4. Identification of Roadside Accident Blackspot. We considered that there was a high frequency of roadside accidents (i.e., accident blackspot) when the probability of roadside

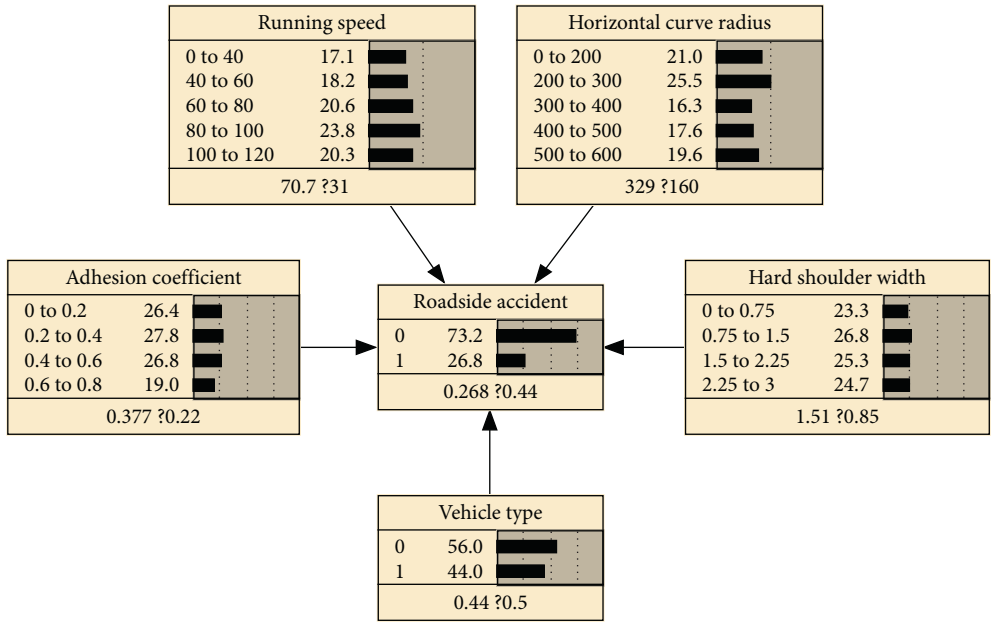


Figure 5: Bayesian network prediction model of roadside accidents. Note: 0 represents "truck" and 1 represents "car" in vehicle type; 0 denotes "no roadside accident occurrence" and 1 denotes "roadside accidents occurrence" in roadside accidents.

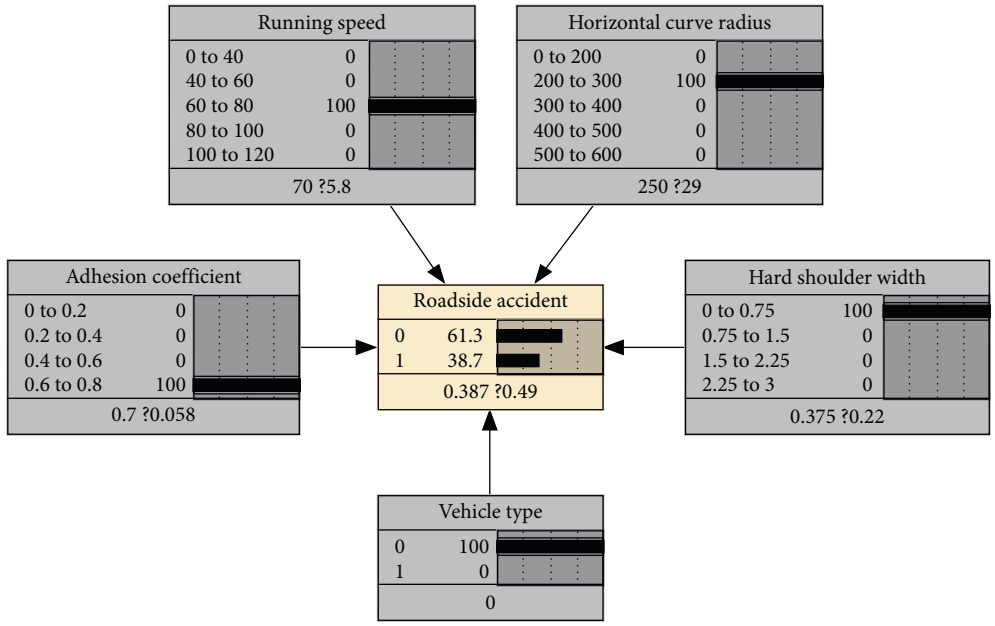
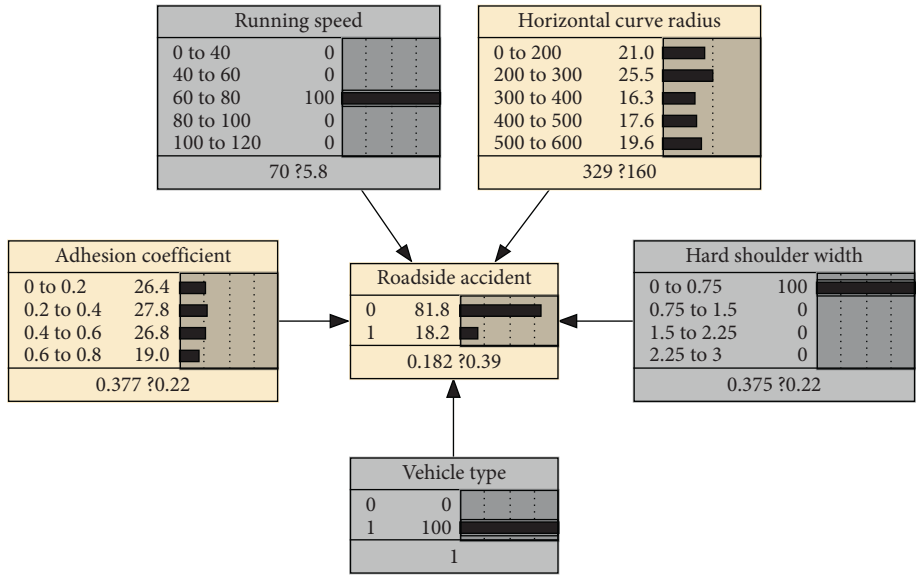


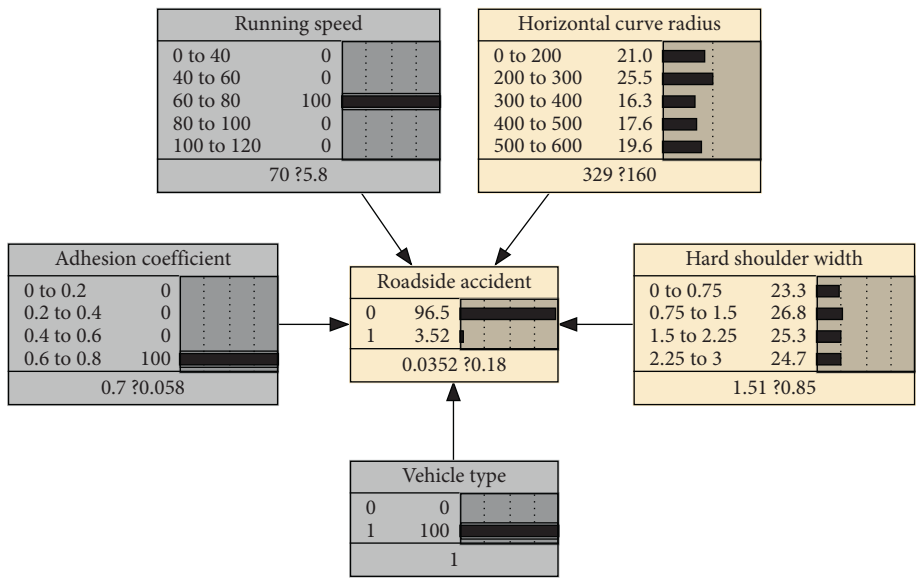
Figure 6: Probability calculation of roadside accidents.

accidents occurrence was greater than that of no roadside accidents occurrence (i.e., the probability of roadside accidents was greater than 50%). According to the results from Table 7, a range of vehicle speeds corresponds to 1 to 4 identification rules for roadside accident blackspots. When the value of each risk factor from a certain road section meets any of these 18 identification rules, this road section is then judged to be the road section with frequent occurrences of roadside accidents.

In this paper, a section (K2639 + 498.02 to K2679 + 170) from G105 was selected to confirm the effectiveness of the proposed method of identification. The G105 is a first-class road with a design speed of 80 km/h. By collecting road design documents and data of annual operating speed, the location of K2669 + 256.378 is determined to be the road section with frequent accidents according to the risk factor threshold, as shown in Table 7. The horizontal curve radius of this location is 280 m, the width of the hard shoulder is

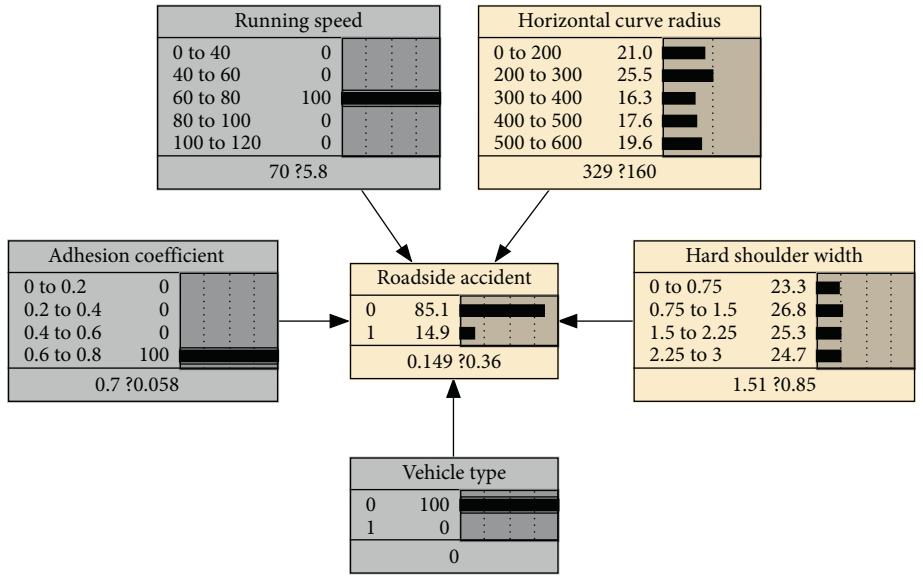


(a)

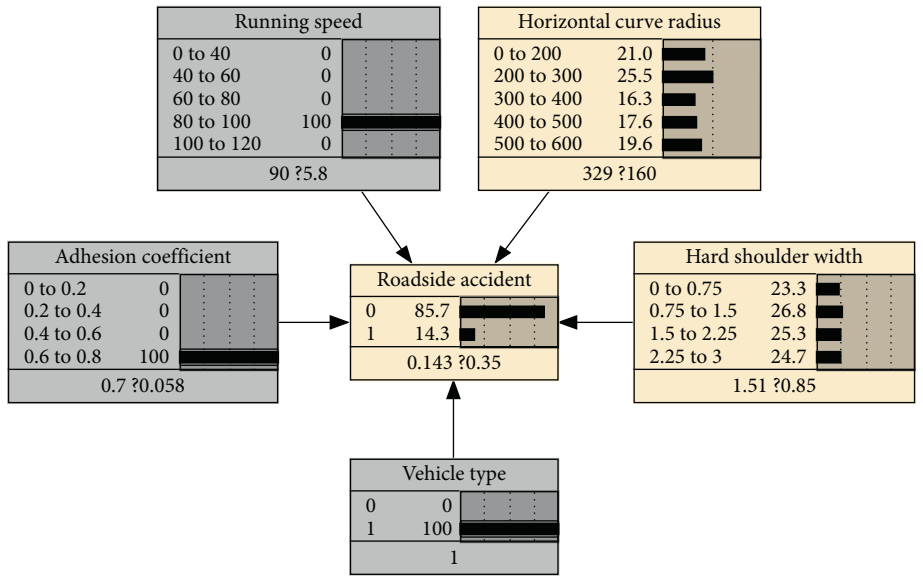


(b)

Figure 7: Continued.

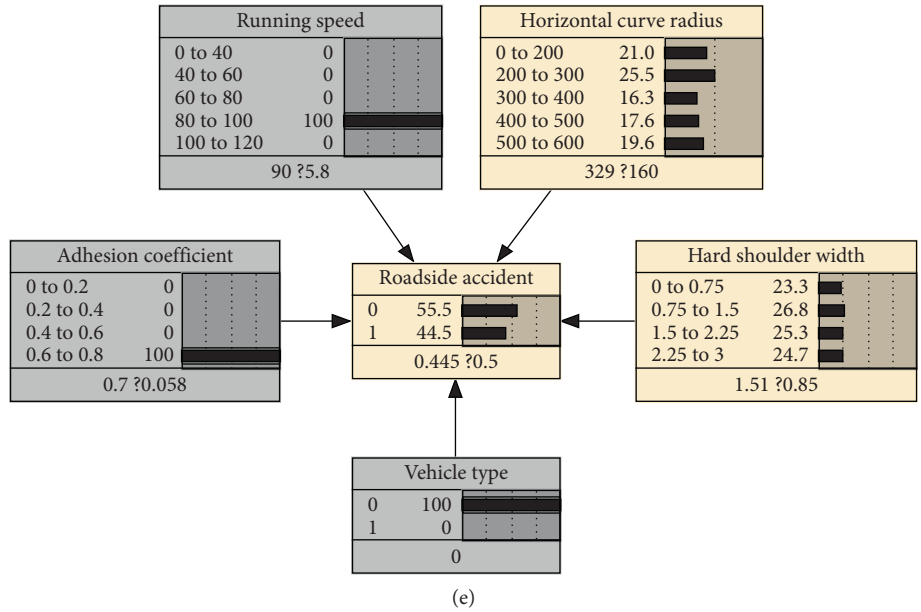


(c)

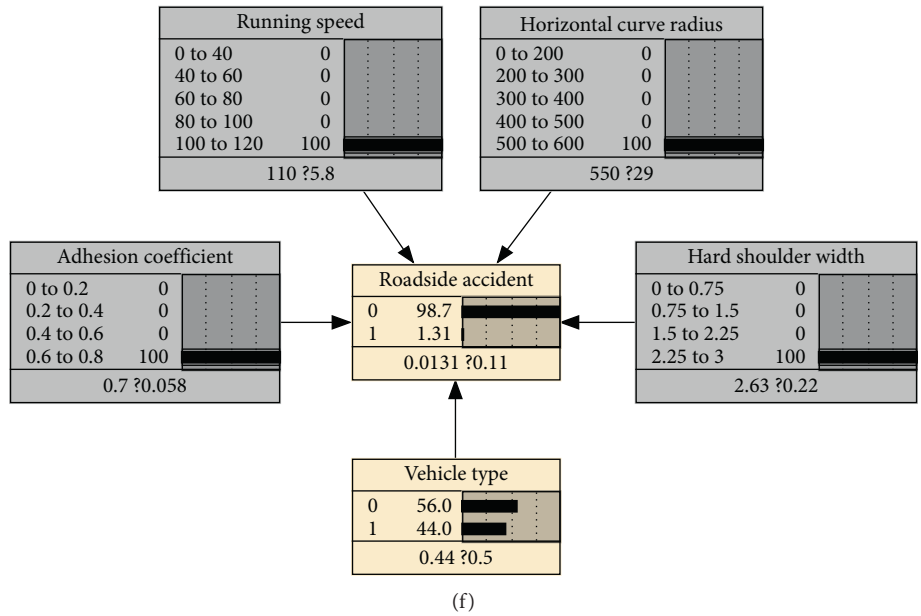


(d)

Figure 7: Continued.



(e)



(f)

Figure 7: Probability calculation of roadside accidents. (a) Result 1. (b) Result 2. (c) Result 3. (d) Result 4. (e) Result 5. (f) Result 6.

1.5 m, the operating speed of cars are mainly distributed in 120 km/h > $V \geq$ 100 km/h (see Figure 8(a)), and that of trucks are mainly distributed in 100 km/h > $V \geq$ 80 km/h (Figure 8(b)). The above indicators, respectively, conform to the 7 and 17 identification rules. According to the traffic police department's accident records, there were more than 80 roadside accidents in the above section from 2014 to 2018, which has been classified as the roadside accident-prone road. Based on the above analysis, the reliability of the proposed identification method for roadside accident blackspot in this paper is, therefore, verified.

The importance of such a study lies in the fact that it can help authorities identify significant risk factors that result in frequent roadside accidents in small curve segments to implement effective countermeasures or optimize alignment design in the process of future road construction and reconstruction. For instance, most of the thresholds for trucks were larger than those for cars at the same vehicle speed in Table 7, which suggests that the higher designed standard of geometric design and pavement condition is required for truck driving safety. Furthermore, for curve sections with truck speeds of no less than 60 km/h or car speeds of no less than 80 km/h, some thresholds of adhesion coefficients had almost reached the maximum 0.8. Therefore, we can reduce the risk of roadside accidents by optimizing other factors according to their respective thresholds.

Table 7: Threshold of significant factors leading to frequent roadside accidents.

No.	Vehicle speed (km/h)	Vehicle type	Horizontal curve radius (m)	Hard shoulder width (m)	Adhesion coefficient	Probability (>50%)
1	$80 \geq V > 60$	Truck	$300 < R \leq 400$	$w \leq 0.75$	$\mu \leq 0.2$	≥55.9
2			$200 < R \leq 300$	$w \leq 0.75$	$\mu \leq 0.6$	≥54.3
3			$R \leq 200$	$w \leq 1.50$	$\mu \leq 0.8$	≥54.9
4	$100 \geq V > 80$		$500 < R \leq 600$	$w \leq 0.75$	$\mu \leq 0.2$	≥54.2
5			$400 < R \leq 500$	$w \leq 0.75$	$\mu \leq 0.8$	≥55.0
6			$300 < R \leq 400$	$w \leq 1.50$	$\mu \leq 0.8$	≥54.8
7			$R \leq 300$	$w \leq 2.25$	$\mu \leq 0.8$	≥55.3
8	$120 \geq V > 100$		$500 < R \leq 600$	$1.50 < w \leq 2.25$	$\mu \leq 0.2$	≥55.3
9				$w \leq 1.50$	$\mu \leq 0.8$	≥54.3
10			$R \leq 500$	$w \leq 2.25$	$\mu \leq 0.8$	≥54.2
11	$80 \geq V > 60$	Car	$R \leq 300$	$w \leq 0.75$	$\mu \leq 0.2$	≥54.2
12	$100 \geq V > 80$		$300 < R \leq 400$	$w \leq 1.50$	$\mu \leq 0.2$	≥55.3
13			$200 < R \leq 300$	$w \leq 1.50$	$\mu \leq 0.4$	≥55.3
14			$R \leq 200$	$w \leq 2.25$	$\mu \leq 0.8$	≥54.2
15	$120 \geq V > 100$		$300 < R \leq 400$	$w \leq 2.25$	$\mu \leq 0.4$	≥54.5
16			$200 < R \leq 300$	$w \leq 1.50$	$0.4 < \mu \leq 0.6$	≥54.6
17				$w \leq 2.25$	$\mu \leq 0.8$	≥55.9
18			$R \leq 200$	$w \leq 3.00$	$\mu \leq 0.8$	≥54.3

Note. V represents the vehicle speed, R represents the horizontal curve radius, μ denotes the adhesion coefficient, and w is the hard shoulder width.

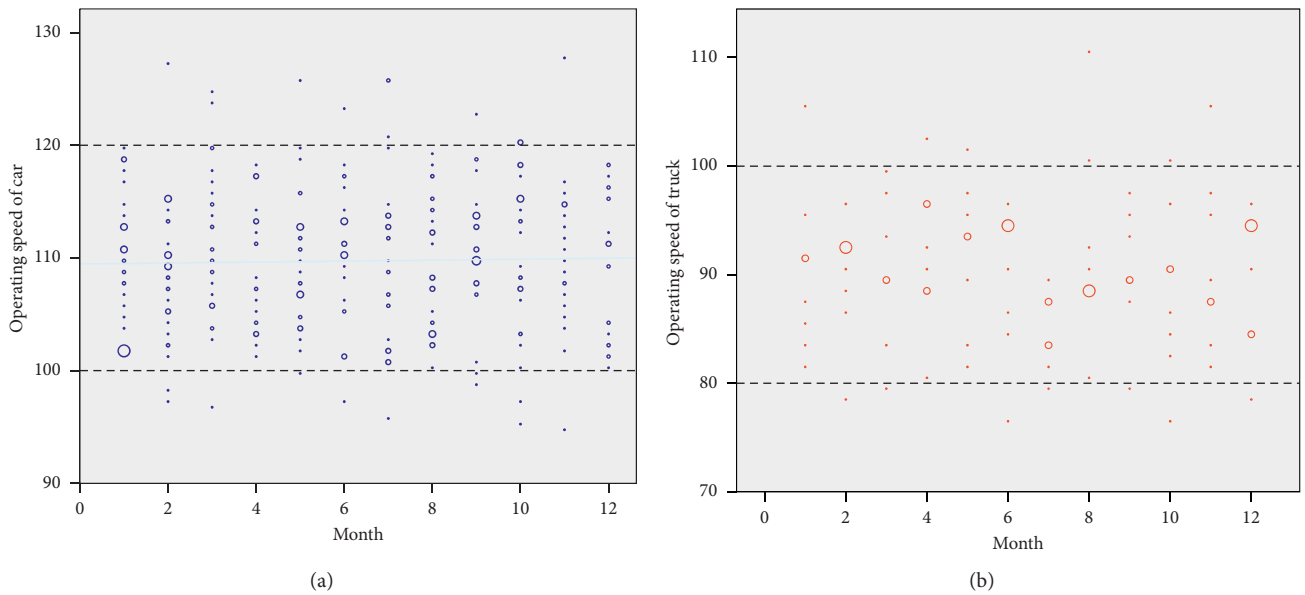


Figure 8: Operating speed distributions. (a) Car. (b) Truck.

4. Conclusions and Recommendations

The issue of roadside safety is crucial, especially for curve sections. In the present study, we employed CHAID decision tree analysis to identify significant risk factors resulting in the occurrence of roadside accidents, explored the impact of different combinations of risk factors on roadside accidents, and then used path analysis to determine the importance of these significant risk factors by investigating their direct and indirect effects on roadside accident occurrence. According to the results of the CHAID technique and path analysis, the significant predictors were in decreasing order of importance, vehicle speed, horizontal curve radius, vehicle type, adhesion coefficient, hard shoulder width, and longitudinal slope. The first five important factors were included as predictors of the probability of roadside accidents in the Bayesian network analysis to establish the probability prediction model of roadside accidents. Based on the results of probabilities of roadside accidents, the thresholds of horizontal curve radius, adhesion coefficient, and hard shoulder width corresponding to different vehicle speeds and vehicle types for accident blackspot identification in curve section were given.

These findings contribute to improving roadside safety in curve sections with a small radius. For instance, we confirmed again that vehicle speed and horizontal curve

radius are still the most critical factors leading to roadside accidents, whether in this study or other previous literature [63–66], and road sections with a high running speed and small radius are usually regarded as accident blackspot areas. Furthermore, based on the results of CHAID analysis, some specific recommended countermeasures as a supplement or reference for the revision of the DSHA of China were proposed as follows:

(i) For the highway with an operating speed of 60 km/h and a horizontal curve radius ≤200 m or an operating speed of 80 km/h and a horizontal curve radius ≤300 m, antislip measures should be strengthened

(ii) For the highway with an operating speed of 100 km/h and a horizontal curve radius of 300 m < R ≤ 600 m, the width of hard shoulder should be set as ≥1.5 m

(iii) For the freeway with an operating speed of 120 km/h and a horizontal curve radius of 300 m < R ≤ 600 m, the width of hard shoulder should be set as ≥2.25 m to ensure driving safety of trucks

Another important findings is that compared with cars, the width of the hard shoulder has a more significant influence on roadside accidents involving trucks, and trucks are more likely to have roadside accidents, especially in case of the vehicle speed >80 km/h. To ensure truck driving safety, the design standards of the horizontal curve radius, adhesion coefficient, and hard shoulder width should be further improved by decision makers in future highways construction. Additionally, limiting the load and running speed can be the most effective measures to mitigate the risk resulting from a higher centre of gravity. In recent years, a real-time monitoring system transmitting warning messages to truck drivers in cases of overload or overspeed has been designed by combining embedded technology and GPRS technology [67]. This system is expected to perform well in reducing truck roadside accidents. Another countermeasure is that regular maintenance of the truck, in case the brake failed in an emergency, also contributes to a decrease of accident rate [68].

The most remarkable result in this paper is that the developed Bayesian network prediction model can achieve the quantitative analysis of the probability of roadside accidents under the effects of any number (from 1 to 5) of factors. The resulting threshold of factors leading to accident blackspot can be a guide for authorities to identify and check roadside accidents prone areas located in small curve sections. In fact, if there are obstacles to promoting safe design standards for the horizontal curve radius, the adhesion coefficient, and the hard shoulder width due to high construction cost or unrealistic issues, many other effective countermeasures, such as setting deceleration strips in the pavement or related warning signs to control running speeds, widening the road in curve sections to provide a fault-tolerant space for drivers [69], and removing roadside hazards to reduce the loss of run-off-road accidents [70], could also be implemented.

Despite these promising results, some limitations exist in this paper. For example, this paper mainly predicts the roadside accident probability for two-way two-lanes or outer lanes of more than two lanes. Therefore, it remains to be further studied whether the prediction model is applicable to other road types (e.g., inner lanes of more than two lanes). In future studies, given the important impact of vehicle speed on roadside accidents, the limitation of maximum safe speed corresponding to different road geometric designs will be an additional research direction.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (no. 2018YFB1600902), MOE Layout Foundation of Humanities and Social Sciences (no. 18YJAZH009), National Natural Science Foundation of China (no. 51778063), and Fundamental Research Funds for the Central Universities (no. 2572019AB26).

References

[1] National Highway Traffic Safety Administration (NHTSA), *Fatality Analysis Reporting System*, National Highway Traffic Safety Administration (NHTSA), Washington, DC, USA, 2018.

[2] Traffic Management Bureau of the Ministry of Public Security, *Annual Road Traffic Accident Statistics*, China Communications Press, Beijing, China, 2019.

[3] S. B. McLaughlin, J. M. Hankey, S. G. Klauer, and T. A. Dingus, *Contributing Factors to Run-Off-Road Crashes and Near-Crashes*, National Highway Traffic Safety Administration (NHTSA), Washington, DC, USA, 2009, http://hdl.handle.net/10919/55073.

[4] C. Roque and M. Jalayer, "Improving roadside design policies for safety enhancement using hazard-based duration modeling," *Accident Analysis & Prevention*, vol. 120, pp. 165–173, 2018.

[5] D. Lord, M. A. Brewer, K. Fitzpatrick et al., *Analysis of Roadway Departure Crashes on Two Lane Rural Roads in Texas*, Texas Transportation Institute, College Station, TX, USA, 2011.

[6] I. Cruzado and E. Donnell, "Models of vehicle operating speeds along two-lane rural highway transition zones: panel and multilevel modeling approaches," *Transportation Letters*, vol. 3, no. 4, pp. 265–278, 2011.

[7] C. Liu and T. J. Ye, "Run-off-road crashes: an on-scene perspective," 2011.

[8] C. Roque, F. Moura, and J. Lourenço Cardoso, "Detecting unforgiving roadside contributors through the severity analysis of ran-off-road crashes," *Accident Analysis & Prevention*, vol. 80, pp. 262–273, 2015.

[9] C. Liu and R. Subramanian, "Factors related to fatal single-vehicle run-off-road crashes," 2009.
[10] M. H. A. Hussein, T. Sayed, K. Ismail, and A. Van Espen, "Calibrating road design guides using risk-based reliability analysis," *Journal of Transportation Engineering*, vol. 140, no. 9, 2013.
[11] C. V. Zegeer and F. M. Council, "Safety relationships associated with cross-sectional roadway elements," *Transportation Research Record*, vol. 1512, 1995.
[12] J. E. Hummer, W. Rasdorf, D. J. Findley, C. V. Zegeer, and C. A. Sundstrom, "Curve collisions: road and collision characteristics and countermeasures," *Journal of Transportation Safety & Security*, vol. 2, no. 3, pp. 203–220, 2010.
[13] C. D. Fitzpatrick, C. P. Harrington, M. A. Knodler, and M. R. E. Romoser, "The influence of clear zone size and roadside vegetation on driver behavior," *Journal of Safety Research*, vol. 49, no. 6, pp. 91–97, 2014.
[14] C. Roque and J. L. Cardoso, "SAFESIDE: a computer-aided procedure for integrating benefits and costs in roadside safety intervention decision making," *Safety Science*, vol. 74, pp. 195–205, 2015.
[15] M. Jalayer and H. Zhou, "Evaluating the safety risk of roadside features for rural two-lane roads using reliability analysis," *Accident Analysis & Prevention*, vol. 93, pp. 101–112, 2016.
[16] G. Cheng, R. Cheng, S. Zhang, and X. Sun, "Risk evaluation method for highway roadside accidents," *Advances in Mechanical Engineering*, vol. 11, no. 1, Article ID 754323857, 2019.
[17] H. Li, F. Pang, H. Chen, and Y. Du, "Vibration analysis of functionally graded porous cylindrical shell with arbitrary boundary restraints by using a semi analytical method," *Composites Part B: Engineering*, vol. 164, pp. 249–264, 2019.
[18] K. L. Stephan and S. V. Newstead, "Characteristics of the road and surrounding environment in metropolitan shopping strips: association with the frequency and severity of single-vehicle crashes," *Traffic Injury Prevention*, vol. 15, pp. S74–S80, 2014.
[19] R. B. Noland, "Traffic fatalities and injuries: the effect of changes in infrastructure and other trends," *Accident Analysis & Prevention*, vol. 35, no. 4, pp. 599–611, 2003.
[20] C. V. Zegeer and J. A. Deacon, "Effect of lane width, shoulder width, and shoulder type on highway safety," *State of the Art Report*, vol. 6, pp. 1–21, 1987.
[21] C. Zegeer, R. Stewart, D. Reinfurt et al., "Cost effefctive geometric improvements for safety upgrading of horizontal curves," Final Report, vol. 6, Federal Highway Administration, McLean, VA, USA, 1990. https://trid.trb.org/view/386551.
[22] S. Miaou, *Development of Relationship between Truck Accidents and Geometric Design: Phase I*, US Department of Transportation, Federal Highway Administration, McLean, Virginia, USA, 1993.
[23] S. C. Joshua and N. J. Garber, "Estimating truck accident rate and involvements using linear and Poisson regression models," *Transportation Planning and Technology*, vol. 15, no. 1, pp. 41–58, 1990.
[24] P. P. Jovanis and H. Chang, "Modeling the relationship of accidents to miles traveled," *Transportation Research Record*, vol. 1068, pp. 42–51, 1986.
[25] S.-P. Miaou, "Estimating vehicle roadside encroachment frequencies by using accident prediction models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1599, no. 1, pp. 64–71, 1997.
[26] R. Rusli, M. M. Haque, M. King, and W. S. Voon, "Single-vehicle crashes along rural mountainous highways in Malaysia: an application of random parameters negative binomial model," *Accident Analysis & Prevention*, vol. 102, pp. 153–164, 2017.
[27] J. Milton and F. Mannering, "The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies," *Transportation*, vol. 25, no. 4, pp. 395–413, 1998.
[28] M. Hosseinpour, A. S. Yahaya, A. F. Sadullah, N. Ismail, and S. M. R. Ghadiri, "Evaluating the effects of road geometry, environment, and traffic volume on rollover crashes," *Transport*, vol. 31, no. 2, pp. 221–232, 2016.
[29] X. Jiang, X. Yan, B. Huang, and S. H. Richards, "Influence of curbs on traffic crash frequency on high-speed roadways," *Traffic Injury Prevention*, vol. 12, no. 4, pp. 412–421, 2011.
[30] F. T. Kibar, F. Celik, and B. P. Aytac, "Statistical analysis of truck accidents for divided multilane interurban roads in Turkey," *KSCE Journal of Civil Engineering*, vol. 22, no. 5, pp. 1927–1936, 2018.
[31] V. Shankar, J. Milton, and F. Mannering, "Modeling accident frequencies as zero-altered probability processes: an empirical inquiry," *Accident Analysis & Prevention*, vol. 29, no. 6, pp. 829–837, 1997.
[32] H. Li, F. Pang, and H. Chen, "A semi-analytical approach to analyze vibration characteristics of uniform and stepped annular-spherical shells with general boundary conditions," *European Journal of Mechanics-A/Solids*, vol. 74, pp. 48–65, 2019.
[33] J. Lee and F. Mannering, "Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis," *Accident Analysis & Prevention*, vol. 34, no. 2, pp. 149–161, 2002.
[34] Y. Peng, R. Li, G. B. Li, X. M. Yang, and D. Zhou, "Method for investigation of child occupant impact dynamics based on real-world accident," *International Journal of Automotive Technology*, vol. 16, no. 5, pp. 791–797, 2015.
[35] J. Mandelík and M. Bundzel, "Application of neural network in order to recognise individuality of course of vehicle and pedestrian body contacts during accidents," *International Journal of Crashworthiness*, vol. 24, no. 2, pp. 221–234, 2019.
[36] M. Gopal, K. Baron, and M. Shah, *Simulation and Testing of a Suite of Field Relevant Rollovers*, SAE Technical Paper, Pennsylvania, MI, USA, 2004.
[37] H. Steffan and A. Moser, *How to Use PC-CRASH to Simulate Rollover Crashes*, SAE Technical Paper, Pennsylvania, MI, USA, 2004.
[38] N. A. Rose and G. Beauchamp, *Analysis of a Dolly Rollover with PC-Crash*, SAE Technical Paper, Pennsylvania, MI, USA, 2009.
[39] C. L. Naing, J. Hill, R. Thomson et al., "Single-vehicle collisions in Europe: analysis using real-world and crash-test data," *International Journal of Crashworthiness*, vol. 13, no. 2, pp. 219–229, 2008.
[40] R. Ootani and C. Pal, "Effective numerical simulation tool for real-world rollover accidents by combining PC-crash and FEA," SAE Technical Paper, Pennsylvania, MI, USA, 2007.
[41] Ministry of transport of the People's Republic of China (MOTO), *Design Specification for Highway Alignment*, Ministry of transport of the People's Republic of China (MOTO), Beijing, China, 2017.
[42] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random

forests," *Psychological Methods*, vol. 14, no. 4, pp. 323–348, 2009.

[43] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[44] G. Ritschard, *CHAID and Earlier Supervised Tree Methods*, pp. 70–96, Routledge, Abingdon, UK, 2011.

[45] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

[46] A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, Guilford Publications, New York, NY, USA, 2017.

[47] S. A. Gargoum and K. El-Basyouny, "Exploring the association between speed and safety: a path analysis approach," *Accident Analysis & Prevention*, vol. 93, pp. 32–40, 2016.

[48] C. Yu, *SPSS Statistical Analysis*, Publishing House of Electronics Industy, Beijing, China, 2007.

[49] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Elsevier, Amsterdam, Netherlands, 2014.

[50] A. Hadayeghi, A. Shalaby, and B. Persaud, "Development of planning-level transportation safety models using full Bayesian semiparametric additive techniques," *Journal of Transportation Safety & Security*, vol. 2, no. 1, pp. 45–68, 2010.

[51] A. Darwiche, *Modeling and Reasoning with Bayesian Networks*, Cambridge University Press, Cambridge, UK, 2009.

[52] L. Uusitalo, "Advantages and challenges of Bayesian networks in environmental modelling," *Ecological Modelling*, vol. 203, no. 3-4, pp. 312–318, 2007.

[53] J. Zhao and W. Deng, "The use of Bayesian network in analysis of urban intersection crashes in China," *Transport*, vol. 30, no. 4, pp. 411–420, 2015.

[54] M. Deublein, M. Schubert, and B. T. Adey, "Prediction of road accidents: comparison of two Bayesian methods," *Structure and Infrastructure Engineering*, vol. 10, no. 11, pp. 1394–1416, 2014.

[55] Q. Zeng, H. Wen, H. Huang, X. Pei, and S. C. Wong, "A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity," *Accident Analysis & Prevention*, vol. 99, pp. 184–191, 2017.

[56] S. H. Lee, Y. D. Lee, and M. Do, "Analysis on safety impact of red light cameras using the Empirical Bayesian approach," *Transportation Letters*, vol. 8, no. 5, pp. 241–249, 2016.

[57] X. Ma, Y. Xing, and J. Lu, "Causation analysis of hazardous material road transportation accidents by bayesian network using genie," *Journal of Advanced Transportation*, vol. 2018, 2018.

[58] C. Chen, X. Liu, H. Chen, M. Li, and L. Zhao, "A rear-end collision risk evaluation and control scheme using a Bayesian network model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 264–284, 2018.

[59] X. Zhu, Y. Yuan, X. Hu, Y.-C. Chiu, and Y.-L. Ma, "A Bayesian Network model for contextual versus non-contextual driving behavior assessment," *Transportation Research Part C: Emerging Technologies*, vol. 81, pp. 172–187, 2017.

[60] F. Zong, H. Xu, and H. Zhang, "Prediction for traffic accident severity: comparing the Bayesian network and regression models," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–9, 2013.

[61] G. Prati, L. Pietrantoni, and F. Fraboni, "Using data mining techniques to predict the severity of bicycle crashes," *Accident Analysis & Prevention*, vol. 101, pp. 44–54, 2017.

[62] S. H. Chen and C. A. Pollino, "Good practice in Bayesian network modelling," *Environmental Modelling & Software*, vol. 37, pp. 134–145, 2012.

[63] J. Xu, X. Luo, and Y. Shao, "Vehicle trajectory at curved sections of two-lane mountain roads: a field study under natural driving conditions," *European Transport Research Review*, vol. 10, no. 1, p. 12, 2018.

[64] H. Farah, A. van Beinum, and W. Daamen, "Empirical speed behavior on horizontal ramp curves in interchanges in The Netherlands," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2618, no. 1, pp. 38–47, 2017.

[65] S. Mavromatis, B. Psarianos, P. Tsekos, G. Kleioutis, and E. Katsanos, "Investigation of vehicle motion on sharp horizontal curves combined with steep longitudinal grades," *Transportation Letters*, vol. 8, no. 4, pp. 220–228, 2016.

[66] S. Alkheder, "Learning from the past: traffic safety in the eyes of affected local community in Abu Dhabi City, United Arab Emirates," *Transportation Letters*, vol. 9, no. 1, pp. 20–38, 2017.

[67] C. Dong, Q. Dong, B. Huang, W. Hu, and S. S. Nambisan, "Estimating factors contributing to frequency and severity of large truck--involved crashes," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 8, Article ID 4017032, 2017.

[68] B. L. Boada, M. J. L. Boada, M. Ramirez, and V. Diaz, "Study of van roadworthiness considering their maintenance and periodic inspection. The Spanish case," *Transportation Letters*, vol. 6, no. 4, pp. 173–184, 2014.

[69] American Association of State Highway and Transportation Official (AASHTO), *Roadside Design Guide*, American Association of State Highway and Transportation Official (AASHTO), Washington, DC, USA, 4th edition, 2011.

[70] J. M. Holdridge, V. N. Shankar, and G. F. Ulfarsson, "The crash severity impacts of fixed roadside objects," *Journal of Safety Research*, vol. 36, no. 2, pp. 139–147, 2005.