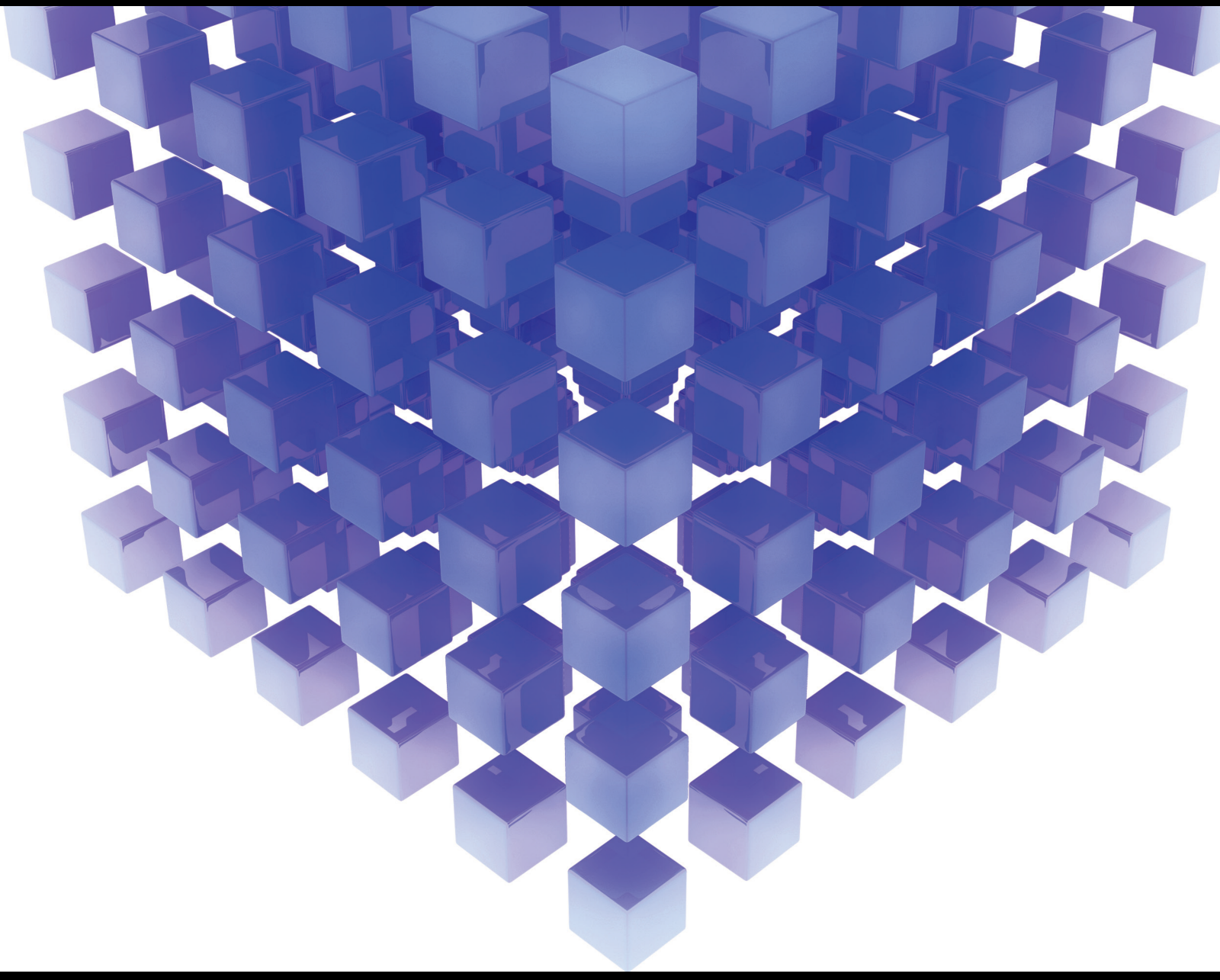
Mathematical Problems in Engineering

Safety Technologies and Fault Tolerant Methods for Engineering 2021

Lead Guest Editor: Yong Chen
Guest Editors: Mahdi Tavakoli, Darong Huang, Chuan Li, and Meng Li



Safety Technologies and Fault Tolerant Methods for Engineering 2021

Mathematical Problems in Engineering

Safety Technologies and Fault Tolerant Methods for Engineering 2021

Lead Guest Editor: Yong Chen
Guest Editors: Mahdi Tavakoli, Darong Huang, Chuan Li, and Meng Li

Chief Editor

Academic Editors

Contents

Compound Fault Diagnosis of Gearbox Based on RLMD and SSA-PNN

Shitong Liang and Jie Ma

Research Article (9 pages), Article ID 3716033, Volume 2021 (2021)

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7688702, 9 pages
https://doi.org/10.1155/2022/7688702

Research Article

On the FAULT-TOLERANT Resolvability in Line Graphs of Dragon and Kayak Paddles Graphs

Muhammad Faheem and Zohaib Zahid

University of Management and Technology (UMT), Lahore, Pakistan

Correspondence should be addressed to Zohaib Zahid; zohaib_zahid@hotmail.com

Received 30 September 2021; Revised 24 February 2022; Accepted 2 March 2022; Published 13 April 2022

Academic Editor: Yong Chen

Because of its wide range of applications, metric resolvability has been used in chemical structures, computer networks, and electrical circuits. It has been applied as a node (sensor) in an electric circuit. The electric circuit will not be able to flow current if one node (sensor) in that chain becomes faulty. The fault-tolerant selfstable circuit is a circuit that permits the current flow even if one of the nodes (sensors) becomes faulty. If the removal of any node from a resolving set (RS) of the circuit is still a RS, then the RS of the circuit is considered a fault-tolerant resolving set (FTRS) and the fault-tolerant metric dimension (FTMD) is its minimum cardinality. Even though the problem of finding the exact values of MD in line graphs seems to be even harder, the FTMD for the line graphs was first discussed by Guo et al. [13]. Ahmad et al. [5] determined the precise value of the MD for the line graph of the kayak paddle graph. We calculate the precise value of the FTMD for the line graph in this family of graphs. The FTMD is a more generalized invariant than the MD. We also consider the problem of obtaining a precise value for this parameter in the line graph of the dragon graph. It is concluded that these families have a constant FTMD.

1. Introduction and Preliminaries

Network topology is the graphical depiction of electric circuits. Because convoluted electric circuits (networks) are difficult to work on and study in their natural state, network topology is developed to make them simple and intelligible. Using this technique, any electric circuit (network) can be changed (moulded) into its corresponding graph; open circuits replace current sources, while short circuits replace passive parts and voltage sources. Short circuits are termed branches in network topology and edges in graph theory conceptualization, while open circuits are called nodes in network topology and vertices in pure mathematical graph theory. The following is the formal definition of a graphical depiction of an electric circuit (network).

Definition 1. Let $\lambda(V(\lambda), E(\lambda))$ be an electric circuit, where $V(\lambda)$ and $E(\lambda)$ are the sets of nodes (vertices) and branches (edges), respectively. The order of the electric circuit is $|V(\lambda)|$ and the size of the circuit is $|E(\lambda))|$.

Figure 1 demonstrates the example of the electric circuit and its equivalent graph.

Slater and Harary described the concept of resolving sets (RSs) independently in graphs [1–3], respectively. Metric basis have been used in robot navigation [4], chemical structures [5], and computed networks [6]. An electric circuit will stop working if one node of the circuit becomes faulty. Hernando et al. [7] established the new invariant FTRS to resolve such complications. If the removal of one node from the RS results in another RS, then the RS is called the FTRS. In this situation, a FTRS solves the problem by efficiently flowing the current in the circuit when one of the nodes stops working. The minimum cardinality of FTRS is known as the FTMD. Due to the generalized invariant, the FTMD produces more efficient results than MD. Due to this fact, researchers started to give attention to computing the exact values of the FTMD for different families of graphs.

Hernando et al. [7] discussed the invariant FTMD for the tree graphs and computed the upper bound $\beta'(G) \leq \beta(G)(1 + 2.5^{\beta(G)-1})$ for any graph G. Raza et al. [8],

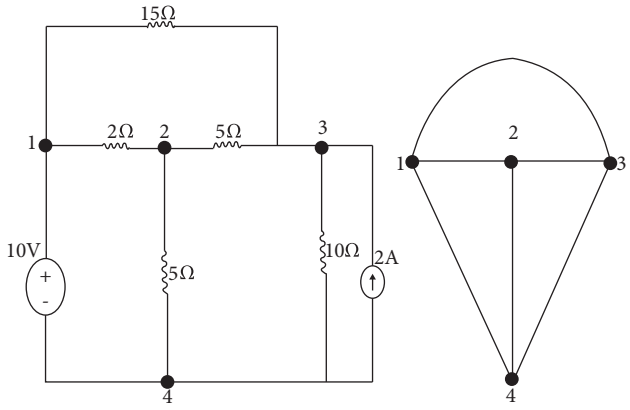


Figure 1: Electric circuit and its equivalent graph.

Zheng et al. [9], and Afzal et al. [10] applied the concept of the FTMD in some families of convex polytopes and computed their exact values. Basak et al. [11] calculated this parameter for the graph $C_n(1,2,3)$, and Saha et al. [12] generalized the results for the graph $C_n(1,2,3,4)$. Hayat et al. [13] and Prabhu et al. [14] applied this invariant to different computer networks and found their upper bounds. Somasundari et al. [15], Azeem et al. [16], Ahmad et al. [17], and Nadeem et al. [18] used the FTMD on different chemical structures and computed the exact values of this parameter. Laxman in [19] computed the lower bound of the FTMD for the cube of the path graph. Koam et al., in [20], calculated the MD and FTMD of the hollow coronoid chemical structure. Wang et al., in [21], considered the problem of finding the FTMD of three types of ladder graphs. Sharma and Bhat, in [22], calculated the FTMD of three families of the double antiprism graphs, which equals to 4. For more applications of the FTMD, see [23–25].

Voronov [26] and Raza et al. [27] determined some important upper bounds for the king's and extended Petersen graphs, respectively. Guo et al. [28] computed the FTMD for the line graphs of the families of necklace and prism graphs. Faheem et al. [29] calculated this invariant for the subdivision graphs of the same families of graphs. Simic et al. [30] and Saha et al. [31] determined the precise value of the FTMD for the grid and square of grid graphs, respectively. Ahmad et al. [32] calculated this parameter for $P(n,2)\odot K_1$ graph, which equals to 4. Hussain et al. [33] applied the idea of the FTMD to some families of gear graphs. For more applications and results about the FTMD in engineering, we refer [34–36].

The following are some essential terminologies and definitions that assist in calculating our primary results.

Definition 2. The degree $\mathrm{II}d_\lambda(\xi)\mathrm{II}$ of $\mathrm{II}\xi\mathrm{II}$ is the cardinality of branches that is incident to a node in $\xi \in V(\lambda)$.

Definition 3. The minimum cardinality of the branches, between $\xi_1 - \xi_2$ path, is known as the distance $d_\lambda(\xi_1,\xi_2)$ between $\xi_1,\xi_2 \in V(\lambda)$.

Definition 4. Let $\kappa = \{\xi_1,\xi_2,\ldots,\xi_t\} \subset V(\lambda)$; then, the absolute difference code have t -vector $(|d_\lambda(\mu_1,\xi_1) - d_\lambda(\mu_2,\xi_1)|,\ldots,|d_\lambda(\mu_1,\xi_t) - d_\lambda(\mu_2,\xi_t)|)$ for any $\mu_1,\mu_2 \in V(\lambda)$ with respect to κ, denoted by $A\ D((\mu_1,\mu_2)|\kappa)$.

Definition 5. Let $\kappa = \{\xi_1,\xi_2,\ldots,\xi_t\} \subset V(\lambda)$; then, the t-order distance code $r(\mu|\kappa)$ for a node $\mu \in V(\lambda)$ is $(d_\lambda(\mu,\xi_1), d_\lambda(\mu,\xi_2),\ldots,d_\lambda(\mu,\xi_t))$ with respect to κ. If the distance codes for every nodes of the circuit are unique, then the set κ is said to be a RS of the circuit λ. Moreover, if the absolute difference codes for any two nodes of the circuit have at least one nonzero with respect to κ, then κ is called the RS. The minimum cardinality of κ is called the MD, denoted by $\beta(\lambda)$.

Definition 6. Any RS $\kappa\prime$ of the circuit λ is known as the FTRS of the circuit if $\kappa' \smallsetminus \{\xi\}$ is again a RS of the circuit, where $\xi \in \kappa'$. Moreover, if the absolute difference codes for any two nodes of the circuit have at least two nonzeros with respect to $\kappa\prime$, then $\kappa\prime$ is called the FTRS. The minimum cardinality of κ' is called the FTMD, denoted by $\beta'(\lambda)$.

Definition 7. The line graph of the circuit λ is a new circuit $L(\lambda)$, whose nodes are the branches of λ and two branches v_1 and v_2 have a common end node in λ if and only if they are connected in $L(\lambda)$.

Following are some important bounds for $\beta'(\lambda)$ which are presented.

Lemma 1 (see [37]). *Let λ be any graph; then, $\beta(\lambda) < \beta'(\lambda)$.*

Lemma 2 (see [37]). *Let $\lambda \neq P_n$ be any graph; then, $\beta'(\lambda) \geq 3$.*

Lemma 3. *If the FTMD of any graph λ is 3 and $\{\xi_1,\xi_2,\xi_3\} \subset V(\lambda)$ is a FTRS in λ, then the degrees of the vertices ξ_1,ξ_2,ξ_3 are no more than 3.*

2. The Fault-Tolerant Resolvability of the Line Graph of Dragon Graph

Let C_n be a cycle with edge set $E(C_n) = \{e_1,e_2,\ldots,e_n\}$; also, let P_{m+1} be a path with edge set $E(P_{m+1}) = \{f_1,f_2,\ldots,f_m\}$. Dragon graph $T_{n,m}$ is shown in Figure 2.

To compute our required results, we convert the graph $T_{n,m}$ into their line graph $L(T_{n,m})$. The line graph $L(T_{n,m})$ of the dragon graph consists of a cycle of nodes $\{e_1,e_2,\ldots,e_n\}$ and the path of nodes $\{f_1,f_2,\ldots,f_m\}$, as shown in Figure 3.

The result of the MD of $L(T_{n,m})$ is presented below.

Theorem 1. *For any integers $m \geq 1$ and $n \geq 4$, we have $\beta(L(T_{n,m})) = 2$.*

Proof. For even integers $n \geq 4$, it can be easily verify that $\kappa = \{e_1, e_{n+4/2}\} \subset V(L(T_{n,m}))$ is the metric generator of $L(T_{n,m})$. For odd integers $n \geq 5$, it can also be verified that $\kappa = \{e_1, e_{n+1/2}\} \subset V(L(T_{n,m}))$ is the metric generator of $L(T_{n,m})$. So, $\beta(L(T_{n,m})) = 2$.

Now, we will compute the FTMD for $L(T_{n,m})$. □

Theorem 2. *For any integers $m \geq 1$ and $n \geq 4$, we have $\beta'(L(T_{n,m})) = 3$.*

Proof. To calculate our required results, the following are the cases. □

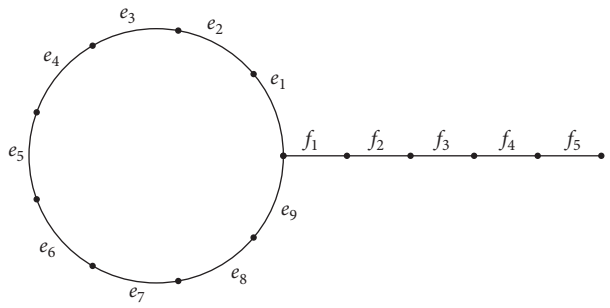


Figure 2: Dragon graph $T_{9,5}$.

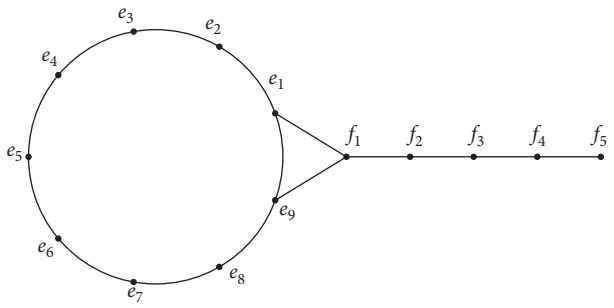


Figure 3: Line graph of dragon graph $T_{9,5}$.

Case 1. If n is odd.

Take $\kappa' = \{e_2, e_{n+1/2}, e_{n+3/2}\} \subset V(L(T_{n,m}))$ for odd integers $n \geq 5$. The distance codes of the nodes e_k, where $1 \leq k \leq n$, are

$$r\left(e_k|\kappa\prime\right) = \begin{cases} \left(1, \dfrac{-1+n}{2}, \dfrac{-1+n}{2}\right), & \text{if } k = 1, \\ \left(-2+k, \left|\dfrac{n+2k-1}{2}\right|, \dfrac{-2k+n+3}{2}\right), & \text{if } 2 \leq k \leq \dfrac{3+n}{2}, \\ \left(-k+n+2, \left|\dfrac{-n+2k-1}{2}\right|, \dfrac{-n+2k-3}{2}\right), & \text{if } \dfrac{5+n}{2} \leq k \leq n. \end{cases} \tag{1}$$

The distance codes for the nodes f_k are $r(f_k|\kappa') = (k+1, -1+n+2k/2, -3+n+2k/2)$, for $1 \leq k \leq m$.

From the above codes, we can conclude that the absolute difference codes for every pair of nodes have at least two nonzero in their 3-vector. So, $\beta'(L(T_{n,m})) \leq 3$. From Lemma 1 and Theorem 1, we have $\beta'(L(T_{n,m})) \geq 3$. Hence, $\beta'(L(T_{n,m})) = 3$.

Case 2. If n is even.

Take $\kappa' = \{e_1, e_{n+4/2}, f_1\} \subset V(L(T_{n,m}))$ for every even integers $n \geq 4$. The distance codes of the nodes e_k, where $1 \leq k \leq n$, are

$$r\left(e_k|\kappa'\right) = \begin{cases} \left(0, \dfrac{-2+n}{2}, 1\right), & \text{if } k = 1, \\ \left(-1+k, \left|\dfrac{-n+2k-4}{2}\right|, k\right), & \text{if } 2 \leq k \leq \dfrac{n}{2}, \\ \left(-k+n+1, \left|\dfrac{-n+2k-4}{2}\right|, -k+n+1\right), & \text{if } \dfrac{n+2}{2} \leq k \leq n. \end{cases} \tag{2}$$

The distance codes for the nodes f_k with respect to W' are $r(f_k|\kappa') = (k, -4+n+2k/2, -1+k)$, for $1 \leq k \leq m$.

From the above codes, we can conclude that the absolute difference codes for every pair of nodes have at least two nonzero in their 3-vector. This shows that $\beta'(L(T_{n,m})) \leq 3$, and from Lemma 2, $\beta'(L(T_{n,m})) \geq 3$. Hence, $\beta'(L(T_{n,m})) = 3$.

3. The Fault-Tolerant Resolvability of the Line Graph of Kayak Paddles Graph

Kayak paddles graph $KP(l, m, n)$ is a graph made up of two cycles C_l and C_m having size l and m joined by a path of length n. We label the branches of cycle C_l by $\{e_1, e_2, \ldots, e_l\}$, the branches of cycle C_m by $\{f_1, f_2, \ldots, f_m\}$, and the

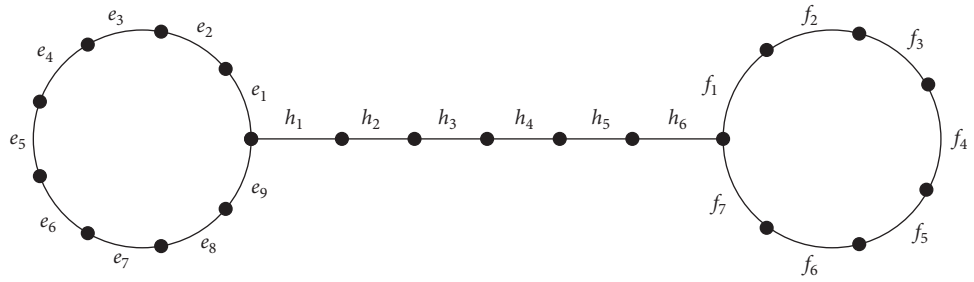


Figure 4: Kayak paddle graph $KP(9,7,6)$.

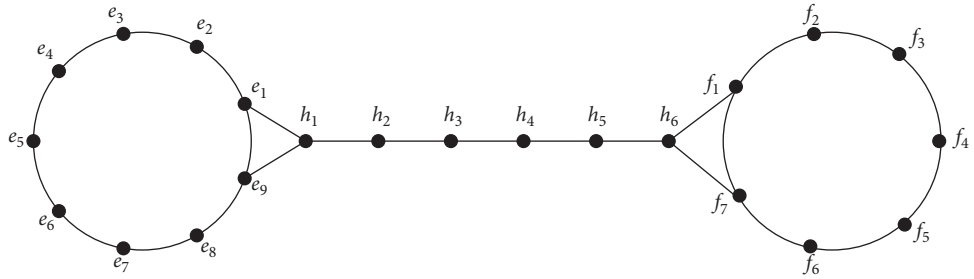


Figure 5: Line graph $L(KP(9,7,6))$.

branches of path joining these cycles by $\{h_1, h_2, \ldots, h_n\}$, as shown in Figure 4.

To compute our required results, we need to convert the graph $KP(l,m,n)$ into the graph $L(KP(l,m,n))$. The line graph of kayak paddles graph $L(KP(l,m,n))$ consists of cycle C_l with nodes $\{e_1, e_2, \ldots, e_l\}$, the cycle C_m with nodes $\{f_1, f_2, \ldots, f_m\}$, and the nodes of path joining these cycles $\{h_1, h_2, \ldots, h_n\}$, as shown in Figure 5.

The known result about $\beta(L(KP(l,m,n)))$ is presented below.

Theorem 3 (see [38]). *For any integers $n \geq 2$ and $l, m \geq 3$, we have $\beta(L(KP(l,m,n))) = 2$.*

Now, we will compute $\beta'(L(KP(l,m,n)))$.

Theorem 4. *For any integers $n \geq 2$ and $l, m \geq 3$, we have $\beta'(L(KP(l,m,n))) = 4$.*

Proof. To calculate our required results, the following are the cases. □

Case 3. If l and m both are even.

Take $\kappa' = \{e_1, e_2, f_1, f_2\} \subset V(L(KP(l,m,n)))$ for both even integers $l, m \geq 4$. The distance codes of the nodes e_k, where $1 \leq k \leq l$, are

$$r(e_k \mid \kappa') = \begin{cases} (-1+k, |-k+2|, k+n, 1+n+k), & \text{if } 1 \leq k \leq \dfrac{l}{2}, \\ \left(\dfrac{l}{2}, \dfrac{-2+l}{2}, \dfrac{l+2n}{2}, \dfrac{2+l+2n}{2}\right), & \text{if } k = \dfrac{2+l}{2}, \\ (-k+l+1, -k+l+2, -k+l+n+1, -k+2+l+n), & \text{if } \dfrac{4+l}{2} \leq k \leq l. \end{cases} \tag{3}$$

The distance codes for the nodes f_k, where $1 \leq k \leq m$, are

$$r(f_k \mid \kappa') = \begin{cases} (k+n, 1+n+k, -1+k, |-2+k|), & \text{if } 1 \le k \le \dfrac{m}{2}, \\ \left(\dfrac{m+2n}{2}, \dfrac{2+m+2n}{2}, \dfrac{m}{2}, \dfrac{-2+m}{2}\right), & \text{if } k = \dfrac{2+m}{2}, \\ (-k+1+m+n, -k+2+m+n, -k+m+1, -k+2+m), & \text{if } \dfrac{m+4}{2} \le k \le m. \end{cases} \tag{4}$$

The distance codes for the nodes h_k are $r(h_k \mid \kappa') = (k, 1+k, -k+1+n, -k+2+n)$, for $1 \le k \le n$.

From the above codes, we can conclude that the absolute difference codes for every pair of nodes have at least two nonzeros in their 4-vector. This shows that $\beta'(L(KP(l,m,n))) \le 4$, but, in Lemma 2, $\beta'(L(KP(l,m,n))) \ge 3$.

Now, to prove that $\beta'(L(KP(l,m,n))) \ge 4$, suppose contrary that $\beta'(L(KP(l,m,n))) = 3$, and according to Lemma 3, we have the following conditions:

(i) Let $\kappa\prime = \{e_i, e_j, e_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i < j < k \le l$; then, $A\,D((f_1, f_m)|\kappa\prime) = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

(ii) Let $\kappa\prime = \{e_i, f_j, f_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i \le l$ and $1 \le j < k \le m$; then, $A\,D((e_1, e_l)|\kappa\prime) = (1,0,0)$. So, $\kappa\prime$ is not FTRS.

(iii) Let $\kappa\prime = \{e_i, e_j, h_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i < j \le l$ and $1 \le k \le n$; then, $A\,D((f_1, f_m)|\kappa') = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

(iv) Let $\kappa' = \{e_i, h_j, h_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i \le l$ and $1 \le j < k \le n$; then, $A\,D((f_1, f_m)|\kappa') = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

(v) Let $\kappa'\{h_i, h_j, h_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i < j < k \le n$; then, $A\,D((e_1, e_l)|\kappa\prime) = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

From the above discussion, we conclude that there is no FTRS with cardinality 3. This shows that $\beta'(L(KP(l,m,n))) \ge 4$. Hence, $\beta'(L(KP(l,m,n))) = 4$.

Case 4. If l and m both are odd.

(1) Let $l = 3$, and for any odd integers $m \ge 3$, take $\kappa\prime = \{e_1, e_3, f_1, f_{m+3/2}\} \subset V(L(KP(3,m,n)))$. The distance codes for the nodes e_k, where $1 \le k \le 3$, are

$$r(e_k \mid \kappa') = \begin{cases} \left(-1+k, 1, k+n, \dfrac{-3+m+2k+2n}{2}\right), & \text{if } 1 \le k \le 2, \\ \left(1, 0, 1+n, \dfrac{-1+m+2n}{2}\right), & \text{if } k = 3. \end{cases} \tag{5}$$

The distance codes for the nodes f_k, where $1 \le k \le m$, are

$$r(f_k \mid \kappa') = \begin{cases} \left(1+n, 1+n, 0, \dfrac{-1+m}{2}\right), & \text{if } k = 1, \\ \left(k+n, n+k, -1+k, \dfrac{-2k+m+3}{2}\right), & \text{if } 2 \le k \le \dfrac{1+m}{2}, \\ \left(-k+1+m+n, -k+1+m+n, -k+m+1, \dfrac{-3-m+2k}{2}\right), & \text{if } \dfrac{3+m}{2} \le k \le m. \end{cases} \tag{6}$$

The distance codes for the nodes h_k are $r(h_k \mid \kappa') = (k, k, -k+n+1, -2k-1m+2n/2)$, for $1 \le k \le n$.

(2) Take $\kappa' = \{e_1, e_2, f_1, f_2\} \subset V(L(KP(l,m,n)))$, for both odd integers $l, m \ge 5$. The distance codes for the nodes e_k, where $1 \le k \le l$, are

$$r(e_k \mid \kappa') = \begin{cases} (-1+k, |-2+k|, n+k, 1+n+k), & \text{if } 1 \le k \le \dfrac{l+1}{2}, \\ \left(\dfrac{-1+l}{2}, \dfrac{-1+l}{2}, \dfrac{-1+l+2n}{2}, \dfrac{1+l+2n}{2}\right), & \text{if } k = \dfrac{3+l}{2}, \\ (-k+1+l, -k+l+2, -k+l+n+1, -k+2+l+n), & \text{if } \dfrac{5+l}{2} \le k \le l. \end{cases} \tag{7}$$

The distance codes for the nodes f_k, where $1 \le k \le m$, are

$$r(f_k \mid \kappa') = \begin{cases} (k+n, 1+n+k, -1+k, |-k+2|), & \text{if } 1 \le k \le \dfrac{1+m}{2}, \\ \left(\dfrac{-1+m+2n}{2}, \dfrac{1+m+2n}{2}, \dfrac{-1+m}{2}, \dfrac{-1+m}{2}\right), & \text{if } k = \dfrac{3+m}{2}, \\ (-k+1+m+n, -k+2+m+n, -k+1+m, -k+2+m), & \text{if } \dfrac{5+m}{2} \le k \le m. \end{cases} \tag{8}$$

The distance codes for the nodes h_k are $r(h_k|\kappa') = (k, 1+k, -k+1+n, -k+2+n)$, for $1 \le k \le n$.

From the above codes, we can conclude that the absolute difference codes for every pair of nodes have at least two nonzeros in their 4-vector. This shows that $\beta'(L(KP(l, m, n))) \le 4$, but in Lemma 2, $\beta'(L(KP(l, m, n))) \ge 3$.

Now, to prove that $\beta'(L(KP(l, m, n))) \ge 4$, suppose contrary that $\beta'(L(KP(l, m, n))) = 3$, and according to Lemma 3, we have the following conditions:

(i) Let $\kappa' = \{e_i, e_j, e_k\} \subset V(L(KP(l, m, n)))$, for $1 \le i < j < k \le l$; then, $A\,D((f_1, f_m)|\kappa') = (0, 0, 0)$. So, $\kappa\prime$ is not FTRS.

(ii) Let $\kappa' = \{e_i, f_j, f_k\} \subset V(L(KP(l, m, n)))$; for $1 \le i \le l$ and $1 \le j < k \le m$, then

$$A\,D\left((e_1, e_l)|\kappa\prime\right) = \begin{cases} (0, 0, 0), & \text{if } i = \dfrac{l+1}{2}, \\ (1, 0, 0), & \text{else.} \end{cases} \tag{9}$$

So, $\kappa\prime$ is not FTRS.

(iii) Let $\kappa' = \{e_i, e_j, h_k\} \subset V(L(KP(l, m, n)))$; for $1 \le i < j \le l$ and $1 \le k \le n$, then $A\,D((f_1, f_m)|\kappa\prime) = (0, 0, 0)$. So, $\kappa\prime$ is not FTRS.

(iv) Let $\kappa' = \{e_i, h_j, h_k\} \subset V(L(KP(l, m, n)))$; for $1 \le i \le l$ and $1 \le j < k \le n$, then $A\,D((f_1, f_m)|\kappa\prime) = (0, 0, 0)$. So, $\kappa\prime$ is not FTRS.

(v) Let $\kappa' = \{h_i, h_j, h_k\} \subset V(L(KP(l, m, n)))$; for $1 \le i < j < k \le n$, then $A\,D((e_1, e_l)|\kappa\prime) = (0, 0, 0)$. So, $\kappa\prime$ is not FTRS.

From the above discussion, we conclude that there is no FTRS with cardinality 3. This shows that $\beta'(L(KP(l, m, n))) \ge 4$. Hence, $\beta'(L(KP(l, m, n))) = 4$.

Case 5. If m is even and l is odd.

(i) Let $l = 3$, and for any even integers $m \ge 4$, take $\kappa\prime = \{e_1, e_3, f_1, f_2\} \subset V(L(KP(3, m, n)))$. The distance codes for the nodes e_k, where $1 \le k \le 3$, are

$$r\left(e_k|\kappa\prime\right) = \begin{cases} (-1+k, 1, k+n, 1+n+k), & \text{if } 1 \le k \le 2, \\ (1, 0, 1+n, 2+n), & \text{if } k = 3. \end{cases} \tag{10}$$

The distance codes for the nodes f_k, where $1 \le k \le m$, are

$$r\left(f_k|\kappa\prime\right) = \begin{cases} (n+k, k+n, -1+k, |-k+2|), & \text{if } 1 \le k \le \dfrac{m}{2}, \\ \left(\dfrac{m+2n}{2}, \dfrac{m+2n}{2}, \dfrac{m}{2}, \dfrac{-2+m}{2}\right), & \text{if } k = \dfrac{2+m}{2}, \\ (-k+1+m+n, -k+1+m+n, -k+m+1, -k+m+2), & \text{if } \dfrac{4+m}{2} \le k \le m. \end{cases} \tag{11}$$

The distance codes for the nodes h_k are
$r(h_k|\kappa') = (k, k, -k+n+1, -k+n+2)$, for $1 \le k \le n$.

(2) Take $\kappa' = \{e_1, e_2, f_1, f_2\} \subset V(L(KP(l,m,n)))$ for any odd $l \ge 5$ and even $m \ge 4$ integers. The distance codes for the nodes e_k, where $1 \le k \le l$, are

$$r\left(e_k|\kappa\prime\right) = \begin{cases} (-1+k, |-k+2|, k+n, 1+n+k), & \text{if } 1 \le k \le \dfrac{1+l}{2}, \\ \left(\dfrac{-1+l}{2}, \dfrac{-1+l}{2}, \dfrac{-1+l+2n}{2}, \dfrac{1+l+2n}{2}\right), & \text{if } k = \dfrac{l+3}{2}, \\ (-k+l+1, -k+l+2, -k+1+l+n, -k+2+l+n), & \text{if } \dfrac{l+5}{2} \le k \le l. \end{cases} \tag{12}$$

The distance codes for the nodes f_k, where $1 \le k \le m$, are

$$r(f_k|\kappa') = \begin{cases} (k+n, 1+n+k, -1+k, |-k+2|), & \text{if } 1 \le k \le \dfrac{m}{2}, \\ \left(\dfrac{m+2n}{2}, \dfrac{2+m+2n}{2}, \dfrac{m}{2}, \dfrac{-2+m}{2}\right), & \text{if } k = \dfrac{2+m}{2}, \\ (-k+1+m+n, -k+2+m+n, -k+1+m, -k+2+m), & \text{if } \dfrac{4+m}{2} \le k \le m. \end{cases} \tag{13}$$

The distance codes for the nodes h_W are
$r(h_k|W\prime) = (k, 1+k, -k+1+n, -k+2+n)$, for $1 \le k \le n$.

From the above codes, we can conclude that the absolute difference codes for every pair of nodes have at least two nonzeros in their 4-vector. This shows that $\beta'(L(KP(l,m,n))) \le 4$, but, in Lemma 2, $\beta'(L(KP(l,m,n))) \ge 3$.

Now, to prove $\beta'(L(KP(l,m,n))) \ge 4$, suppose contrary that $\beta'(L(KP(l,m,n))) = 3$, and according to Lemma 3, we have the following conditions:

(i) Let $\kappa\prime = \{e_i, e_j, e_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i < j < k \le l$; then, $A\,D((f_1, f_m)|\kappa\prime) = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

(ii) Let $\kappa\prime = \{e_i, f_j, f_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i \le l$ and $1 \le j < k \le m$; then,

$$A\,D\left((e_1, e_l)|\kappa\prime\right) = \begin{cases} (0,0,0), & \text{if } i = \dfrac{l+1}{2}, \\ (1,0,0), & \text{else.} \end{cases} \tag{14}$$

So, $\kappa\prime$ is not FTRS.

(iii) Let $\kappa' = \{e_i, e_j, h_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i < j \le l$ and $1 \le k \le n$; then, $A\,D((f_1, f_m)|\kappa\prime) = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

(iv) Let $\kappa\prime = \{e_i, h_j, h_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i \le l$ and $1 \le j < k \le n$; then, $A\,D((f_1, f_m)|\kappa\prime) = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

(v) Let $\kappa\prime = \{h_i, h_j, h_k\} \subset V(L(KP(l,m,n)))$, for $1 \le i < j < k \le n$; then, $A\,D((e_1, e_l)|\kappa\prime) = (0,0,0)$. So, $\kappa\prime$ is not FTRS.

From the above discussion, we conclude that there is no FTRS with cardinality 3. This shows that $\beta'(L(KP(l,m,n))) \ge 4$. Hence, $\beta'(L(KP(l,m,n))) = 4$.

4. Conclusion

We conclude that, in the case of the line graph for the dragon graph, the FTMD is exactly one more than its MD, and it exactly doubles the MD in the case of the line graph of kayak paddles graph [39–44].

Data Availability

No data were used to support the findings of the study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. J. Slater, "Leaves of trees," *Congressus Numerantium*, vol. 14, pp. 549–559, 1975.

[2] P. J. Slater, "Dominating and referece set in a graph," *Journal of Mathematical and Physical Sciences*, vol. 22, no. 4, pp. 445–455, 1988.

[3] F. Harary and R. A. Melter, "On the metric dimension of a graph," *Ars Combinatoria*, vol. 2, pp. 191–195, 1976.

[4] G. Chartrand, L. Eroh, M. A. Johnson, and O. R. Oellermann, "Resolvability in graphs and the metric dimension of a graph," *Discrete Applied Mathematics*, vol. 105, pp. 99–113, 2000.

[5] G. Chartrand and P. Zhang, "The theory and applications of resolvability in graphs," *A Survey. Congressus Numerantium*.vol. 160, pp. 47–68, 2003.

[6] H. Raza, S. Hayat, and X. F. Pan, "On the fault-tolerant metric dimension of certain interconnection networks," *Journal of Applied Mathematics and Computing*, vol. 60, no. 1, pp. 517–535, 2019.

[7] C. Hernando, M. Mora, P. J. Slater, and D. R. Wood, "fault-Tolerant metric dimension of graph," *Proc. Internat. Conf. Convexity in Discrete Structure, Ramanujan Math. Society Lecture Notes*, vol. 5, pp. 81–85, 2008.

[8] H. Raza, S. Hayat, and X. F. Pan, "On the fault-tolerant metric dimension of convex polytopes," *Applied Mathematics and Computation*, vol. 339, pp. 172–185, 2018.

[9] Z. B. Zheng, A. Ahmad, Z. Hussain et al., "fault-tolerant metric dimension of generalized wheels and convex polytopes," *Mathematical Problems in Engineering*, vol. 2020, Article ID 1216542, 8 pages, 2020.

[10] H. M. A. Siddiqui, S. Hayat, A. Khan, M. Imran, A. Razzaq, and J. B. Liu, "Resolvability and fault-tolerant resolvability structures of convex polytopes," *Theoretical Computer Science*, vol. 796, pp. 114–128, 2019.

[11] M. Basak, L. Saha, G. K. Das, and K. Tiwary, "Fault-tolerant metric dimension of circulant graphs Cn (1, 2, 3)," *Theoretical Computer Science*, vol. 817, pp. 66–79, 2020.

[12] L. Saha, R. Lama, K. Tiwary, K. C. Das, and Y. Shang, "fault-tolerant metric dimension of circulant graphs," *Mathematics*, vol. 10, no. 1, p. 124, 2022.

[13] S. Hayat, A. Khan, M. Y. H. Malik, M. Imran, and M. K. Siddiqui, "fault-tolerant metric dimension of interconnection networks," *IEEE Access*, vol. 8, Article ID 145435, 2020.

[14] S. Prabhu, V. Manimozhi, M. Arulperumjothi, and S. Klavar, "Twin vertices in fault-tolerant metric sets and fault-tolerant metric dimension of multistage interconnection networks," *Applied Mathematics and Computation*, vol. 420, Article ID 126897, 2022.

[15] M. Somasundari and F. S. Raj, "Fault-tolerant resolvability of oxide interconnections," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 2278–3075, 2019.

[16] M. Azeem and M. F. Nadeem, "Metric-based resolvability of polycyclic aromatic hydrocarbons," *The European Physical Journal Plus*, vol. 136, no. 4, pp. 1–14, 2021.

[17] A. Ahmad, A. N. A. Koam, M. H. F. Siddiqui, and M. Azeem, "Resolvability of the starphene structure and applications in electronics," *Ain Shams Engineering Journal*, vol. 13, no. 2, Article ID 101587, 2022.

[18] M. F. Nadeem, A. Shabbir, and M. Azeem, "On metric dimension and fault tolerant metric dimension of some chemical structures," *Polycyclic Aromatic Compounds*, vol. 1-13, 2021.

[19] L. Saha, "fault-tolerant metric dimension of cube of paths," in *Journal of Physics: Conference Series*vol. 1714, no. 1, IOP Publishing, Article ID 012029, 2021.

[20] A. N. Koam, A. Ahmad, M. E. Abdelhag, and M. Azeem, "Metric and fault-tolerant metric dimension of hollow coronoid," *IEEE Access*, vol. 9, 2021.

[21] H. Wang, M. Azeem, M. F. Nadeem, A. Ur-Rehman, and A. Aslam, "On fault-tolerant resolving sets of some families of ladder networks," *Complexity*, vol. 2021, 2021.

[22] S. K. Sharma and V. K. Bhat, "On fault-tolerant resolvability of double antiprism and its related graphs," 2021, https://arxiv.org/abs/2104.09167.

[23] M. Kordestani, A. A. Safavi, N. Sharafi, and M. Saif, "Fault tolerant control of rhine-meuse delta water system: a performance assessment based approach," in *Proceedings of the International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET)*, pp. 1–8, IEEE, Istanbul, Turkey, August 2019.

[24] R. M. Souza, E. G. Nascimento, U. A. Miranda, W. J. Silva, and H. A. Lepikson, "Deep learning for diagnosis and classification of faults in industrial rotating machinery," *Computers & Industrial Engineering*, vol. 153, Article ID 107060, 2021.

[25] A. Nadeem, A. Kashif, A. Aljaedi, and S. Zafar, "On the fault tolerant partition resolvability of toeplitz networks," *Mathematical Problems in Engineering*, vol. 2022, Article ID 3429091, 8 pages, 2022.

[26] R. V. Voronov, "The fault-tolerant metric dimension of the king's graph," *Vestnik of Saint Petersburg University Applied Mathematics Computer Science Control Processes*, vol. 13, no. 3, pp. 241–249, 2017.

[27] H. Raza, S. Hayat, M. Imran, and X. F. Pan, "Fault-tolerant resolvability and extremal structures of graphs," *Mathematics*, vol. 7, no. 1, 2019.

[28] X. Guo, M. Faheem, Z. Zahid, W. Nazeer, and J. Li, "Fault-tolerant resolvability in some classes of line graphs," *Mathmatical Problems in Engineering*, vol. 2020, Article ID 1436872, 8 pages, 2020.

[29] M. Faheem, Z. Zahid, D. Alrowaili, I. Siddique, and A. Iampan, "fault-tolerant resolvability in some classes of subdivision graphs," *Journal of Mathematics*, vol. 2022, Article ID 5784800, 15 pages, 2022.

[30] A. Simic, M. Bogdanovic, Z. Maksimovic, and J. Milosevic, "fault-tolerant metric dimension problem: a new integer linear programming formulation and exact formula for grid graph," *Krag. Journal of Math.*vol. 42, no. 4, pp. 495–503, 2018.

[31] L. Saha, M. Basak, and K. Tiwary, "All metric bases and fault-tolerant metric dimension for square of grid," *Opuscula Mathematica*, vol. 42, no. 1, pp. 93–111, 2022.

[32] Z. Ahmad, M. A. Chaudhary, A. Q. Baig, and M. A. Zahid, "On metric dimension of P(n, 2)⊙K1 graph," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 24, no. 2, pp. 629–645, 2021.

[33] Z. Hussain, M. Munir, I. Ali, A. Ahmad, and S. Kang, *Fault-Tolerant Metric Dimension of Wheel Related Graphs*, 2019.

[34] Y. Tan, J. Zhang, H. Tian et al., "Multi-label classification for simultaneous fault diagnosis of marine machinery: a comparative study," *Ocean Engineering*, vol. 239, Article ID 109723, 2021.

[35] A. Dineva, A. Mosavi, M. Gyimesi, I. Vajda, N. Nabipour, and T. Rabczuk, "Fault diagnosis of rotating electrical machines using multi-label classification," *Applied Sciences*, vol. 9, no. 23, 2019.

[36] I. Javaid, M. Salman, M. A. Chaudhry, and S. Shokat, "Fault-Tolerance in resolvability," *Utilitas Mathematica*, vol. 80, pp. 263–275, 2009.
[37] A. Estrado-Moreno, J. A. Rodriguez-Velaquez, and I. G. Yero, "The k– metric dimension of a graph," *Applied Mathematics and Information Sciences*, vol. 9, no. 6, pp. 2829–2840, 2015.
[38] M. Ahmad, N. Ameen, and Z. Zahid, "Computing edge version of metric dimension and boubly resolving sets of a graph," *Discrete Mathematics, Algorithms and Applications*, vol. 12, no. 5, Article ID 2050070, 2020.
[39] Z. Ahmad, M. O. Ahmad, A. Q. Baig, and M. Naeem, "fault-tolerant metric dimension of $P(n, 2)$ with prism graph," 2018, https://arxiv.org/abs/1811.05973.
[40] A. Ahmad, M. Bača, and S. Sultan, "Computing the metric dimension of Kayak Paddles graph and Cycles with chord," *Proyecciones (Antofagasta)*, vol. 39, no. 2, pp. 287–300, 2020.
[41] S. Khuller, B. Raghavachari, and A. Rosenfeld, "Landmarks in graphs," *Discrete Applied Mathematics*, vol. 70, pp. 217–229, 1996.
[42] J. B. Liu, Z. Zahid, R. Nasir, and W. Nazeer, "Edge version of metric dimension and doubly resolving sets of the necklace graph," *Mathematics*, vol. 6, no. 11, p. 243, 2018.
[43] A. Murtaza, A. Gohar, A. Usman, and M. T. Rahim, "On cycle related graphs with constant metric dimension," *Open Journal of Discrete Mathematics*, vol. 2, 2012.
[44] R. Nasir, S. Zafar, and Z. Zahid, "Edge metric dimension of graphs," *Ars Combinatoria*, vol. 147, pp. 143–155, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7648288, 16 pages
https://doi.org/10.1155/2022/7648288

Research Article

Application of FIFD-OMGD-1.5D Teager Energy to Extract Microfault Features of Rolling Bearing

Fangming Liu and Jie Ma

School of Mechatronics Engineering, Beijing Information Science and Technology University, Beijing, China

Correspondence should be addressed to Jie Ma; mjbeijing@163.com

Received 18 October 2021; Revised 18 January 2022; Accepted 29 January 2022; Published 27 February 2022

Academic Editor: Meng Li

Deconvolution-related methods are the mainstream choice when it comes to enhancing the pulse impact of bearing fault and reducing noise interference. Kurtogram algorithm is used to optimize the minimum generalized Lp/Lq deconvolution to improve the nonconvexity of other optimization criteria. However, it has low computational efficiency and poor diagnostic accuracy under strong background noise. The paper proposes an optimized method using protrugram algorithm that combines fast iterative filter decomposition (FIFD) with minimum generalized Lp/Lq deconvolution (OMGD) for the 1.5-dimension Teager energy spectrum demodulation. Here is the specific process of the application: Fast iterative filtering (FIF) was used to reduce noise interference before using the maximum kurtosis to obtain the center frequency and frequency band and optimize the filter design, which was for the MGD initialization operation to prevent the result from falling into the local optimal solution and check the interference of impulse noise to a certain extent. The 1.5-dimension Teager energy spectrum was then used for demodulation analysis to extract small fault features of rolling bearings. The verification of simulation signals and actual data showed that this method was better in terms of extraction effect and efficiency than the use of fast kurtogram algorithm to optimize minimum generalized Lp/Lq deconvolution when it comes to extracting microfault features with high interference of background noise.

1. Introduction

The bearing is an indispensable component in the mechanical system. It is prone to all types of failure because of its fast motion speed, large load, and complex structure [1]. All failures start from small faults (early faults or initial faults) according to the final analysis. However, when the bearing signals are being collected from sensors, the signals will inevitably be mixed with other vibration interference components. Besides, one of the characteristics of microfault is small amplitude, which can make the fault impulse response very weak. Thus, it is easy to drown in the strong background noise, and it is too hard to find and solve timely. Over time, the equipment may be damaged or shut down, even bringing significant personal casualty and huge economic loss. For example, from September 2003 to October 2004, major accidents of freight train derailment caused by bearing fatigue fracture occurred many times in some important sections of China, resulting in economic losses of more than 2 billion yuan [2]; in 2010, a Russian airliner had an accident during landing due to motor shaft failure, resulting in 132 deaths [3]; in 2015, a serious leakage accident occurred in a petrochemical company in China due to the rupture of the ball caused by the distortion and fracture of the bearing inner ring [4]. In addition, the operational safety of modern engineering system which has large scale and complexity has gradually reduced [5]. Therefore, the study on microfaults of rolling bearings is helpful to prevent equipment accidents and ensure the operation safety of mechanical systems [6–9]. However, how to detect and enhance fault impulse response and extract fault features effectively has always been an important concern in bearing microfault diagnosis, and it is also a popular subject in the study of nonlinear vibration signals.

Due to the background noise and interference, noise reduction filtering is a necessary step before extracting the

features of microfaults of rolling bearings. Wavelet can be used to denoise nonstationary signals [10]. However, its transform results depend on wavelet basis and lack adaptability. Furthermore, energy leakage occurs easily during the signal transformation [11], which leads to the nonideal results of denoising. Therefore, some decomposition methods are proposed to denoise the signal, such as empirical mode decomposition (EMD) [12], ensemble empirical mode decomposition (EEMD) [13], complementary ensemble empirical mode decomposition (CEEMD) [14], complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [15], and variational mode decomposition (VMD) [16]. Most decomposition technologies break down the signal into some IMFs. In order to achieve noise reduction, the components with more useful information are retained and the those with more noise are discarded. However, such decomposition methods only work better for the signals that have very short data points. When long data points are involved, they will have various problems, such as poor decomposition effect and long decomposition time [17]. Considering that the iterative filtering (IF) method is an iterative decomposition method inspired by EMD and that it can ensure convergence and stability [18], paper [19] has proposed an iterative decomposition method based on fast Fourier transform (FFT), namely, fast iterative filtering decomposition (FIFD) method. This method has excellent rapidity and accuracy in signal decomposition.

Furthermore, because the generalized Lp/Lq norm applied in blind deconvolution performs well in extracting sparse features from noise signals [20], the generalized Lp/Lq norm sparse filtering method for pulse feature enhancement is applied to rolling bearing fault diagnosis successfully [21]. Based on the contributions of the above papers, paper [22] has proposed the minimum correlation generalized Lp/Lq deconvolution method, which can successfully extract the fault features and complete the detection of composite faults. Because the objective function of the minimum generalized Lp/Lq deconvolution (MGD) is nonconvex, the result may fall into the local optimal solution if improper initial values are selected. Therefore, paper [9] has used the filter designed by fast kurtogram to provide appropriate initialization and achieve adaptive adjustment. It is helpful to find the frequency band of the best filter and optimize MGD. Therefore, it is very important to adopt an appropriate method based on deconvolution to enhance the impact of bearing fault pulse.

Moreover, since the nonlinear signal can modulate amplitude and modulation frequency, the nonlinear broadband component can be converted into narrowband component through demodulation technology; that is, the concentrated spectrum can be obtained [23]. Therefore, the traditional demodulation technology, such as Hilbert transform (HT) [24] and Teager energy operator (TEO) demodulation [25], is usually applied to the bearing fault signal that can modulate amplitude and modulation frequency after noise reduction and brings out the fault features. However, considering the poor noise resistance of these demodulation methods, the paper [26] has proposed that the 1.5-dimension energy spectrum be applied to analyze the bearing fault, since it can reduce noise well and recover nonlinear features. Thus, it works well in extracting the frequency doubling information of fault feature frequency.

To sum up, in order to solve the computational efficiency loss in the above optimization methods, this paper has proposed MGD optimization based on the protrugram algorithm (referred to as OMGD herein). It is used to obtain the center frequency and the corresponding frequency band at the maximal kurtosis [27, 28]. Moreover, we can use the filter parameters obtained from this algorithm to design the filter and then to complete MGD optimization. This method does not achieve calculation accuracy at the expense of calculation efficiency. On the contrary, it ensures not only the accuracy of the diagnosis results, but also the calculation speed. However, the optimization method alone is not enough to process the signal under strong background noise. Therefore, in order to maintain the performance of deconvolution when the signal is affected by strong background noise, this paper has proposed a combination of OMGD and FIFD. First, we used FIFD to eliminate the noise and other signal interference, improve the signal-to-noise ratio, and reduce the number of objective functions converging to the local optimal solution. Then, the deconvolution method was used to make it easier to converge to the global optimal solution and then to enhance the pulse characteristics of small faults. Finally, we used the 1.5-dimension Teager energy spectrum for demodulation analysis to complete the feature extraction of microfaults. With the proposed FIFD-OMGD-1.5-dimension Teager energy spectrum, we could extract the microfault features of rolling bearing efficiently and accurately even with serious interference of noise.

The rest of the content is arranged in the following order: first, an introduction of the formula derivation of FIFD method is presented in Section 2. Second, an introduction of the derivation process of OMGD is provided in Section 3. Finally, an introduction of the 1.5-dimension energy spectrum is given in Section 4. In Sections 3 and 4, we use the abovementioned methods to analyze and verify the analog signal and actual signal, respectively. Finally, we summarize the study in Section 5.

2. Methodology

2.1. Fast Iterative Filtering Decomposition Method. Fast iterative filter decomposition (FIFD) is essentially the discrete version of iterative filter decomposition (IFD) based on fast Fourier transform (FFT).

Assuming that the signal is $s(x), x \in R$, we sample it at n points $x_j = j/n - 1$ to ensure that the sampling rate captures all its details. The main objective of FIFD is that vector IMFs are obtained by vector $[s(x_j)]_{j=0}^{n-1}$. In addition, it is assumed that $\|[s(x_j)]\|_2 = 1$ to make the method more general.

For convenience, $[s(x_j)]_{j=0}^{n-1}$ is replaced by symbol s to simplify the formula.

There are two "while" loops in the discrete version of the iterative filter decomposition (DIFD) algorithm. The first "while" loop is known as an outer loop, and the second one is an inner loop.

In the inner loop, the first IMF can be obtained through $\mathrm{IMF}_1 = \lim_{m\longrightarrow\infty}(I - C_m)s_m$, where the matrix $C_m = [\omega_m(x_i - x_j)]_{i,j=0}^{n-1}$ represents the discrete convolution operator, and its structure is determined by the semi-support length l_m at each step m. A nonnegative number is used for ω_m. The closely supported symmetric filter ω_m within area 1 is generated by the convolution of the symmetric filter h_m and itself. We can use the following equation to calculate $\omega_m(u)$:

$$\omega_m(u) = \frac{\omega\left(g^{-1}(u)\right)}{\int_{-1}^{1}\omega(t)\left|g_m'(t)\right|\mathrm{d}t}, \tag{1}$$

where g_m is the scaling function, which is considered monotonically reversible. By linear scaling, $g^{-1}(u) = u/l_m$, $g_m'(t) = l_m \geq 0$. Thus, we can rewrite (1) as

$$\omega_m(u) = \frac{\omega(u/l_m)}{l_m}. \tag{2}$$

Suppose that for any number m and $m \geq 1$, there is $C = C_m$. Given the semi-simple eigenvalue $\{\lambda_p\}_{p=0,\ldots,n-1}$ of C and its eigenvector $\{u_p\}_{p=0,\ldots,n-1}$, we can define the column of matrix U as the eigenvector u_p. In addition, suppose there are k zero eigenvalues in C, where $k \in \{0, 1, \ldots, n-1\}$.

Therefore, in the external circulation, the first IMF is obtained by the following formula:

$$\begin{aligned}\mathrm{IMF}_1 &= \lim_{m\longrightarrow 0}(I - C)^m s \\ &= UQU^T s,\end{aligned} \tag{3}$$

where U is known as unitary matrix. Besides, the k elements in the diagonal of the diagonal matrix Q are equal to 1.

To make DIFD reach the accuracy δ in calculating an IMF, we need to approximate the number of iterations through the following equation:

$$\frac{N_0^{N_0}}{(N_0+1)^{N_0+1}} < \frac{\delta}{\|\tilde{s}\|_\infty\sqrt{n-1-k}}, \quad N_0 \in N, \tag{4}$$

where $\tilde{s} = U^T s$.

Therefore, we can calculate IMF from the following formula:

$$\mathrm{IMF} = UP\begin{bmatrix} 0 & & & & & & \\ & (1-\lambda_1)^{N_0} & & & & & \\ & & \ddots & & & & \\ & & & (1-\lambda_{n-1-k})^{N_0} & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \times P^T U^T s, \tag{5}$$

where the permutation matrix obtained by reordering the columns of U is represented by P. Since the column of U corresponds to the eigenvector of C, λ represents the corresponding eigenvalues in descending order.

$U^T s$ is obtained by DFT of s using FFT algorithm. In addition, its computational complexity is $n\log(n)$. Multiplying the matrix U on the left side of $U^T s$ is equivalent to calculating the IDFT, which can be completed by using the IFFT. Therefore, IMF can use the following formula for rapid calculation:

$$\mathrm{IMF} = \sum_{k=0}^{n-1} u_k\left(1-\lambda_k\right)_{\sigma_k}^{N_0} = \mathrm{IDFT}\left((I-D)^{N_0}\mathrm{DFT}(s)\right), \tag{6}$$

where D is a diagonal matrix in which its diagonal includes the eigenvalue of C, and σ_k is the $k-$th element obtained after DFT of s.

In order to shorten the execution time of the algorithm, a stop criterion is introduced here, that is,

$$\left\|M^I(s)(x) - M^{I+1}(s)(x)\right\|_{L^2} < \delta, \forall I \geq N_0, \tag{7}$$

where $M(s_m) = s_m - L_m(s_m) = s_{m+1}$ and $L_m(s_m)(x) = \int_{-l_m}^{l_m} s_m(x+t)\omega_m(t)\mathrm{d}t$.

In order to avoid m unlimited growth, we use the above stopping criterion to intervene, shown as follows:

$$S\,D := \frac{\left\|s_{m+1} - s_m\right\|_2}{\left\|s_m\right\|_2}. \tag{8}$$

When the value of $S\,D$ is less than or equal to δ, the process can be stopped, so as to greatly shorten the calculation time.

2.2. OMGD Method

2.2.1. Generalized Lp/Lq Norm. In some current literature, the concise sparse representation of signal is expressed by the following formula:

$$\begin{aligned}&\min_\theta \|\theta\|_{l_0}, \\ &\|x - A\theta\| \leq \varepsilon,\end{aligned} \tag{9}$$

where A represents the dictionary matrix or base matrix, θ represents the expected sparse representation, x represents the input signal, and ε represents the estimation error which is an infinitesimal positive number.

However, because it is difficult to optimize l_0, some literature has proposed to use l_p norm $(0 < p \leq 1)$ to represent the sparsity of signal. Because signal sparsity regularization was not applicable in the study, the study in [29] has proposed to make the signal sparse due to several advantages of l_1/l_2 in calculation. Its advantages are as follows: First, the scale remains unchanged. Second, it has all the input gradient information and will give an easy minimization algorithm. According to that, l_1/l_2 norm is a modified sparse index. We further extend l_1/l_2 to the generalized l_p/l_q norm, which is a generalized sparse measure. Its definition can be expressed by the following formula:

$$\begin{aligned}J_{p,q}(x) &= \log\left(\frac{q}{p}\right)\cdot\left(\frac{\|x\|_{l_p}}{\|x\|_{l_q}}\right)^p \\ &= \log\left(\frac{q}{p}\right)\cdot\frac{\sum_i^N |x_i|^p}{\left(\sum_i^N |x_i|^q\right)^{p/q}},\end{aligned} \tag{10}$$

where $\log(q/p)$ is a constant with a sign. According to the above explanations, the minimization method can be used to replace the optimization method.

2.2.2. OMGD. The method based on the generalized l_p/l_q norm can optimize the sparsity of the input characteristic matrix. Each increasing diagonal element in the characteristic matrix of the input signal obtained by this method is a constant. This matrix is called Hankel matrix, shown as follows:

$$H = \begin{bmatrix} x_1 & x_2 & \cdots & x_l \\ x_2 & x_3 & \cdots & x_{l+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N-l+1} & x_{N-l+2} & \cdots & x_N \end{bmatrix}, \tag{11}$$

where $\mathbf{x} \in R^N$ represents the input signal vector and $\mathbf{H} \in R^{(N-L)\times L}$ represents the Hankel matrix derived from $\mathbf{x}$. We use $\mathbf{f} = \mathbf{H} \cdot \mathbf{W}$ to describe the signal features, where $\mathbf{W} \in R^L$ is the weighted vector and $\mathbf{f} \in R^{N-L+1}$ represents the feature vector after sparse learning. The sparse constraint applied to the eigenvector $\mathbf{f}$ is expressed by the following equation:

$$\begin{cases} \min\limits_{\mathbf{f}} & J_{p,q}(\mathbf{f}) = \min\limits_{f} \dfrac{\sum_i^N |f_i|^p}{\left(\sum_i^N |f_i|^q\right)^{p/q}} \\ & \mathbf{f} = \mathbf{H} \cdot \mathbf{W}, \\ \text{subject to:} & \\ & \|\mathbf{W}\|_{L_2} = 1, p < q. \end{cases} \tag{12}$$

Because sparse filtering and deconvolution can be equated mathematically, the sparse filtering of generalized l_p/l_q norm is renamed as minimum generalized l_p/l_q deconvolution (hereinafter referred to as MGD) in [8, 9]. We can use the following equation to represent deconvolution:

$$y = x * g_{\text{inv}}, \tag{13}$$

where x indicates the input signal, g_{inv} is the inverse filter that we desire, y is the filtered signal, and $*$ is the convolution operator.

In MGD, according to the relationship between sparse filtering and deconvolution, we can rewrite the convolution process of the above formula as the following formula by using Hankel matrix:

$$\mathbf{y} = \mathbf{H} \cdot \mathbf{W}, \tag{14}$$

where $\mathbf{W} \in R^L$ is equivalent to g_{inv} in the formula. At this time, the filtered signal $\mathbf{y}$ to which the sparse constraint is applied has sparse features, which can be indicated as follows:

$$\begin{cases} \min\limits_{\mathbf{y},\mathbf{w}} & J_{p,q}(\mathbf{y}), \\ & \mathbf{y} = \mathbf{H} \cdot \mathbf{W}, \\ \text{subject to} & \|\mathbf{W}\|_2 = 1, p < q. \end{cases} \tag{15}$$

The reason why we cannot minimize $J_{p,q}(\mathbf{y})$ is that the objective function in (15) is not smooth and the function also has nonconvexity. Therefore, MGD needs to be optimized:

(1) The nonsmooth problem can be solved by replacing $|y_n|$ with the soft absolute function $c_i = \sqrt{y_n^2 + \xi}$, where ξ is a small number, which is greater than 0 and is generally specified as $\xi = 10^{-8}$.

(2) We choose to use the gradient descent algorithm to optimize the nonconvexity of the function. The gradient term of the optimization algorithm is as follows:

$$\frac{\partial J_{p,q}}{\partial w_j} = \sum_i \frac{\partial J_{p,q}}{\partial c_i} \cdot \frac{\partial c_i}{\partial y_i} \cdot \frac{\partial y_i}{\partial w_j} = p \sum_i \left\{ \left[\frac{c_i^{p-1}}{\sum_i (c_i^q)^{p/q}} - \frac{(\sum_i c_i) \cdot c_i^{q-1}}{(\sum_i c_i^q)^{p/q+1}} \right] \cdot \frac{f_i}{\sqrt{f_i^2 + \varepsilon}} \cdot x_{i+j-1} \right\}. \tag{16}$$

However, it is indispensable to confirm the initial value before gradient descent. A good initial value is very important to solve the optimal solution. The initial value optimization method adopted in [9] is to realize the design of the filter through the fast kurtogram algorithm and then take the filter coefficient as the initial value. However, when the environmental noise has a great impact and the signal-to-noise ratio is low, the fast kurtogram can only provide a large bandwidth. Therefore, it may result in the lack of accuracy of the center frequency and filter bandwidth and then make it unable to find the parameters correctly.

Therefore, in this paper, we used the protrugram algorithm to find the optimal initial value of MGD. The algorithm selects the optimal frequency band based on the kurtosis of the signal narrowband envelope spectrum, which solves the problem of low parameter accuracy obtained by fast kurtogram algorithm. The general steps are as follows:

(1) The p and q parameter values of the generalized Lp/Lq norm, the number of decomposition layers k, and the filter length L are set; the coefficients of the stage 0 all-pass filter are initialized to $[0, 1, 0, \ldots, 0]$

(2) Taking three times of the fault characteristic frequency of the rolling bearing as the fixed bandwidth value of the protrugram method, the kurtosis value corresponding to each center frequency is calculated according to the protrugram algorithm

(3) The bandwidth and center frequency with the largest kurtosis are selected as filter parameters to design the corresponding filter

(4) The designed filter coefficient is used as the initial value of MGD to complete the optimization work

2.3. 1.5-Dimension Teager Energy Spectrum

2.3.1. Review of the Teager Energy Operator. We use $y(t)$ to represent the signal and define the energy operator ψ by

$$\Psi[y(t)] = \left[\frac{dy(t)}{dt}\right]^2 - y(t)\frac{d^2y(t)}{dt^2}, \tag{17}$$

where $\Psi[y(t)]$ is the instantaneous energy signal, and $dy(t)/dt$ and $d^2y(t)/dt^2$ are the first derivative and second derivative of signal $y(t)$, respectively. Teager energy operator has good time resolution. Its output can quickly and accurately track the change of total energy required by the signal and enhance the transient features.

2.3.2. Definitions of 1.5-Dimension Spectrum. For the stationary random signal $y(t)$, the third-order cumulant diagonal slice $R_{3y}(T,T)(T_1 = T_2 = T)$ is

$$R_{3y}(T,T) = E[y(n)y(n+T)y(n+T)]. \tag{18}$$

By performing one-dimension Fourier transform on $R_{3y}(T,T)$, the 1.5-dimension spectrum $B(\omega)$ of signal $y(t)$ is obtained as

$$B(\omega) = \int_{-\infty}^{\infty} R_{3y}(T,T)e^{-j\omega T}\mathrm{d}T. \tag{19}$$

In addition, the 1.5-dimension spectrum has the properties of strengthening the fundamental frequency component, suppressing Gaussian white noise, and detecting harmonic components [30].

2.3.3. 1.5-Dimension Teager Energy Spectrum. 1.5-dimension energy spectrum combines Teager energy operator and 1.5-dimension spectrum. Its principle is as follows: Firstly, the energy operator is used to demodulate the signal $y(t)$ to obtain the instantaneous energy signal so that the transient impact of the signal becomes obvious. Then, a 1.5-dimensional spectrum of $\Psi[y(t)]$ is made to suppress the noise component. Therefore, the 1.5-dimension Teager energy spectrum of $y(t)$ is

$$E(\omega) = \int_{-\infty}^{\infty} R_{3\Psi}(\tau,\tau)e^{-j\omega\tau}\mathrm{d}\tau. \tag{20}$$

Because the 1.5-dimension energy spectrum can not only effectively highlight the transient impact features, but also suppress the interference of noise, it is very appropriate for extracting the impact features of bearing fault signal.

2.4. Proposed Method for Bearing Microfault Diagnosis under Strong Background Noise. In this paper, a bearing microfault feature extraction method under strong background noise has been proposed. The flowchart of the proposed method is shown in Figure 1. It consists of three main steps: First, the original vibration signal polluted by interference and noise is decomposed by FIFD method, and the useful components are selected by the combination of correlation coefficient and kurtosis criterion to complete the signal reconstruction. In the second step, OMGD is performed on the reconstructed signal to enhance the pulse characteristics. During this step, we use the parameters obtained by the protrugram algorithm to optimize the filter and then take the filter coefficient as the initial value of MGD to optimize MGD. In the third step, the 1.5-dimension Teager energy spectrum is used for demodulation analysis to extract the characteristics of bearing microfaults under strong background noise.

3. Simulation Validation

3.1. Construction of Bearing Fault Simulation Signal. In this part, we use the same method to diagnose the analog signal to further verify its effectiveness. In this part, we construct the simulation model of defective bearing as follows:

$$Y = y_0 e^{-2\pi g f_n t_0}\sin\left(2\pi f_s\sqrt{1-g^2}\right)t_0 + n(t), \tag{21}$$

where y_0 is the displacement constant, the natural frequency is shown by f_n, g is the damping coefficient, f_s is the sampling frequency, and t_0 is the single cycle sampling time. Besides, $n(t)$ is Gaussian white noise and its amplitude is 1.5 [31].

Set $y_0 = 2, g = 0.1, f_n = 2kHz$, and $f_s = 20kHz$; the repetition period is 0.1 s; the number of sampling points is 10000; and the fault feature frequency of the simulation signal is 100 Hz.

The waveform diagram and envelope spectrum from the above simulation model are shown in Figure 2. Due to the interference of noise, we cannot identify the fault impact component of the signal. Direct envelope analysis of simulation signals cannot directly find the correct fault frequency.

3.2. Verification of the Proposed Method. Now, the proposed method is used to extract the fault characteristic frequency from the severely damaged signal. Fast iterative filter decomposition (FIFD) is implemented on the original simulation signal to obtain 8 IMFs. Besides, the correlation degree between each component and bearing signal is evaluated by calculating the cross-correlation coefficient. The effective component with more information has more correlation with the signal, and the calculated correlation value is large [32]. Of course, kurtosis is often used to characterize bearing faults. The larger the kurtosis value is, the more obvious the fault impact component is, and the easier it is to extract fault

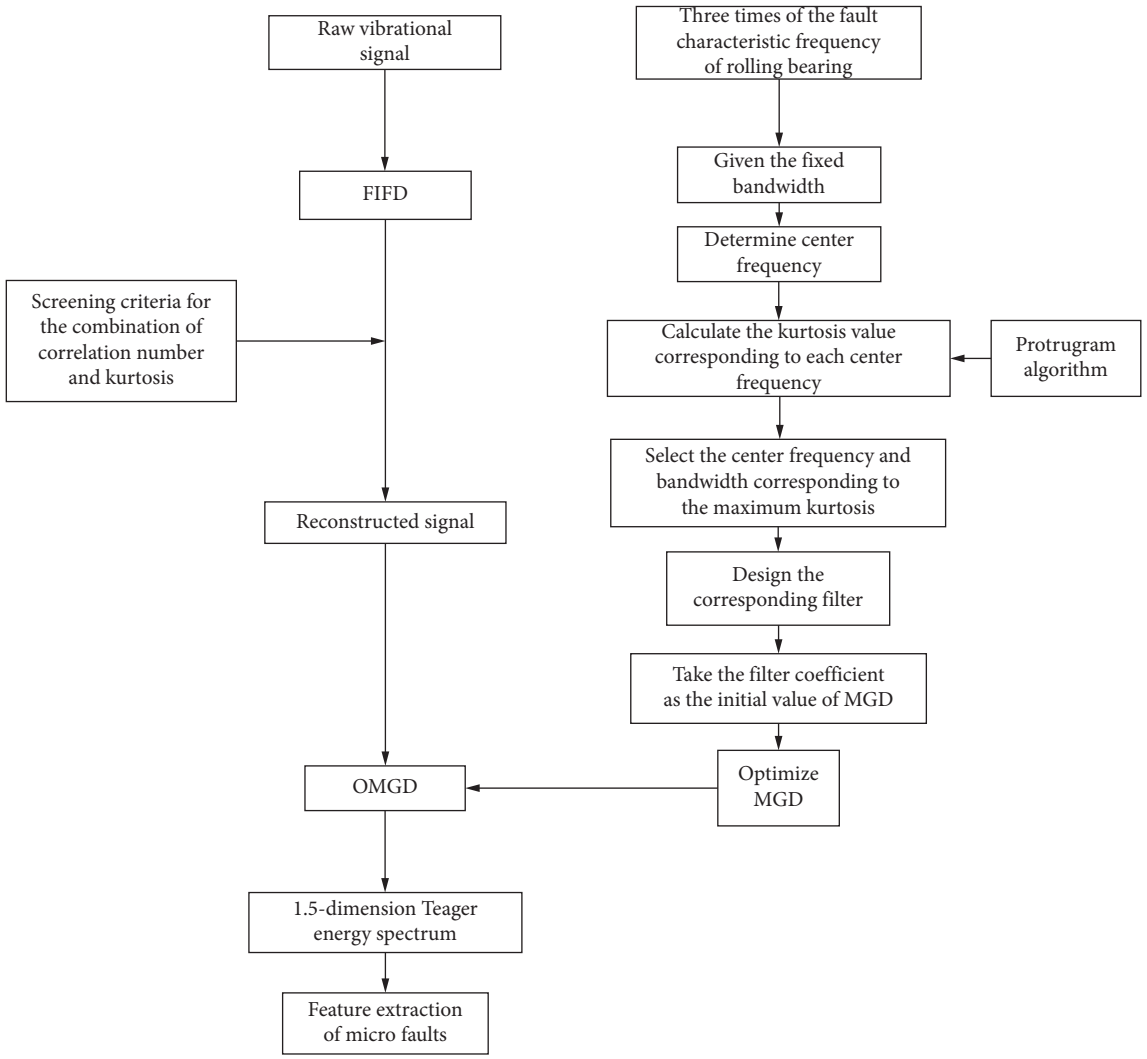


Figure 1: Flowchart of the proposed bearing microfault diagnosis.

information [33]. However, because kurtosis is particularly susceptible to abnormal components, it is difficult to distinguish random pulses from cycle pulses [9, 34, 35]. Therefore, we use the screening method on cross-correlation coefficient and kurtosis value to select the sensitive IMF and then to complete the filtered signal reconstruction.

The cross-correlation coefficients and kurtosis values are shown in Tables 1 and 2. According to the screening criteria, we select IMF1 and IMF2 components for signal reconstruction after filtering. The waveform of the reconstructed signal is shown in Figure 3. We can find out that compared with the original simulation model, the reconstructed signal after preliminary filtering shows obvious impact, which proves the effect of the fast iterative filtering decomposition method in noise reduction filtering.

However, we can also see that after fast iterative filtering, there are still some noise and interference in the reconstructed signal. Therefore, its periodicity is not shown, and the amplitude of the filtered signal is significantly reduced. To solve these problems, we will use OMGD based on protrugram algorithm to filter the reconstructed signal again and enhance the fault pulse features, so as to reduce the difficulty of extracting small faults in the filtering process. Firstly, we set the parameters of the protrugram algorithm responsible for MGD initialization. We set the iteration step to 100; select the triple fault characteristic frequency as the fixed bandwidth, that is, bw = $300Hz$; and determine the center frequency as $CF = bw/2$: step: $Fs/2 - bw/2$. According to the protrugram algorithm, the center frequency corresponding to the maximum kurtosis value is 1050 Hz. According to the above values, the corresponding band-pass filter is designed, and the coefficient of the band-pass filter is taken as the initial value of MGD iteration. Then, we set the filter length to 100 in MGD, $p = 1$, $q = 2$. Analyzing the reconstructed signal by using OMGD based on protrugram algorithm, the processing result is shown in Figure 4(a). We can see from the figure that after OMGD processing the signal shows obvious periodicity and impact, the filtering effect is very obvious and the signal amplitude has been significantly improved. In addition, the signal processed by OMGD is analyzed by 1.5-dimension Teager energy spectrum. We can see that the fault feature frequency

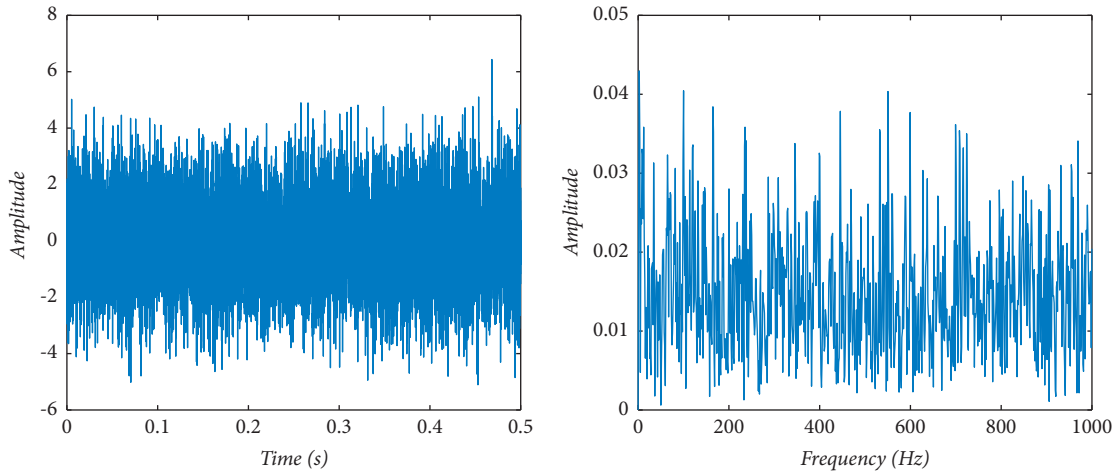


Figure 2: Simulation signal: (a) its waveform and (b) its corresponding envelope spectrum.

Table 1: Kurtosis values of components obtained by decomposing the simulation model.

IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8
2.8779	3.3622	3.0615	3.0491	3.3131	3.0418	4.6211	2.6637

Table 2: Correlation values of components obtained by decomposing the simulation model.

IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8
0.8329	0.5976	0.3352	0.2169	0.1481	0.0847	0.0732	0.0506

and frequency doubling feature of the simulation signal are effectively extracted. From the final envelope spectrum, we can know that using the FIFD-OMGD combined method for noise reduction filtering in addition to 1.5-dimension Teager energy spectrum analysis has a better effect in extracting small fault features of rolling bearing.

3.3. The Comparison of the Proposed Method with FIFD-OMGD (Based on Fast Kurtogram Algorithm), CEEMDAN-OMGD, and FIFD-MCKD. We compare the proposed method with FIFD-OMGD (based on fast kurtogram algorithm), CEEMDAN-OMGD, and FIFD-MCKD and analyze the results to evaluate the effectiveness of the proposed method.

Referring to the idea of control variable method, firstly, we apply OMGD based on fast kurtogram algorithm to the reconstructed signal after FIFD decomposition. The processing result is shown in Figure 5. In this method, we keep the parameter settings in MGD consistent with the parameters of OMGD based on protrugram algorithm; that is, the filter length in MGD is set to 100, $p = 1$, and $q = 2$. The only difference is that the two methods provide two different initialization techniques for the initial value of MGD. Among them, OMGD based on fast kurtogram algorithm uses fast kurtogram to search the resonance band of the bearing, designs the filter by laying the frequency plane in the case of 1/3 binary tree, and finally completes the initialization operation via taking the filter coefficient as the initial value.

By comparing Figures 4 and 5, we can see that the OMGD method using the fast kurtogram algorithm has exposed many problems in feature extraction of signals seriously polluted by noise. For example, when the measured signal is tampered with by some strong interference signals, the fault impulse enhancement effect has been reduced greatly. In contrast, it achieves more accuracy and robustness in extracting microfault features under strong background interference.

Through the comparison of the above contents, we can see that OMGD method based on protrugram algorithm has more advantages in enhancing the fault pulse impact of contaminated signal. Therefore, in the following comparison, we choose OMGD method based on protrugram algorithm. In order to highlight the advantages of FIFD algorithm in the proposed method, this paper selects the method of combining CEEMDAN and OMGD to process the contaminated signal. The processing result is shown in Figure 6.

Before using the above decomposition tool, we need to set some predefined parameters. The key parameters are as follows: for CEEMDAN, the noise standard deviation is set to 0.2, the number of times noise is added is 100, and the maximum number of iterations is 100. The number of IMFs is set to 8. For FIFD, the mask length is 2 in this paper and the number of IMFs is 8. Unless otherwise specified, these parameter values no longer change.

In order to reflect the decomposition accuracy of the two decomposition methods better, we calculate the RMSE (root mean square error) between the eight IMF components of the two methods and the ground truth. It can be clearly seen from Figure 7 that the RMSE of IMFs decomposed by FIFD is significantly smaller than that of IMFs decomposed by CEEMDAN algorithm. Therefore, FIFD is obviously better than CEEMDAN algorithm in decomposition accuracy.

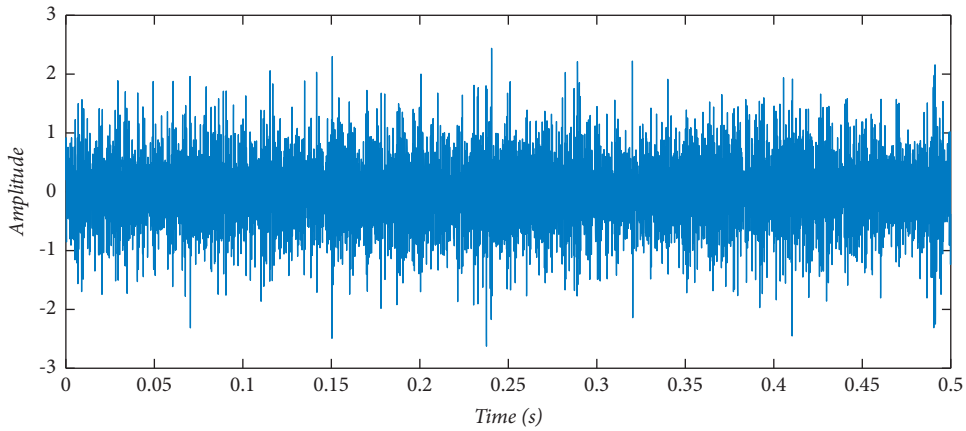


Figure 3: Reconstructed signal after FIFD filtering.

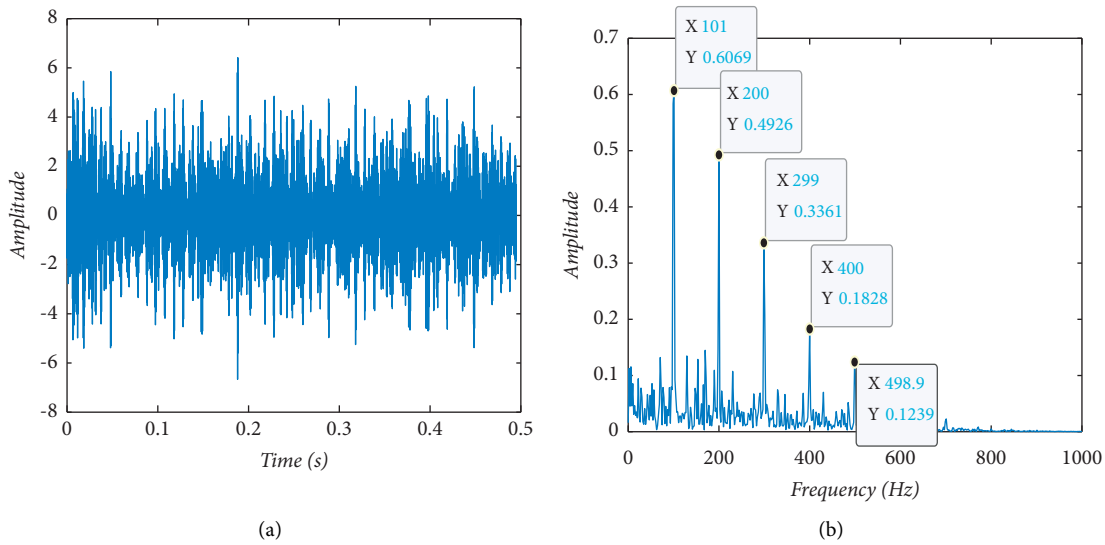


Figure 4: The reconstructed signal based on protrugram algorithm processing after OMGD: (a) its waveform and (b) its 1.5D Teager energy spectrum.

In addition, by comparing Figures 4 and 6, we see that the signal after being denoised and filtered using the CEEMDAN-OMGD method presents obvious impact and periodicity. Compared with the signal filtered by the other method, the signal is sparser, which makes it sensible to suspect that using this method may result in losing some useful information in the process of denoising and filtering. Finally, the filtered signal was analyzed using the 1.5-dimension Teager energy spectrum. It can be observed from the energy spectrum that only the triple frequency was extracted using the method. In addition, by comparison with the effect shown from the quintuple fault feature frequency extracted in Figure 4(b), we have more ground to suspect that the above method has lost some useful information in filtering. This also reveals the robustness of the FIFD-OMGD method.

Next, we choose to use FIFD algorithm to denoise the simulation signal, and then compare the denoised signal with MCKD method and OMGD method based on protrugram algorithm to verify the performance of OMGD. Similarly, we need to set the parameters before using MCKD. For MCKD, the number of iterations is set to 30, and the filter length is set to 100. Then, we use MCKD to enhance the fault pulse impact of the reconstructed signal processed by FIFD algorithm. From Figure 8, we see that the signal impact and periodicity after the signal was filtered by FIFD-MCKD are not obvious, which indicates that filtering using FIFD-OMGD combined with noise reduction is better than that using FIFD-MCKD. To further expand on that, in Figure 8(b), the feature frequency extraction effect of this method and the performance of frequency doubling are not ideal. In addition, the amplitudes of the first and second frequency doubling of the extracted fault feature frequency are extremely small compared with the frequency amplitude extracted by the FIFD-OMGD method. Thus, we conclude that the performance of OMGD is better in fault pulse features enhancement.

In order to illustrate the effectiveness of 1.5-dimension Teager energy spectrum, we performed envelope demodulation analysis on the signal after FIFD-OMGD noise

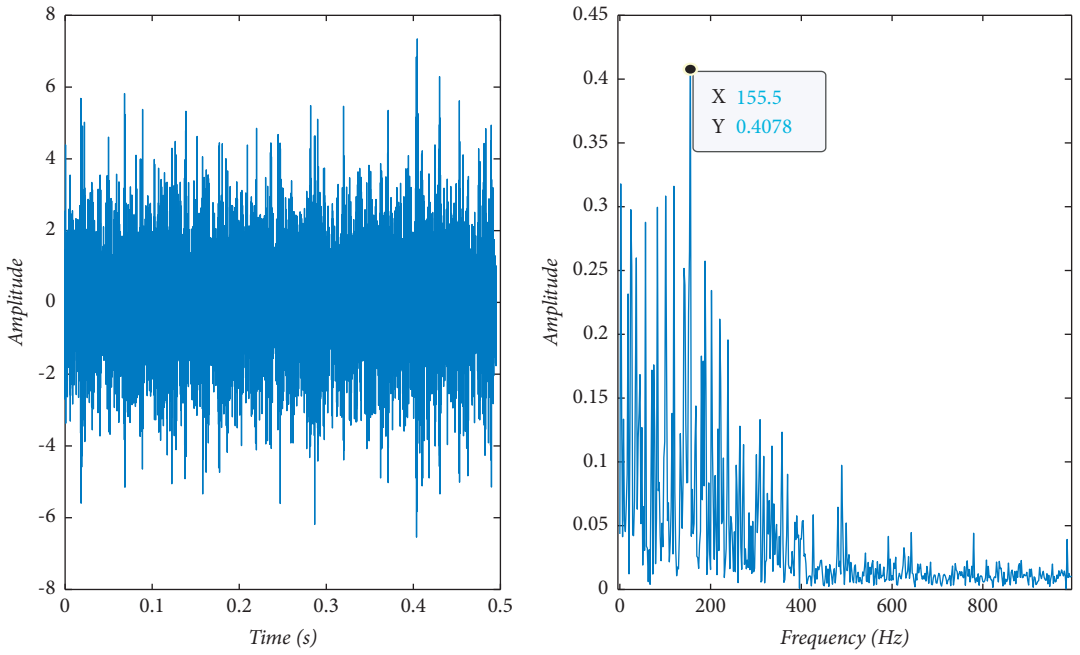


Figure 5: The signal processed by OMGD using fast kurtogram algorithm: (a) its waveform and (b) envelope spectrum.

reduction directly. The analysis result is shown in Figure 9. According to the comparison between Figures 4(b) and 9, we see that the effect of enveloping the joint filtered signal directly is not very satisfactory. However, using the 1.5-dimension Teager energy spectrum can well suppress the interference of noise and highlight the transient impact characteristics of the signal. The superiority of demodulation analysis using 1.5-dimension Teager energy spectrum is confirmed. Due to spatial constraints, we use 1.5-dimension Teager energy spectrum for demodulation analysis directly in the next experimental data verification part.

3.4. Computational Complexity. This section will compare the three methods mentioned with the proposed method in terms of calculation time. Figure 10 shows the relationship between data length and execution time, where the sampling data interval is 10000. It can be clearly seen from the figure that the calculation efficiency of the proposed method is the highest among the four methods. Moreover, when the data length increases from 10000 to 50000, the calculation time curve of the proposed method is always below the other three curves, which also shows that the calculation efficiency of the proposed method is still in the leading position among the four methods no matter how long the data length is.

3.5. Conservativeness of the Proposed Method. Because the proposed method in this paper is aimed at the signal polluted by strong background noise, we add different degrees of noise to the simulation signal in this section. The conservativeness of the proposed method is verified by comparing the fault feature frequency extracted by different methods with the ideal fault feature frequency under different SNR and calculating the average relative error between them. The results are obtained as shown in Table 3. It can be seen from Table 3 that with different SNR, the accuracy of fault feature frequency extraction by the proposed method is higher than that of other algorithms. In addition, as more and more noise is added, the ability of the proposed method to extract fault feature frequency is not affected, which also verifies that the method has good conservativeness.

4. Experimental Data Validation

The experimental data in this paper come from the accelerated life test data set of XJTU-SY rolling bearing conducted by Xi'an Jiao tong University. The data set collects the vibration signals of bearings switching from normal state to failure state, which can be used to diagnose the microfaults of rolling bearings [36]. The bearing tested is LDK UER204 rolling bearing, and the geometric parameters are shown in Table 4. During data acquisition, the sampling frequency was set to 25.6 kHz, the sampling interval was 1 min, and the sampling duration of each time was 1.28 s [36].

In this section, the fault vibration signal of the bearing outer ring was selected under the working conditions of 2400 r/min of rotation speed and 10 kN radial force. The fault feature frequency of the used bearing outer ring was 123.32 Hz, which is calculated by the calculation formula in Table 5.

The waveform and envelope of the outer ring vibration signal are shown intuitively in Figure 11. It can be seen that the original vibration signal contains a lot of random noise and interference and that the signal features are weak. The bearing fault feature frequency extracted by envelope analysis is not obvious.

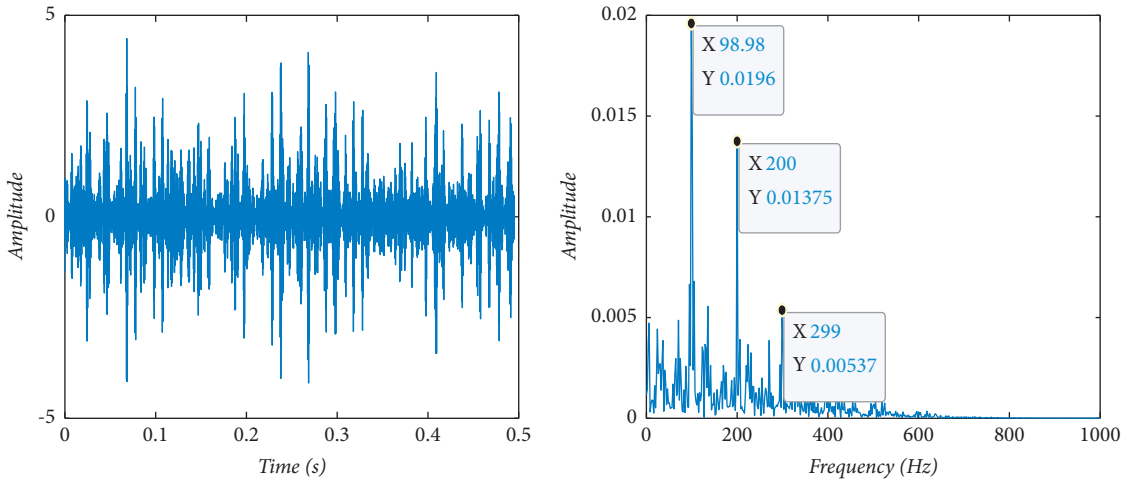


Figure 6: Filtered signal by CEEMDAN-OMGD: (a) its waveform and (b) its 1.5D Teager energy spectrum.

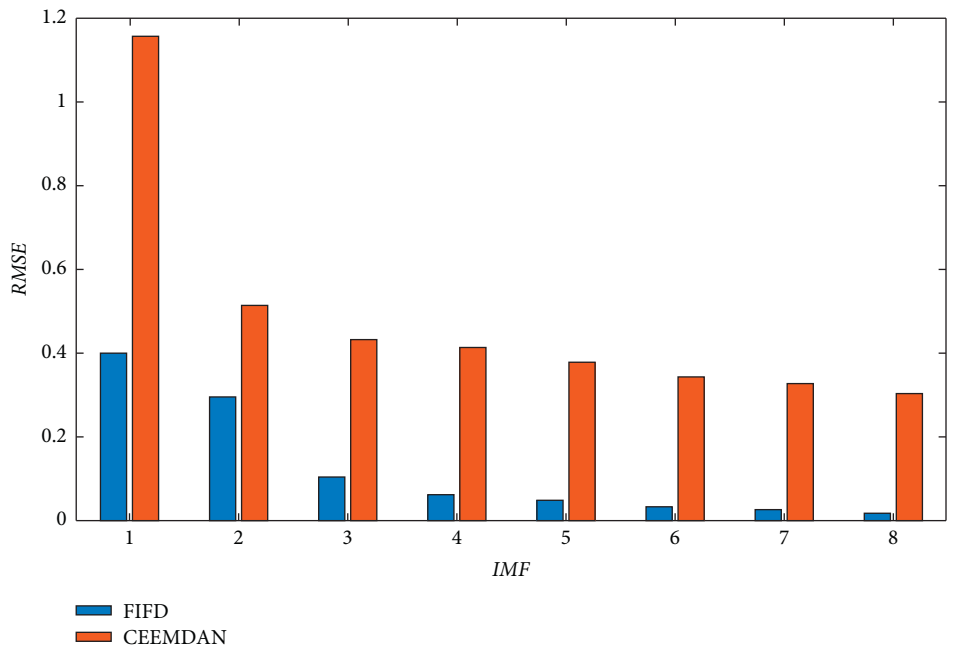


Figure 7: RMSE between 8 IMFs and the ground truth.

As in the previous section, the proposed method was used for fast iterative filter decomposition (FIFD) to obtain 8 IMFs on the original outer ring signal. The cross-correlation coefficients and kurtosis values of the IMFs from decomposing the outer ring signal are shown in Tables 6 and 7. According to the screening criteria, we select IMF1 and IMF3 components for signal reconstruction after filtering. The waveform of the reconstructed signal is shown in Figure 12. It is obvious that most random interference in the reconstructed signal after fast iterative filtering has been removed. It also shows obvious periodicity and impact, which also shows the effectiveness of fast iterative filtering (FIF) in noise reduction filtering.

However, in Figure 12, we see that the amplitude of the reconstructed signal after fast iterative filtering is small. To solve this problem, we use OMGD based on protrugram algorithm to increase the fault feature of reconstructed signal. In Figure 13, the signal amplitude was significantly improved, and the 1.5-dimension envelope of the signal shows that the fault feature frequency and its doubled frequency are effectively and obviously extracted. The extracted fault feature frequency is 124.1 Hz.

Similarly, we apply the proposed method, FIFD-OMGD (based on fast kurtogram algorithm), CEEMDAN-OMGD, and FIFD-MCKD to the actual fault data of bearing outer ring and compare their processing results (the parameter

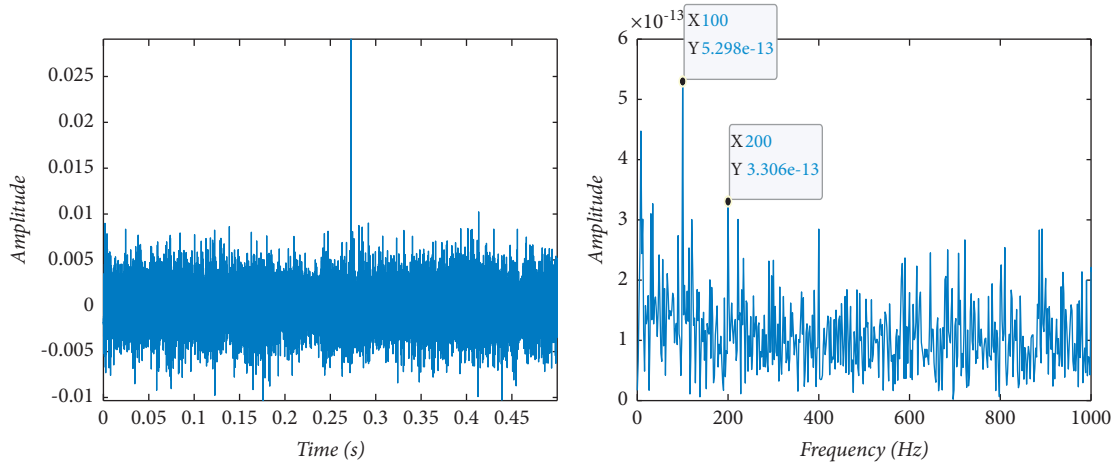


Figure 8: Filtered signal by FIFD-MCKD: (a) its waveform and (b) its 1.5D Teager energy spectrum.

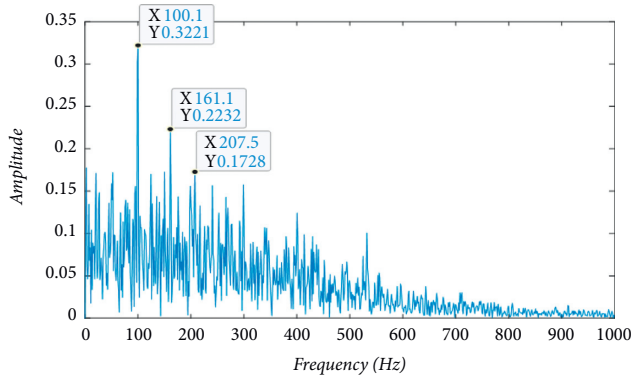


Figure 9: Envelope spectrum of the filtered signal after FIFD-OMGD processing.

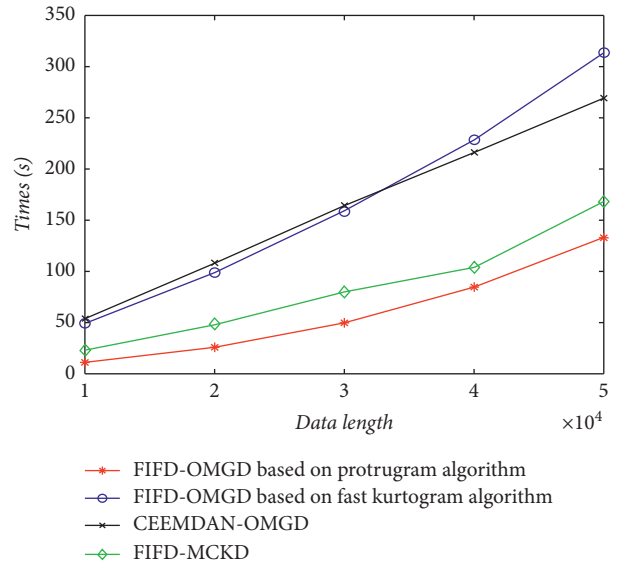


Figure 10: The relationship between data length and execution time.

settings of these algorithms have been described earlier and will not be described here). The effectiveness of the proposed method is further verified by analyzing the results.

Firstly, we apply FIFD-OMGD (based on fast kurtogram algorithm) to the bearing outer ring signal, and the processing results are shown in Figure 14. By comparing Figures 13 and 14, we can find that FIFD-OMGD (based on fast kurtogram algorithm) is not outstanding in feature extraction of signals seriously polluted by noise. This also verifies the superiority of the proposed method indirectly.

Secondly, in order to highlight the advantages of FIFD algorithm in the proposed method, this paper selects the method of combining CEEMDAN with OMGD based on protrugram algorithm to process the bearing outer ring signal. The processing results are shown in Figure 15. It can be seen from Figure 15(b) that the fault feature frequency extracted by CEEMDAN-OMGD method is not complete, and the frequency doubling of the fault feature frequency is submerged in the noise. Comparing this with Figure 13, we can see that FIFD-OMGD method is significantly better than CEEMDAN-OMGD method.

Finally, we choose to use the FIFD algorithm to denoise the bearing outer ring signal, and then use the MCKD method to process the denoised signal. The processing results are shown in Figure 16. It can be seen from Figure 16 that the signal impact and periodicity after filtering with FIFD-MCKD are not obvious, which shows that the effect of FIFD-OMGD combined with noise reduction is better than that of FIFD-MCKD. Moreover, in Figure 16(b), the extraction effect of this method on the fault feature frequency and its frequency doubling are not very ideal. Furthermore, the fault feature frequency and its frequency doubling of the outer ring signal are not clearly extracted. It can be seen that OMGD based on protrugram algorithm is obviously better in enhancing fault pulse feature.

In addition, we can clearly see from Table 8 that the computational efficiency of FIFD-OMGD (based on protrugram algorithm) is still better than that of the other three methods. Although the relative error of FIFD-MCKD in

Table 3: Fault feature frequency extracted by different methods and average relative error with different SNRs.

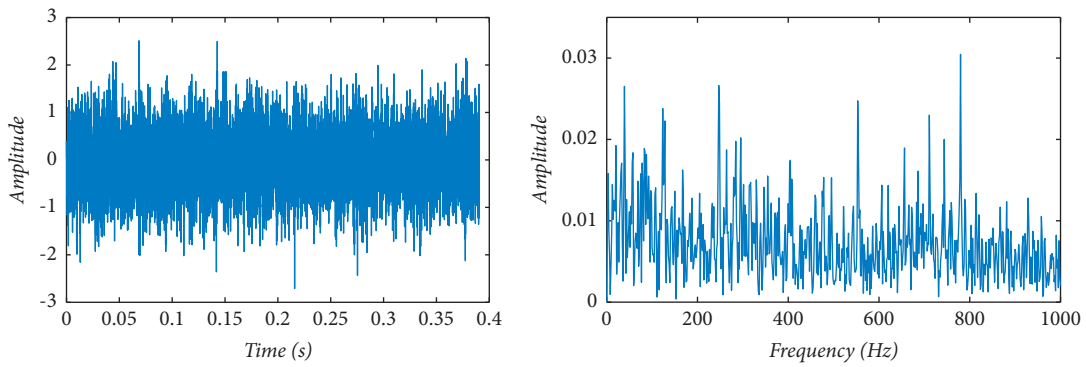
Fault feature Frequency (Hz)	SNR							
Algorithm	−3	−5	−7	−9	−11	−13	−15	Average relative error
True	100	100	100	100	100	100	100	
FIFD-OMGD (based on protrugram algorithm)	99.8	99.8	99.8	100.2	100.2	99.8	100.2	0.000200
FIFD-OMGD (based on fast kurtogram algorithm)	101	98.98	98.98	101	98.98	98.98	98.98	0.010143
CEEMDAN-OMGD	101	101	102.2	101	101.6	98.98	102	0.014029
FIFD-MCKD	100.1	100.1	101	98.6	107.3	111.8	89.4	0.046142

Table 4: Geometric parameters of LDK UER204 rolling bearing.

Number of balls m (piece)	Ball diameter d (mm)	Bearing pitch diameter D (mm)	Contact angle α (°)
8	7.92	34.55	0

Table 5: Calculation formula of fault characteristic frequency of rolling bearing.

Fault type	Calculation formula of fault characteristic frequency of rolling bearing
Outer ring fault	$f_0 = n/60\,(1 - d/D\mathrm{COS}\alpha)m/2$



(a) (b)

Figure 11: Original signal: (a) its waveform and (b) its envelope spectrum.

Table 6: Kurtosis values of components obtained by decomposing the outer ring signal.

IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8
3.2036	3.0818	3.2593	2.3082	1.7219	7.7511	3.0594	1.8473

Table 7: Correlation values of the components obtained by decomposing the outer ring signal.

IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8
0.8116	0.3801	0.4740	0.3051	0.1194	0.0174	0.0072	0.0056

extracting fault feature frequency is slightly less than that of the proposed method, the extraction effect of the proposed method is still the best in combination with the energy spectrum of the above method. Therefore, taking everything into consideration, the proposed method in this paper has good accuracy, high efficiency, and good conservativeness in extracting the microfault feature of seriously polluted signals.

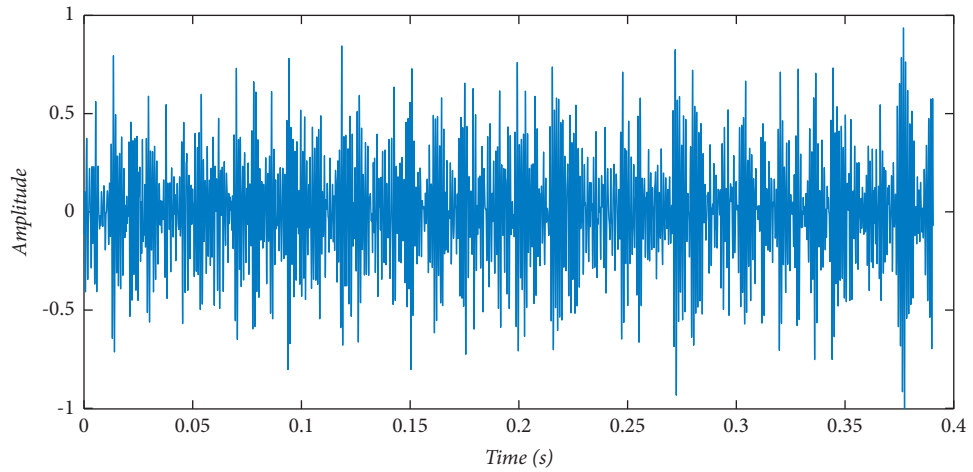


Figure 12: Reconstructed signal.

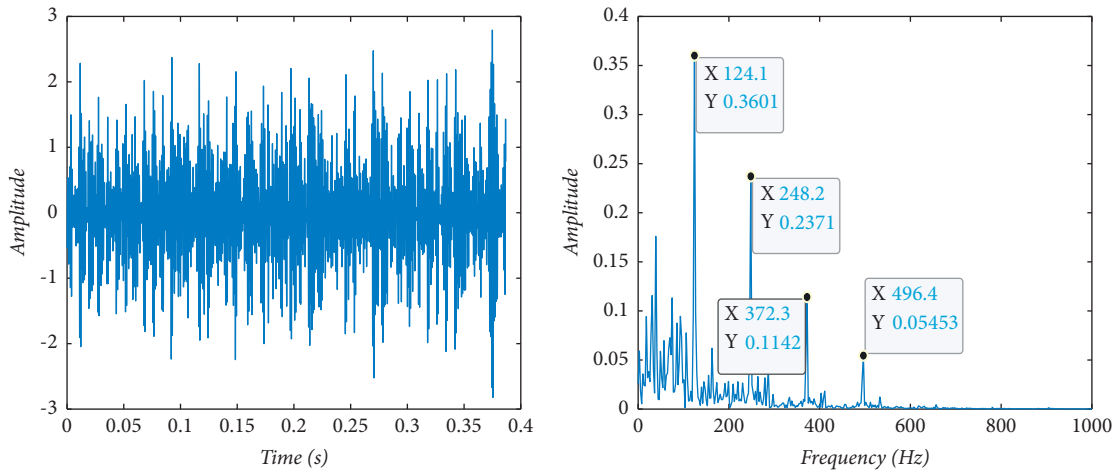


Figure 13: Filtered signal after processing reconstructed signal by OMGD: (a) its waveform and (b) its 1.5D Teager energy spectrum.

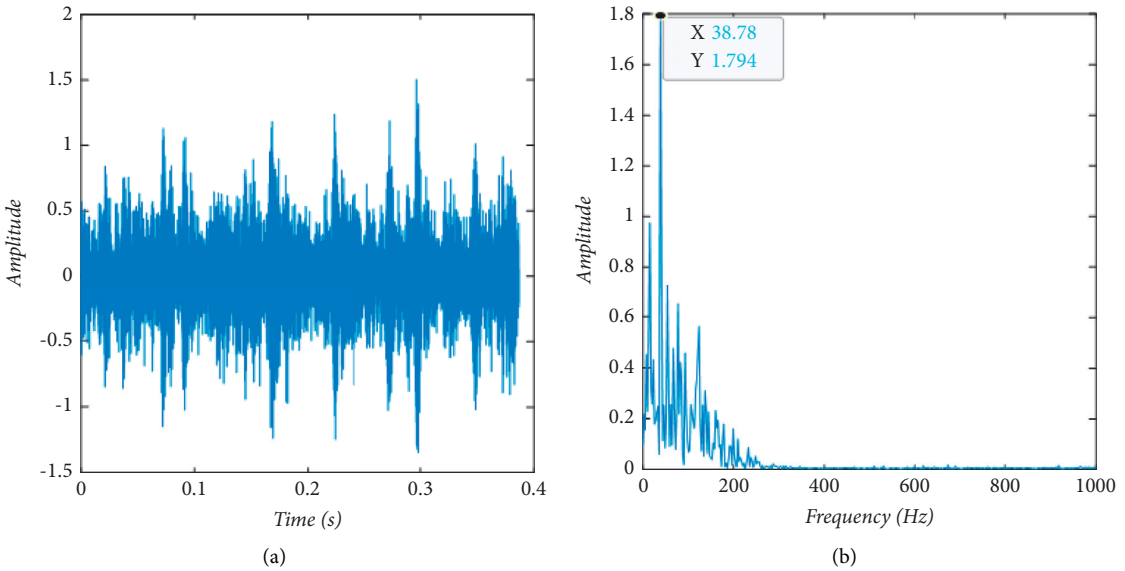


Figure 14: Filtered signal by FIFD-OMGD based on fast kurtogram algorithm: (a) its waveform and (b) its 1.5D Teager energy spectrum.

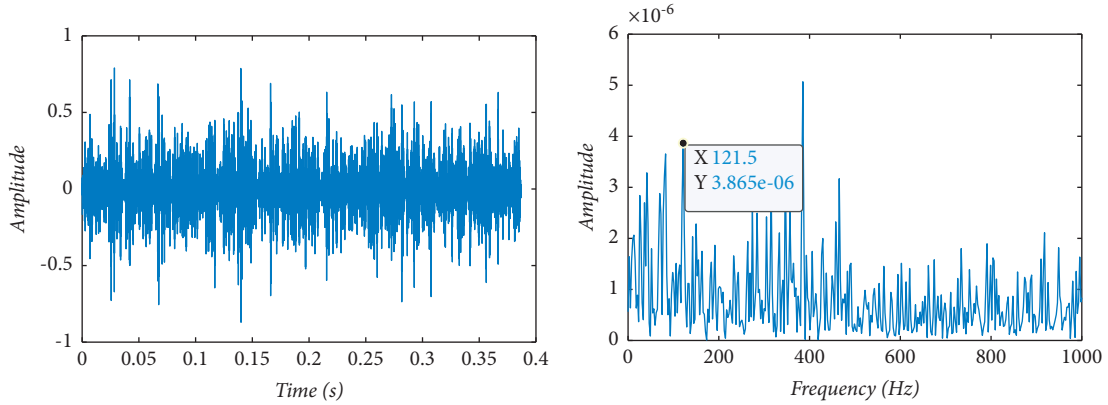


Figure 15: Filtered signal by CEEMDAN-OMGD: (a) its waveform and (b) its 1.5D Teager energy spectrum.

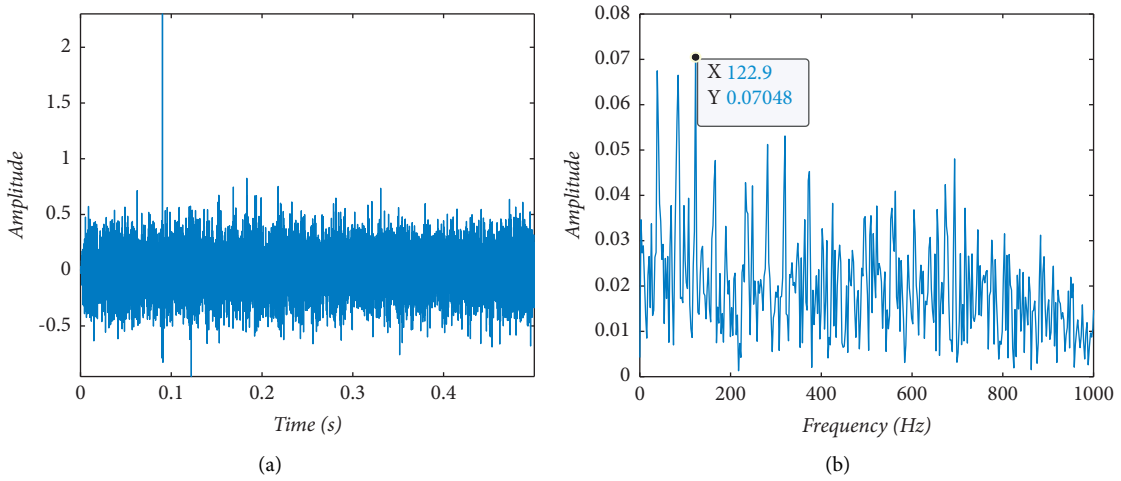


Figure 16: Filtered signal by FIFD-MCKD: (a) its waveform and (b) its 1.5D Teager energy spectrum.

Table 8: Computational time and error analysis of fault feature frequency for different algorithms.

Algorithm	Computational time (s)	FCF (Hz)	Relative error
True	—	123.32	
FIFD-OMGD (based on protrugram algorithm)	42.653022	124.1	0.006325
FIFD-OMGD (based on fast kurtogram algorithm)	133.733360	38.78	0.685534
CEEMDAN-OMGD	174.489326	121.5	0.014758
FIFD-MCKD	107.101329	122.9	0.003406

5. Conclusion

In this paper, the combination of FIFD-OMGD and 1.5-dimension energy spectrum was used to extract the features of microfault pulse under massive noise. In order to solve the problem of low precision and low efficiency of OMGD method using the fast kurtogram algorithm to extract bearing fault features under strong background noise, the protrugram algorithm was proposed to determine the filter parameters for the filter design, and the designed filter coefficient was adopted as the initial value of MGD to achieve more efficient pulse feature enhancement. After using different methods to analyze and compare the signal of the outer ring of the simulation bearing and the actual bearing, the following conclusions are drawn:

(1) By comparing the FIFD-OMGD with the OMGD method that uses fast kurtogram algorithm, we can see that the proposed method has better accuracy, noise robustness, and efficiency in fault feature extraction.

(2) By comparing CEEMDAN-OMGD and FIFD-OMGD, it can be seen that the decomposition speed of FIFD is significantly faster than CEEMDAN. This efficiency is very important for processing large-scale data. In addition, for strong background noise, FIFD

method has better noise reduction effect and stronger robustness.

(3) By comparing FIFD-OMGD with FIFD-MCKD, we can see that the effect of noise reduction of FIFD-OMGD combined with noise reduction filter is better. Moreover, FIFD-OMGD is obviously better in enhancing fault pulse features.

(4) By comparing the analysis results of direct envelope demodulation and 1.5-dimension Teager energy spectrum demodulation, we can see that the performance of the latter method is better in that it highlights the transient impact features of signal and restrains the interference of noise, which further affirms the superiority of 1.5-dimension Teager energy spectrum in demodulation analysis.

To sum up, to combine FIFD-OMGD with noise reduction filter and 1.5-dimension Teager energy spectrum is more effective as a practical tool in the microfault feature detection of rolling bearings. In addition, because it is outstanding in microfault feature extraction with strong background interference, this method can be applied to a more general range of targets, such as planetary gear vibration signal seriously polluted by noise. Moreover, signal processing is combined with optimized deconvolution to enhance the performance of the latter, which provides a new idea for optimizing the deconvolution during the processing of strong interference signals in the future.

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

This work was supported by the National Natural Science Foundation of China (Grant no. 61973041) and National Key Research and Development Program of China (Grant no. 2019YFB1705403).

References

[1] X. Li, Z. Ma, D. Kang, and X. Li, "Fault diagnosis for rolling bearing based on VMD-FRFT," *Measurement*, vol. 155, 2020.

[2] S. J. Wang, *Method Research of Weak Fault Feature Extraction of Rolling Bearing Based on EEMD*, Shijiazhuang Tie Dao University, Shaoxing, China, 2016.

[3] J. J. Li, *Research and Application of the Fault Diagnosis of Rolling Bearing Based on the Sound Signal*, Shijiazhuang Tie Dao University, Shaoxing, China, 2017.

[4] Y. G. Yang, *Research on Rolling Bearing Fault Diagnosis Method Based on Neural Network and Support Vector Machine*, Lanzhou Jiao Tong University, Lanzhou, China, 2021.

[5] J. Ma and J. A. Xu, "Fault prediction algorithm for multiple mode process based on reconstruction technique," *Mathematical Problems in Engineering*, vol. 2015, Article ID 348729, 8 pages, 2015.

[6] J. Li, D. H. Zhou, X. S. Si, M. Y. Chen, and C. H. Xu, "Review of incipient fault diagnosis methods," *Control Theory & Applications*, vol. 29, no. 12, pp. 1517–1529, 2012.

[7] S. J. Dong, X. W. Pei, W. L. Wu, B. P. Tang, and X. X. Zhao, "Rolling bearing fault diagnosis method based on multilayer noise reduction technology and improved convolutional neural network," *Journal of Mechanical Engineering*, vol. 155, no. 12, pp. 148–156, 2021.

[8] G. Wang, Z. J. He, X. F. Chen, and Y. N. Lai, "Basic Research on machinery fault diagnosis-what is the prescription," *Journal of Mechanical Engineering*, vol. 49, no. 01, pp. 63–72, 2013.

[9] H. Liu, Y. Cai, W. Dong, W. Feng, and J. H. Lin, "Optimized minimum generalized lp/lq deconvolution for recovering repetitive impacts from a vibration mixture," *Measurement*, vol. 168, 2020.

[10] S. Gaci, "The use of wavelet-based denoising techniques to enhance the first-arrival picking on seismic traces," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4558–4563, 2014.

[11] Z. H. Luo, X. N. Xue, X. Z. Wang, B. H. Wu, and Z. He, "Study on the method of incipient motor bearing fault diagnosis based on wavelet transform and EMD," *Proceedings of the CSEE*, vol. 25, no. 14, pp. 125–129, 2005.

[12] 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 Mathematical Physical & Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[13] Z. H. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 101 pages, 2011.

[14] J.-R. Yeh, J.-S. Shieh, and N. E. Huang, "Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method," *Advances in Adaptive Data Analysis*, vol. 02, no. 2, pp. 135–156, 2010.

[15] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," *IEEE International Conference on Acoustics*, vol. 125, no. 3, pp. 4144–4147, 2011.

[16] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[17] Y. Xu, F. Fan, and X. Jiang, "A fast iterative filtering decomposition and symmetric difference analytic energy operator for bearing fault extraction," *ISA Transactions*, vol. 108, pp. 317–332, 2021.

[18] A. Stallone, A. Cicone, and M. Materassi, "New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms," *Scientific Reports*, vol. 10, no. 1, p. 15161, 2020.

[19] A. Cicone and H. Zhou, "Numerical analysis for iterative filtering with new efficient implementations based on FFT," *Numerische Mathematik*, vol. 147, no. 1, pp. 1–28, 2021.

[20] X. Jia, M. Zhao, M. Buzza, Y. Di, and J. Lee, "A geometrical investigation on the generalized l p/l q norm for blind deconvolution," *Signal Processing*, vol. 134, pp. 63–69, 2017.

[21] X. Jia, M. Zhao, Y. Di, P. Li, and J. Lee, "Sparse filtering with the generalized lp/lq norm and its applications to the condition monitoring of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 102C, pp. 198–213, 2017.

[22] L. He, Y. Li, Y. Liu, and J. Lin, "Minimum correlated generalized lp/lq deconvolution for recovering repetitive impacts from a vibration mixture," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 2043–2054, 2021.

[23] S. Wei, D. Wang, Z. K. Peng, and Z. P. Feng, "Variational nonlinear component decomposition for fault diagnosis of planetary gearboxes under variable speed conditions," *Mechanical Systems and Signal Processing*, vol. 162, 2021.

[24] M. Feldman, "Hilbert transform in vibration analysis," *Mechanical Systems and Signal Processing*, vol. 25, no. 3, pp. 735–802, 2011.

[25] P. Henríquez Rodríguez, J. B. Alonso, M. A. Ferrer, and C. M. Travieso, "Application of the Teager-Kaiser energy operator in bearing fault diagnosis," *ISA Transactions*, vol. 52, no. 2, pp. 278–284, 2013.

[26] M. T. Ge, X. Wang, and A. R. Liu, "Rolling bearing fault diagnosis based on 1.5-dimensional Teager energy spectrum," *Machine Design and Research*, vol. 31, no. 5, pp. 62–66, 2015.

[27] X. T. Zhang, L. W. Tang, P. Wang, and S. J. Deng, "Multi-band resonance demodulation method for bearing fault acoustic emission signal," *Journal of Vibration, Measurement & Diagnosis*, vol. 35, no. 2, 2015.

[28] J. Du, L. Cui, J. Zhang, J. Li, and J. Huang, "The method of quantitative trend diagnosis of rolling bearing fault based on protrugram and lempel-ziv," *Shock and Vibration*, vol. 2018, pp. 1–8, 2018.

[29] X. D. Jia, M. Zhao, Y. Di, C. Jin, and J. Lee, "Investigation on the kurtosis filter and the derivation of convolutional sparse filter for impulsive signature enhancement," *Journal of Sound and Vibration*, 2017.

[30] X. P. Ren, H. J. Huang, and P. Li, "Early fault diagnosis of rolling bearing based on ELMD and 1.5 dimension spectrum," *Machinery Design & Manufacture*, vol. 11, 2019.

[31] Y. A. Zhang, Z. G. Lu, C. Tao, and X. C. Wang, "Rolling bearing fault diagnosis method based on improved adaptive filtering and 1.5-d spectrum," *Machine Design and Research*, vol. 36, no. 6, 2020.

[32] Z. C. Xu, J. Wang, and J. W. Liu, "Bearing fault feature extraction method based on multi - layer noise re-duction," *Machine Tool & Hydraulics*, vol. 49, no. 16, pp. 174–179, 2021.

[33] A. J. H, L. Ma, and G. J. Tang, "Rolling bearing fault feature extraction method based on ensemble empirical mode decomposition and kurtosis criterion," *Proceeding of the CSEE*, vol. 32, no. 11, pp. 106–111, 2012.

[34] Y. Cheng, Z. Wang, W. Zhang, and G. Huang, "Particle swarm optimization algorithm to solve the deconvolution problem for rolling element bearing fault diagnosis," *ISA Transactions*, vol. 90, 2019.

[35] G. L. Mcdonald, Q. Zhao, and M. J. Zuo, "Maximum correlated Kurtosis deconvolution and application on gear tooth chip fault detection," *Mechanical Systems and Signal Processing*, vol. 33, pp. 237–255, 2012.

[36] B. Wang, Y. G. Lei, N. P. Li, and N. B. Li, "A Hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability*, vol. 9, pp. 1–12, 2018.

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4648653, 14 pages
https://doi.org/10.1155/2022/4648653

Research Article

Fault Diagnosis of Planetary Gear Based on FRWT and 2D-CNN

Jie Ma[1] **and Lei Jiao**[2]

[1]*School of Mechanical and Electrical Engineering, Beijing Information Science and Technology University, Beijing, China*
[2]*School of Automation, Beijing University of Information Science and Technology, Beijing, China*

Correspondence should be addressed to Jie Ma; mjbeijing@163.com

Received 23 August 2021; Accepted 6 January 2022; Published 10 February 2022

Academic Editor: Meng Li

The fault signals of planetary gears are nonstationary and nonlinear signals. It is difficult to extract weak fault features under strong background noise. This paper adopts a new filtering method, fractional Wavelet transform (FRWT). Compared with the traditional fractional Fourier transform (FRFT), it can improve the effect of noise reduction. This paper adopts a planetary gear fault diagnosis method combining fractional wavelet transform (FRWT) and two-dimensional convolutional neural network (2D-CNN). Firstly, several intrinsic mode component functions (IMFs) are obtained from the original vibration signal by AFSA-VMD decomposition, and the two components with the largest correlation coefficient are selected for signal reconstruction. Then, the reconstructed signal is filtered in fractional wavelet domain. By analyzing the wavelet energy entropy of the filtered signal, a two-dimensional normalized energy characteristic matrix is constructed and the two-dimensional features are input into the two-dimensional convolution neural network model for training. The simulation results show that the training effect of this method is better than that of FRFT-2D-CNN. Through the verification of the test set, we can know that the fault diagnosis of planetary gears can be realized accurately based on FRWT and 2D-CNN.

1. Introduction

As an important part of rotating machinery and equipment, planetary gears usually operate in a high-speed and high-power environment. They are widely used in aircraft manufacturing, coal mining machinery, wind power generation, ship manufacturing, and other industries. It is very easy to appear in the long-term operation process, smooth vibration phenomenon. Since the 1980s, many serious accidents have been caused by the fault of rotating equipment around the world, causing huge economic losses. About 80% of the faults occurred on the planetary gears [1]. Therefore, how to accurately diagnose the fault of planetary gears has important research significance.

At present, many achievements have been made in the research on fault diagnosis of planetary gears. Yu Jun and others proposed a planetary gear fault identification method that combines a stacked denoising autoencoder (SDAE) and a gated recurrent unit neural network (GRUNN) to solve the problem of low planetary gear fault recognition rate [1]. Gao Hongying and others proposed a planetary gear fault identification method combining complementary set empirical mode decomposition (CEEMD) and chaotic particle swarm kernel extreme learning machine (CPSO-ELM), which reduces the influence of external disturbances on planetary gear fault diagnosis [2]. Wang Zhenya and others proposed a fault diagnosis method based on optimized variational modal decomposition and multidomain manifold learning of the salvia group, which solved the problem of difficult feature extraction and identification of planetary gears [3]. Li Haiping proposed an intelligent diagnosis method combining Fast Fourier Transform (FFT) and Deep Confidence Network (DBN) to improve the accuracy of planetary gear fault diagnosis [4]. Li Yuheng proposed a fault diagnosis method that combines the ensemble empirical mode (EEMD) and the symmetrical differential energy operator to achieve accurate diagnosis of planetary gears and accurately obtain the fault characteristic frequency value of planetary gears [5]. Zhang et al. proposed a fault diagnosis method based on time-frequency characteristics and PSO-SVM, and verified that the method can quickly and

accurately identify the fault type of planetary gears from nonstationary signals [6]. Wang et al. proposed a gear fault diagnosis method based on multicriteria fault feature selection and heterogeneous integrated learning classification, which improved the accuracy and robustness of diagnosis [7]. Aiming at a kind of multimode process with hidden degenerate faults, a fault prediction algorithm based on the combination of multi-PCA model and fault reconstruction technology is proposed, which can well solve the fault prediction problem of multimode process data [8].

In order to realize the planetary gear fault diagnosis under strong background noise, this paper adopts the planetary gear fault diagnosis method combining fractional wavelet transform and two-dimensional convolutional neural network. Firstly, the planetary gear fault signal is denoised by fractional wavelet transform. Secondly, use wavelet packet to extract the one-dimensional normalized energy value of the filtered signal, and convert the obtained one-dimensional energy value into a two-dimensional energy feature map. Finally, use a two-dimensional convolutional neural network to establish a fault diagnosis model to achieve accurate identification of different faults under different working conditions.

2. The Theoretical Basis of Fractional Wavelet Transform

2.1. Discrete Wavelet Transform. In signal processing, the continuous wavelet is discretized. After the discretization, the continuous wavelet and its corresponding wavelet transform become the discrete wavelet transform. The discrete wavelet transform [8] is the second of the displacement and scale of the continuous wavelet transform. The power is discretized, which is essentially binary wavelet transform. In order to reduce the complexity of wavelet coefficients, the wavelet coefficients are taken at some discrete points, and the scale is discretized first. In order to reduce the wavelet transform coefficients of the remainder, we set the wavelet system. In order to reduce the wavelet transform coefficients of the remainder, we limit the values of a and b of the wavelet coefficient $\psi_{a,b}(t) = 1/\sqrt{a}\psi(t - b/a)$ to some discrete points and first discretize the scale, that is, let $a = a_0^j a_0 > 0$. At this time, the corresponding wavelet function is $a_0^{-j/2}\psi[a_0^{-j}(t - b)]$, where $j = 0, 1, 2, \ldots, b$ is a uniform discrete value. At scale j, since the width of $\psi(a_0^{-j}t)$ is a_0^j times $\psi(t)$, the interval can be used to expand a_0^j without changing $\psi(t)$. At this time, $\psi_{a,b}(t)$ is changed to

$$\begin{aligned}\psi_{a,b}(t) &= a_0^{-j/2}\psi\left[a_0^{-j}\left(t - ka_0^b\tau b_0\right)\right] \\ &= a_0^{-j}\psi\left[a_0^{-j}t - kb_0\right].\end{aligned} \tag{1}$$

The discrete wavelet transform is

$$\begin{aligned}WT\left(a_0^{-j}t - kb_0\right) = \int f(t)\psi_{a_0^j, kb_0}(t)\mathrm{d}t \\ j = 0, 1, 2, \ldots\ldots, k \in z.\end{aligned} \tag{2}$$

The continuous wavelet transform at $a = 2^j (j \in z)$ is called discrete binary wavelet transform, and its expression is

$$\begin{aligned}&\omega_j\left(2^j, b\right) \le f(t), \\ &\psi_{2^j,b}(t) \ge 2^{-1/2}\int f(t)\psi^*\left(\frac{t-b}{2^j}\right)\mathrm{d}t.\end{aligned} \tag{3}$$

2.2. Discrete Fractional Fourier Transform. The fractional Fourier transform is

$$X_p(u) = \int_{-\infty}^{+\infty} K_p(u, u')x(u')\mathrm{d}u'. \tag{4}$$

$K_P(u, u') = A_\alpha \exp[j\pi(u^2 \cot\alpha - 2uu' \csc\alpha + u'^2 \cot\alpha)]$ is called the kernel function of FRFT, $A_\alpha = \sqrt{1 - j\cot\alpha}$, $\alpha = p\pi/2$, $p \ne 2n, n$ is integers.

According to the definition given by formula (4), the formula of Ozaktas sampling fractional Fourier transform can be obtained as

$$\begin{aligned}X_P(u) = A_\alpha \int_{-\infty}^{+\infty} \exp\left[j\pi\left(u^2 \cot\alpha - 2uu' \csc\alpha + u'^2 \cot\alpha\right)\right] \\ x(u')\mathrm{d}u'.\end{aligned} \tag{5}$$

In formula (5),

$$\begin{aligned}A_\alpha &= \frac{\exp(-j\pi\mathrm{sgn}(\sin\alpha)/4 + j\alpha/2)}{|\sin\alpha|^{1/2}}, \\ \alpha &= \frac{p\pi}{2}.\end{aligned} \tag{6}$$

When the order $p\varepsilon[1, -1]$, formula (6) is decomposed into the calculation process of the following formulas:

$$g(u') = \exp\left[-j\pi u'^2 \tan\left(\frac{\alpha}{2}\right)\right]x(u'), \tag{7}$$

$$g'(u) = A_\alpha \int_{-\infty}^{\infty} \exp\left[j\pi\beta(u - u')^2\right]g(u')\mathrm{d}u', \tag{8}$$

$$X_p(u) = \exp\left[-j\pi u^2 \tan\left(\frac{\alpha}{2}\right)\right]g'(u). \tag{9}$$

Here, $g(u')$ and $g'(u)$ are just two intermediate results $\beta = \csc\alpha$, $-\pi/2 \le \alpha \le \pi/2$. Discretize equations (7)–(9) to obtain the numerical calculation method of discrete fractional Fourier transform [9].

2.3. Fractional Wavelet Transform. The scale factors $a = a_0^k$, $k \in z$ (where $a_0 > 1$) and the time shift factor $\Delta b = a_0^k b_0$ in the continuous fractional wavelet transform expression are discretized and sampled in the displacement domain, and the value corresponding to the sampling point can be expressed by the discrete fractional wavelet transform formula.

Discretize the scale factors $a = a_0^k$, $k \in z$ to get

$$\begin{aligned}\psi_{p:\,k,n}(t) &= e^{-jt^2-\left(na_0^k b_0\right)^2/2\cot\alpha}\frac{1}{a_0^k}\psi\left(\frac{t-na_0^k b_0}{a_0^k}\right)\\ &= e^{-jjt^2-\left(na_0^k b_0\right)^2/2\cot\alpha}a_0^{-k/2}\psi\left[a_0^{-k}\left(t-na_0^k b_0\right)\right]\\ &= e^{-jt^2-\left(na_0^k b_0\right)^2/2\cot\alpha}a_0^{-k/2}\psi\left[a_0^{-k}t-nb_0\right].\end{aligned} \tag{10}$$

When $a = a_0^0 = 1$, the expression of discrete fractional wavelet transform is

$$\begin{aligned}\mathrm{DFRWT}_f(k,n) &= \langle f(t),\psi_{p:\,k,n}(t)\rangle = \int_{-\infty}^{+\infty} f(t)\psi_{p:\,k,n}(t)\mathrm{d}t\\ &= 2^{-k/2}\int_{-\infty}^{+\infty} f(t)e^{jt^2-\left(n2^k\right)^2/2\cot\alpha}\psi\left[2^{-k}(t-n)\right]\mathrm{d}t.\end{aligned} \tag{11}$$

The reconstruction of the fractional wavelet transform is the inverse process of the decomposition process of the fractional wavelet. In the known k-th layer, the fractional wavelet coefficients are $\left\{c_m^{'k}\right\}_{m\in z}$ and $\left\{d_m^{'k}\right\}_{m\in z}$, and the original signal is $\{c_n^0\}$ through the reconstruction. $\{V_k^\alpha\}_{k\in z}$ is the multiresolution analysis, which can be seen from the relationship between $\psi_{p:\,k,n}(t)$, $\phi_{p:\,k,n}(t)$, and the function projection:

$$\begin{aligned}c_n^{'k+1} &= c_n^{k+1}e^{-i1/2\left(n\cdot 2^{k+1}\right)^2\cot\alpha}\\ &= \langle f(t),\phi_{p:\,k+1,n}(t)\rangle\\ &= \langle f_k(t),\phi_{p:\,k+1,n}(t)\rangle + \langle w_k(t),\phi_{p:\,k+1,n}(t)\rangle\\ &= \sum_{m\in z} c_m^{'k}\langle\phi_{p:\,k,m}(t),\phi_{p:\,k+1,n}(t)\rangle + \sum_{m\in z} d_m^{'k}\langle\psi_{p:\,k,m}(t),\phi_{p:\,k+1,n}(t)\rangle\\ &= \sum_{m\in z} c_m^{'k}e^{j1/2\left(m2^{-k}\right)^2\cot\alpha}e^{-j1/2\left(n2^{-(k+1)}\right)^2\cot\alpha}\langle\phi_{k,m}(t),\phi_{k+1,n}(t)\rangle + \sum_{m\in z} d_m^{'k}e^{j1/2\left(m2^{-k}\right)^2\cot\alpha}e^{-j1/2\left(n2^{-(k+1)}\right)^2\cot\alpha}\langle\psi_{k,m}(t),\phi_{k+1,n}(t)\rangle\\ &= \sum_{m\in z} c_m^{'k}e^{j1/2\left(m2^{-k}\right)^2\cot\alpha}\langle\phi_{k,m}(t),\phi_{k+1,n}(t)\rangle + \sum_{m\in z} d_m^{'k}e^{j1/2\left(m2^{-k}\right)^2\cot\alpha}\langle\psi_{k,m}(t),\phi_{k+1,n}(t)\rangle.\end{aligned} \tag{12}$$

Hence,

$$\begin{aligned}c_n^{k+1} &= \sum_{m\in z} c_m^k\langle\phi_{k,m}(t),\phi_{k+1,n}(t)\rangle + \sum_{m\in z} d_m^k e^{i(1/2)\left(n2^{j+1}\right)^2\cot\alpha}\langle\psi_{k,m}(t),\phi_{k+1,n}(t)\rangle\\ &= 2^{-1/2}\left[\sum_{m\in z} c_m^k h_0(n-2m) + \sum_{m\in z} d_m^k h_1(n-2m)\right].\end{aligned} \tag{13}$$

Equation (13) is the reconstruction process of traditional discrete wavelet coefficients. Firstly, the fractional coefficients $c_m^{'k}$ and $d_m^{'k}$ of the k layers are modulated, and then one-dimensional wavelet inverse transformation is performed in the wavelet domain to obtain c_n^{k+1}, and then c_n^{k+1} is modulated into the fractional wavelet domain to obtain $c_n^{'k+1}$, and so on, to restore the original signal $\{c_n^0\}$ step by step.

2.4. Realization Process of Fractional Wavelet Transform. With a one-dimensional signal $f(x)$, using the definition of fractional wavelet transform proposed by Menlovevic, the realization process of one-dimensional fractional wavelet transform can be obtained as follows:

(1) Input one-dimensional signal $f(x)$

(2) Select the appropriate fractional order change range p, and use the minimum output energy to search for the best transformation order

(3) Perform p-order fractional Fourier transform on the input signal $f(x)$ to obtain a signal in the fractional domain

(4) Perform wavelet decomposition on the obtained signal in the fractional domain to obtain a signal in the fractional wavelet domain

(5) Perform wavelet reconstruction on the signal in the fractional wavelet domain to recover the signal in the fractional domain

(6) Perform $-p$-order fractional Fourier transform on the signal in the fractional domain to recover the filtered signal [10, 11]

The realization process of fractional wavelet transform is shown in Figure 1 [8].

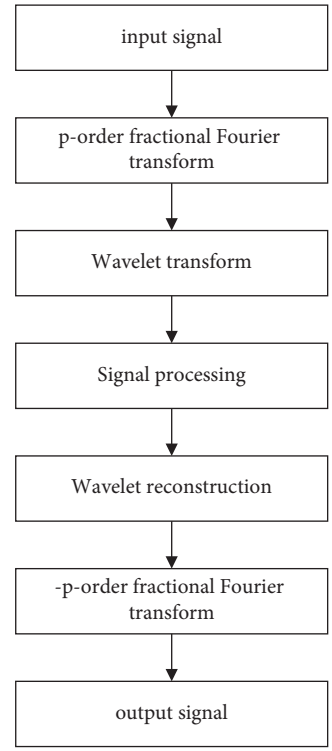


Figure 1: The basic process of fractional wavelet transform.

3. Wavelet Energy and Two-Dimensional Convolutional Neural Network

3.1. Feature Extraction Process Based on Wavelet Energy. The main steps are as follows:

(1) The signal is decomposed by n-layer wavelet packet, the j-th layer has 2^n frequency band signals, and then 2^n features of the n-th layer are extracted.

(2) In order to improve the denoising ability of the signal, select the low-frequency coefficients and high-frequency coefficients of each frequency band decomposed in (1) to reconstruct the signal, denoted as f.

(3) Solve the energy $E_{i,j}$ of each signal, and the calculation formula for the energy value of each frequency band is as follows:

$$E_{i,j}(t_j) = \int \left|f_{i,j}(t_j)\right|^2 \mathrm{d}t = \sum_{k=1}^{m} \left|x_{j,k}\right|^2 \quad j = 0, 1, 2, \ldots, 2^i - 1. \tag{14}$$

Here, x is the decomposition coefficient of the wavelet packet, $E_{i,j}(t_j)$ is the energy value of the j-th node in the i-th layer after the signal $x(t)$ undergoes wavelet decomposition, $k = 1, 2, \ldots, N_c$, and $x_{j,k}$ is the wavelet packet reconstruction coefficient of $f_{i,j}$ [12].

(4) Construct feature vector.

In the process of wavelet decomposition, the energy of each layer is equal to the total energy, and the total energy of the signal is

$$\begin{aligned} E_{\text{sum}} &= E_i \\ &= \sum_j E_{i,j}. \end{aligned} \tag{15}$$

The wavelet packet energy of each frequency band is

$$M_{i,j} = \frac{E_{i,j}}{E_{\text{sum}}}. \tag{16}$$

The wavelet packet energy feature vector is

$$W_n = \left(M_{n,0}, M_{n,1}, \ldots, M_{n,2^n-1}\right). \tag{17}$$

3.2. Structure of a Two-Dimensional Convolutional Neural Network. The current typical two-dimensional convolutional neural network structure is composed of input layer, convolution layer, pooling layer, fully connected layer, and output layer. The network structure of LeNet − 5 is shown in Figure 2. The input of the convolutional neural network is mainly in the form of a two-dimensional grayscale image or a color image. Its output layer uses the Softmax classifier to output the classification and recognition results of a two-dimensional grayscale image or a color image. In other image processing fields such as target detection, other forms of network output layers need to be set up [13, 14].

The convolutional layer is composed of multiple convolutional neurons. The parameters of the convolutional neuron are obtained by using the backpropagation algorithm. The convolutional layer is a key part of the entire convolutional neural network, which is mainly used for input data to extract different features [15]; the process of convolution operation is composed of continuous convolution and discrete convolution.

The process of discrete convolution operation is as follows:

$$y(n) = x(n) * h(n) = \sum_{i=0}^{N-1} x(i)h(n-i). \tag{18}$$

When the image convolution operation is performed, it is the operation between the image pixels. The pixels of the image can be understood as a matrix, and the pixels are not continuous. The process of the convolution operation is the selected convolution kernel and the image. Input for

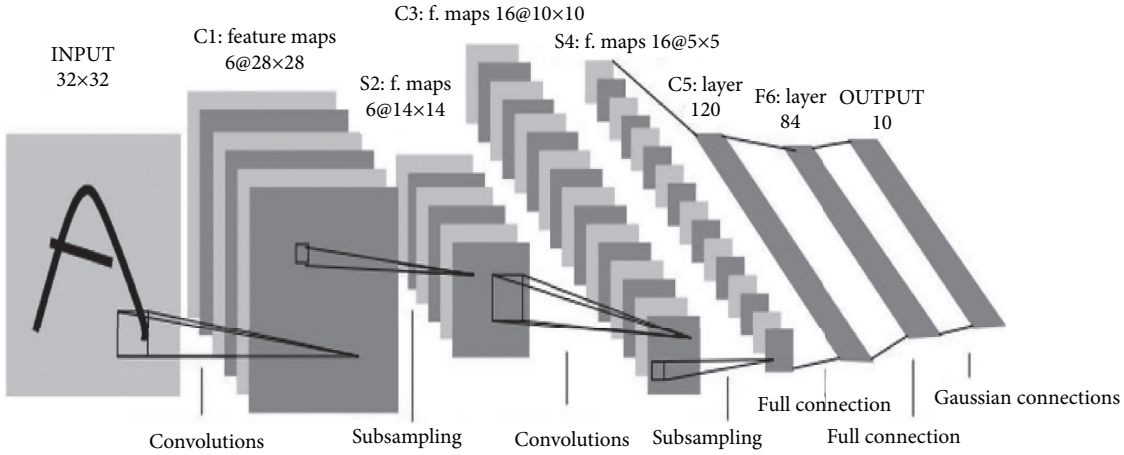


Figure 2: Schematic diagram of LeNet-5 structure.

convolution operation: Assuming that the two-dimensional image input is $I(i, j)$ and the two-dimensional convolution kernel is $K(m, n)$, the image convolution operation process can be expressed as

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n) K(i - m, j - n). \quad (19)$$

Convolution operation is alternating, so

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n). \quad (20)$$

Here, m, n is the size of the convolution kernel. After the feature is extracted by the convolution operation, the offset operation needs to be performed after the convolution operation. The calculation formula is as follows:

$$x_j^i = f\left(\sum_{i \in M_j} x_j^{i-1} * k_{ij}^l + b_j^l\right). \quad (21)$$

Here, x_j^i is the first feature map output by the first layer; $f(x)$ is the activation function used by the convolutional layer; k_{ij}^l is the convolution matrix used by the convolution kernel; and b is the offset of the convolution operation.

The pooling layer is also commonly referred to as the downsampling layer. The pooling layer can reduce the training time of the model, improve the robustness of feature extraction, and avoid overfitting of the model. There are usually three ways of pooling: average pooling process, maximum pooling process, and random pooling process. In actual applications, the pooling process is dominated by maximum pooling.

Maximum pooling calculation formula is

$$p_i = \underset{k \in M_j}{\text{Max}}\{a_i(k)\}. \quad (22)$$

In the actual application process, the classifier needs to be trained in the fully connected layer. The commonly used classifier is the Softmax classifier. The fully connected process is shown in the following formula:

$$y^k = f\left(w^k x^{k-1} + b^k\right). \quad (23)$$

In formula (23), y^k is the output of the fully connected layer; w^k is the weight value; x^{k-1} is the input of the fully connected layer; b^k is the bias term; $f(x)$ is the classification function; k is the network layer number.

In image classification, Softmax is generally used as the classifier. If there are K classifications, the output of Softmax can be expressed as

$$\sigma(x)_i = \frac{e^{zi}}{\sum_j^k e^{zj}}, \quad i = 1, 2, \ldots, k. \quad (24)$$

3.3. Procedure. In order to accurately classify planetary gear faults in a complex actual industrial environment, this paper proposes a planetary gear fault diagnosis method based on FRWT and 2D-CNN. A flowchart can be drawn as shown in Figure 3.

The specific steps are as follows:

(1) Use fractional wavelet transform to separately denoise the gear fault signals

(2) Use Shannon entropy to extract energy from the signal after noise reduction and calculate the normalized energy value

(3) Convert the obtained wavelet energy value into a two-dimensional matrix feature sample set

(4) Initialize the two-dimensional convolutional neural network and use the sample set to extract the characteristics of the signal

(5) Train and establish a two-dimensional convolutional neural network model to identify planetary gear faults

4. Experimental Verification

4.1. Introduction to the Experimental Sample Set. The planetary gear fault experiment data used in this article is collected by the QPZZ-II mechanical fault simulation and test platform produced by Jiangsu Qianpeng Diagnostic Engineering Co., Ltd. The test platform includes drive motors,

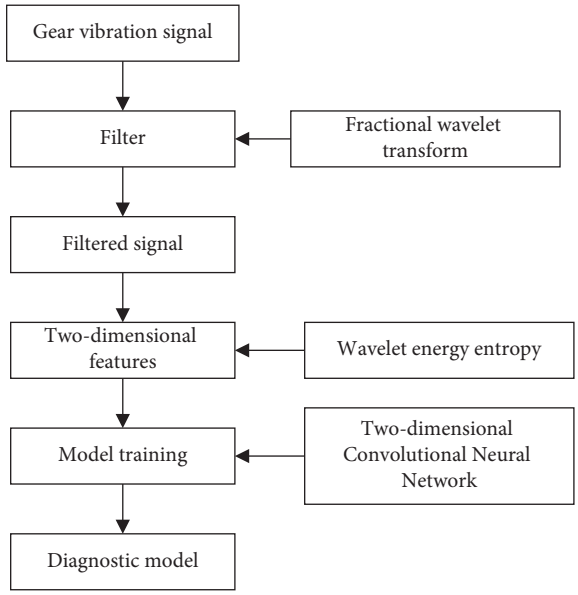


Figure 3: Fault diagnosis flowchart of planetary gear.

planetary gears, rolling bearings, speed control equipment, etc. The fault components used in the entire test platform include planetary gear pitting fault, broken tooth fault, wear fault, and other faults. The sampling frequency of the experimental data is 5120 Hz. A total of 10 fault samples under fault states were selected, 1700 samples were selected for each fault, and the number of sample points for each sample was 512. Among them, the ratio of the number of training samples to the number of test samples is 10 : 7, and the specific sample classification is shown in Table 1.

4.2. FRWT Filter Analysis

4.2.1. Determination of the Optimal Order of FRWT. First, the minimum output energy is used as the objective function to optimize the optimal order. The order optimization process of pitting fault reconstruction signal, broken tooth fault reconstruction signal, and wear fault reconstruction signal is shown in Figure 4. It can be clearly seen from Figure 4 that the minimum value of the FRFT output energy of the pitting fault (Dianshi880-1) reconstructed signal is 11670, and the corresponding order is 1.57; that is, the best order is 1.57; in the broken tooth fault (Duanchi1500) the minimum value of the FRFT output energy of the reconstructed signal is 20020, and the corresponding order is 1.646; that is, the best order is 1.646; the minimum value of the FRFT output energy of the reconstructed signal FRFT for wear fault (Mosun880-1) is 50360. At this time, the corresponding order is 1.558; that is, the best order is 1.558. The output energy value and the corresponding optimal order of the remaining faults are shown in Table 2.

4.2.2. Determination of the Number of FRWT Wavelet Bases and Decomposition Layers. In the fractional wavelet transform, when the selected wavelet base and the number of decomposition layers are different, the noise reduction effect of the signal will be different. Therefore, the wavelet bases are selected as db1 ~ db4 and sym1 ~ sym4, respectively, and the number of decomposition levels is 1 to 5, and the optimal wavelet base and decomposition level are selected by calculating the output signal-to-noise ratio (SNR) of the denoising signal. The specific results are shown in Figure 5–Figure 7. The wavelet basis and decomposition layer settings for each fault are shown in Table 3.

4.2.3. FRWT and FRFT Filtering Effect Analysis. The pitting fault reconstruction signal, wear fault reconstruction signal, and broken tooth fault reconstruction signal are, respectively, subjected to FRFT filtering and FRWT filtering, and the filtering results of each fault signal are shown in Figure 8, Figure 9, and Figure 10. Using the found optimal fractional order $p = 1.57$, 1.646, and 1.558, the pitting reconstruction signal, broken tooth reconstruction signal, and wear reconstruction signal are, respectively, subjected to fractional Fourier transform filtering. The filtering results are shown in Figure 8(b), as shown in Figure 9(b) and Figure 10(b), and then perform wavelet transform on the signal after the fractional Fourier transform in the corresponding fractional domain and finally carry out the transformed signal, $p = -1.57$, -1.646, -1.558-order fractional Fourier transform to obtain the corresponding output signal time domain diagram as shown in Figure 8(c), Figure 9(c), Figure 10(c).

In order to analyze the influence of the fractional order on the signal filtering effect, this paper calculates the output signal-to-noise ratio of the two filtering methods, respectively. The input signal-to-noise ratio of the pitting fault (Dianshi880-1) signal is -12.25 dB; the broken tooth fault (the input signal-to-noise ratio of Duanchi1500) signal is -13.15 dB; the input signal-to-noise ratio of wear fault (Mosun880-1) signal is -16.47 dB. The comparison result is shown in Figure 11.

It can be seen from Figure 11 that the output signal-to-noise ratios (SNR) of FRWT for pitting faults, wear faults, and broken teeth faults are all greater than the output signal-to-noise ratio (SNR) of FRFT. According to the larger output signal-to-noise ratio (SNR), the signal will be distorted. The smaller the degree and the noise interference, the better the filtering effect of FRWT compared to the filtering effect of FRFT.

4.3. Wavelet Packet Extraction Features.

Set the decomposition level of the wavelet packet to 8, which will generate a total of 256 frequency bands, and use the wavelet basis *db* 3 to decompose the fault vibration signals of 10 gears into eight layers, and generate a total of 256 wavelet packet components. Then use Shannon entropy to extract the wavelet energy, and then process the energy of the frequency band, that is, obtain the sum of the norm squares of each node of each layer of neurons, and finally obtain the normalized energy amplitude of each node. The corresponding normalized energy value of each frequency band is shown in Figure 12.

Table 1: Sample set.

Data set name	Fault type	Number of training set samples	Number of test machine samples	Label
Mosun880	Wear	1000	700	1
Mosun880-1	Wear	1000	700	2
Mosun880-2	Wear	1000	700	3
Mosun880-3	Wear	1000	700	4
Dianshi880	Pitting	1000	700	5
Dianshi880-1	Pitting	1000	700	6
Dianshi880-2	Pitting	1000	700	7
Dianshi880-3	Pitting	1000	700	8
Duanchi1500	Broken tooth	1000	700	9
Normal880	Normal	1000	700	10

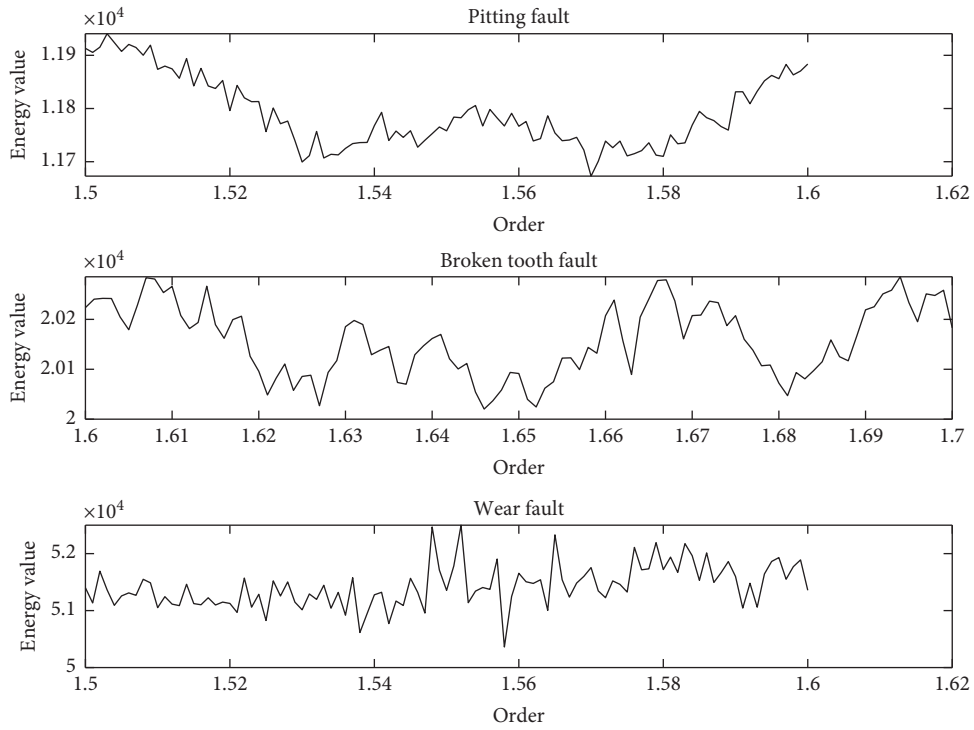


Figure 4: FRWT order optimization.

Table 2: Fault input signal-to-noise ratio, FRWT optimal order, and energy value.

Fault type	Enter SNR (db)	Best order	Minimum energy
Mosun880	−13.58	1.1	35250
Mosun880-2	−20.65	1.42	62135
Mosun880-3	−25.47	1.25	70259
Dianshi880	−10.02	1.291	9875
Dianshi880-2	−15.63	1.432	13026
Dianshi880-3	−20.87	1.45	16548
Normal880	−10.25	1.31	7458
Dianshi880-2	−15.63	1.432	13026

Among them, the characteristics of each sample are 256 frequency band energy spectra, and the characteristics of 1700 samples are converted into a matrix form with a two-dimensional form with a dimension of $16 * 16$. Figure 13 shows the converted two-dimensional frequency band energy characteristic distribution. Finally, the One – hot code is used to set the label category for each type of fault.

4.4. Training and Classification of Fault Models. The specific parameter selection for experimental verification is as follows: the number of layers of the two-dimensional CNN network is set to 6 layers, the convolutional layer and the pooling layer are each two layers, the number of convolution kernels in the first layer is 8, and its size is $3 * 3$. The number of convolution kernels in the second layer is 16, and its size is

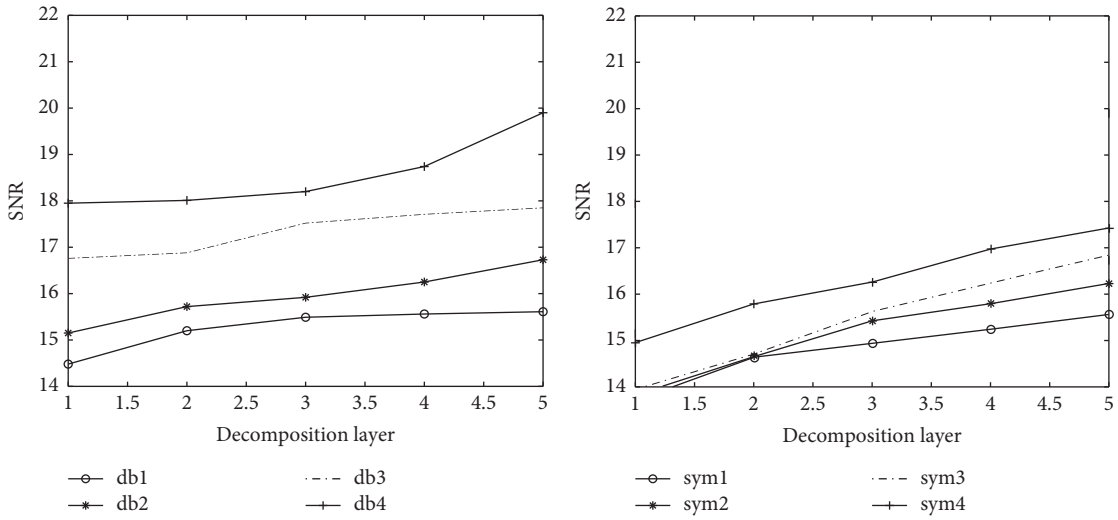


Figure 5: The relationship between the signal-to-noise ratio of the pitting signal and the wavelet basis and the number of decomposition layers.

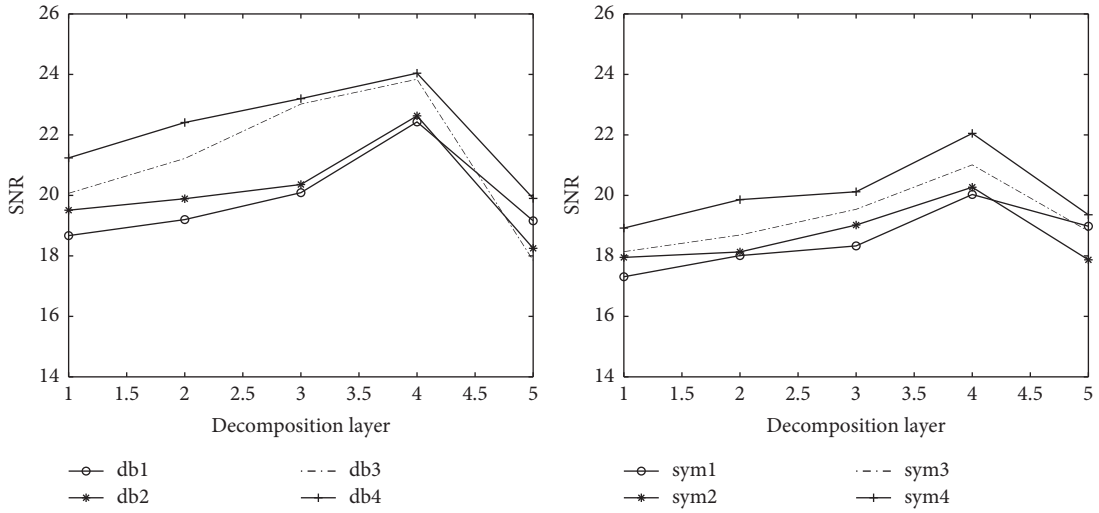


Figure 6: The relationship between the signal-to-noise ratio of the broken tooth signal and the wavelet basis and the number of decomposition layers.

$4 * 4$; the batch size is 10, and the maximum number of iterations is 1500; the pooling layer uses the maximum pooling method, and its size is $2 * 2$; using Dropout regularization reduces overfitting. Extract the wavelet energy values of the signals after FRFT filtering and FRWT filtering to construct a two-dimensional feature matrix as input; randomly select 1000 samples of each type of fault as the training set for model training, and 700 samples as the test set for the two-dimensional convolutional neural. The training model of the network is verified, and the training error curve is shown in Figure 14.

From the analysis in Figure 14, it can be seen that, regardless of whether the fractional Fourier transform or the fractional wavelet transform is used, when the number of iterations is less than or equal to 120, the training error of the two is equal; when the number of iterations is 120, the training error is 0.6667. The effect is extremely poor; when the number of iterations is greater than 120, the training error of the fractional wavelet transform filtering signal is obviously smaller than the training error of the fractional Fourier transform filtering signal; when the number of iterations is 1500, the training error of the fractional wavelet transform filtering method is 0.01623, and the training error of the fractional Fourier transform filtering method is 0.06514, that is; the training error of the fractional wavelet transform filter signal is significantly smaller than the training error of the fractional Fourier transform filter signal. It can be seen that the training effect of FRWT+2D-CNN is better than that of FRWT+2D-CNN.

The classification results of each fault in the test set using the FRWT+2D-CNN and FRFT+2D-CNN models are shown in Figure 15 and Figure 16. The abscissa is the

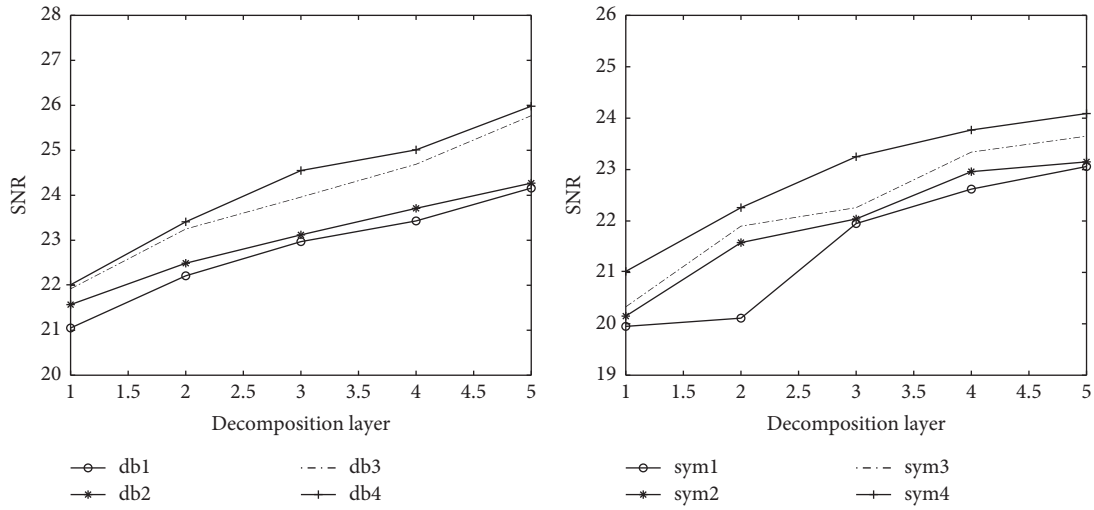


Figure 7: The relationship between the signal-to-noise ratio of the wear signal and the wavelet basis and the number of decomposition layers.

Table 3: Wavelet basis and decomposition layer settings for each fault.

Fault type	Wavelet base	Decomposition layer
Mosun880	db4	4
Mosun880-2	db4	4
Mosun880-3	db4	4
Dianshi880	db4	5
Dianshi880-2	db4	5
Dianshi880-3	db4	5
Normal880	db3	4

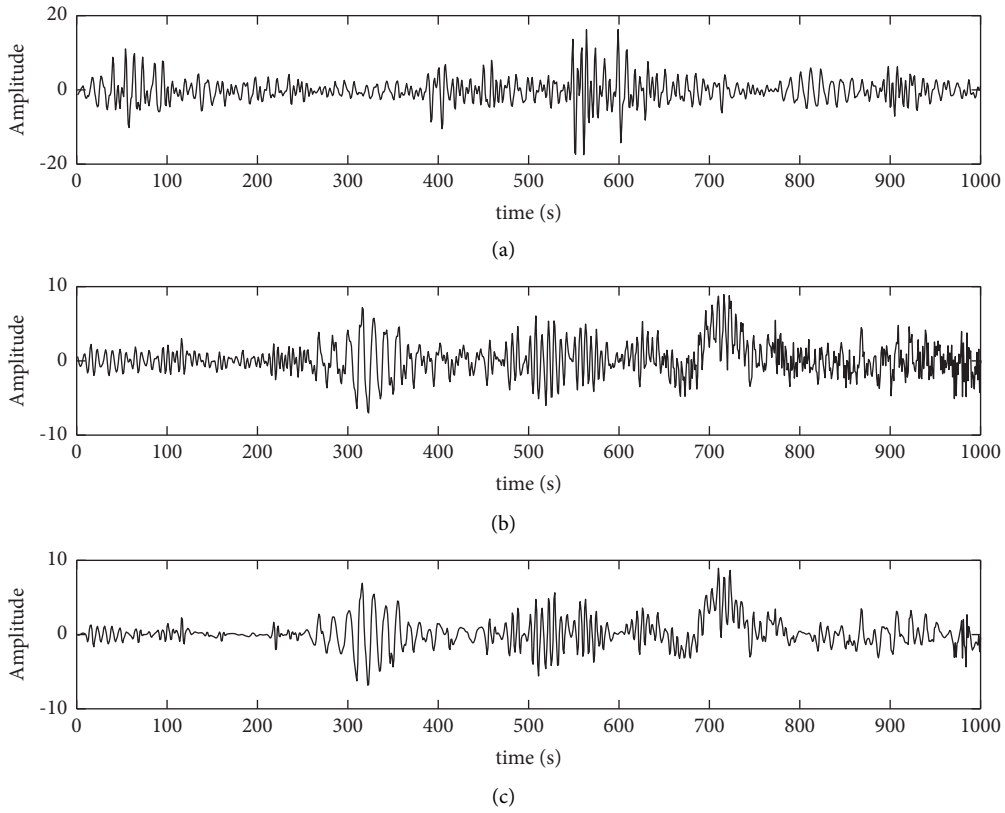


Figure 8: Time domain diagram of pitting signal filtering. (a) Pitting corrosion reconstruction signal. (b) FRFT filtered signal. (c) FRWT filtered signal.

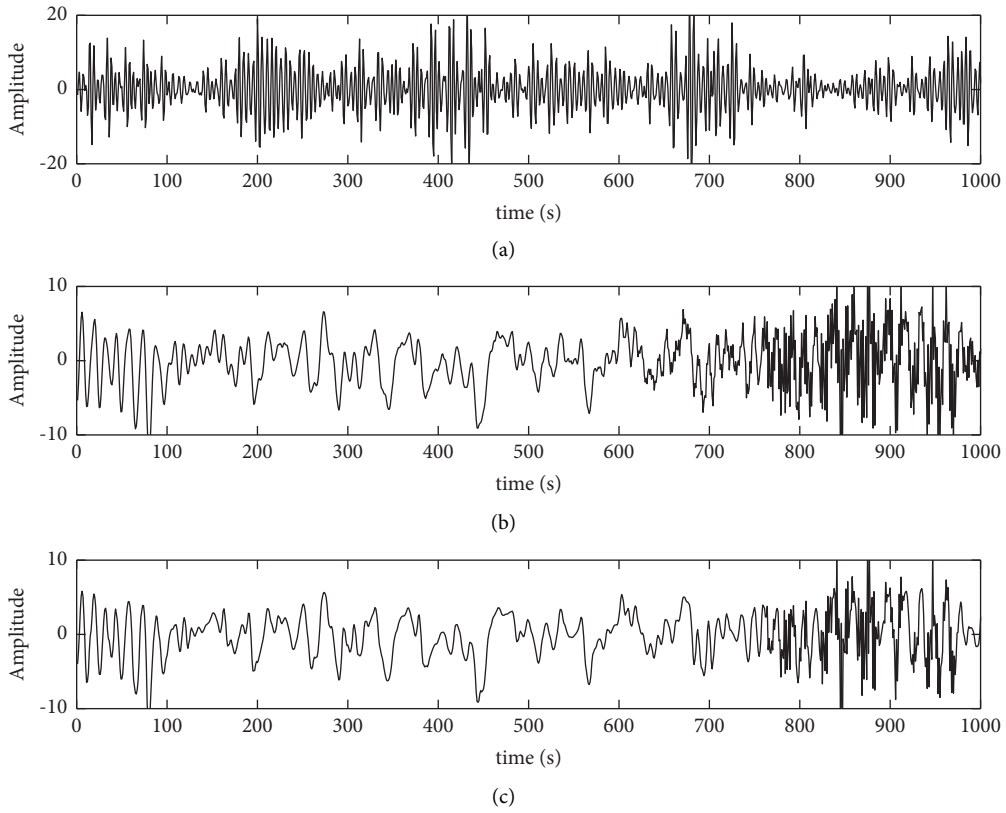


Figure 9: Time domain diagram of broken tooth signal filtering. (a) Tooth reconstruction signal. (b) FRFT filtered signal. (c) FRWT filtered signal.

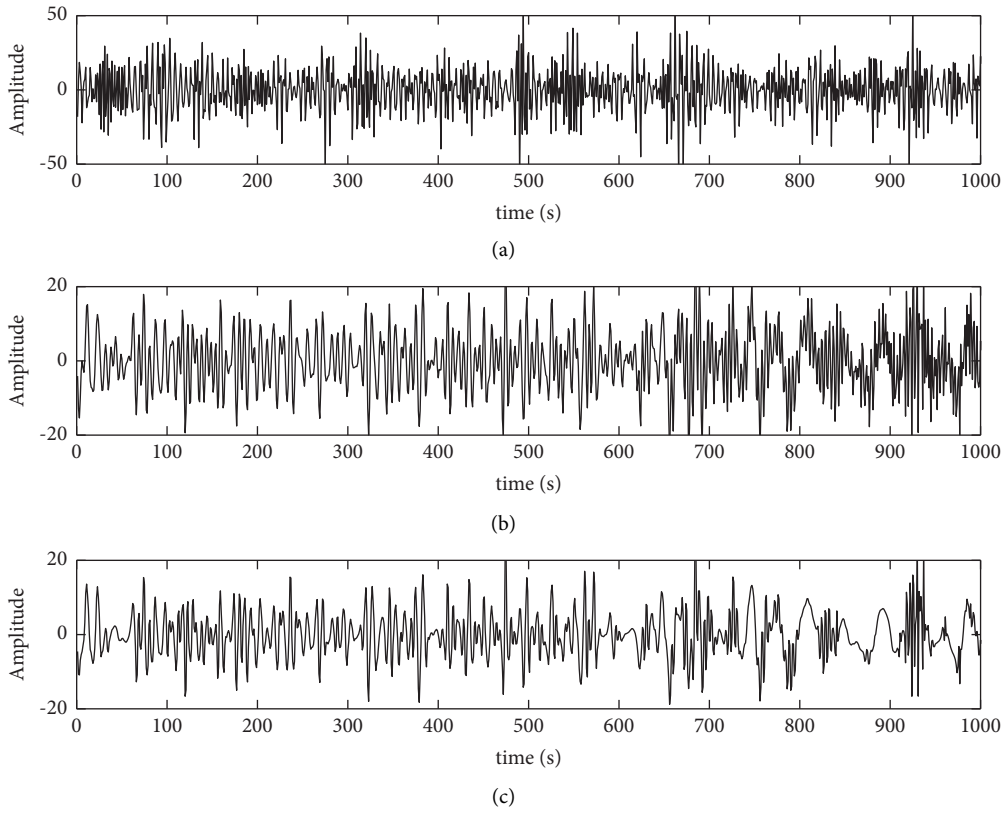


Figure 10: Time domain diagram of wear signal filtering. (a) Wear reconstruction signal. (b) FRFT filtered signal. (c) FRWT filtered signal.

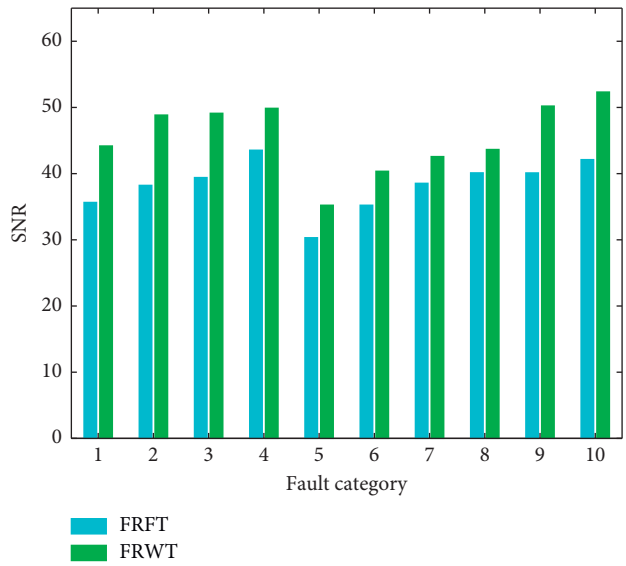


Figure 11: FRFT and FRWT filter output signal-to-noise ratio (SNR).

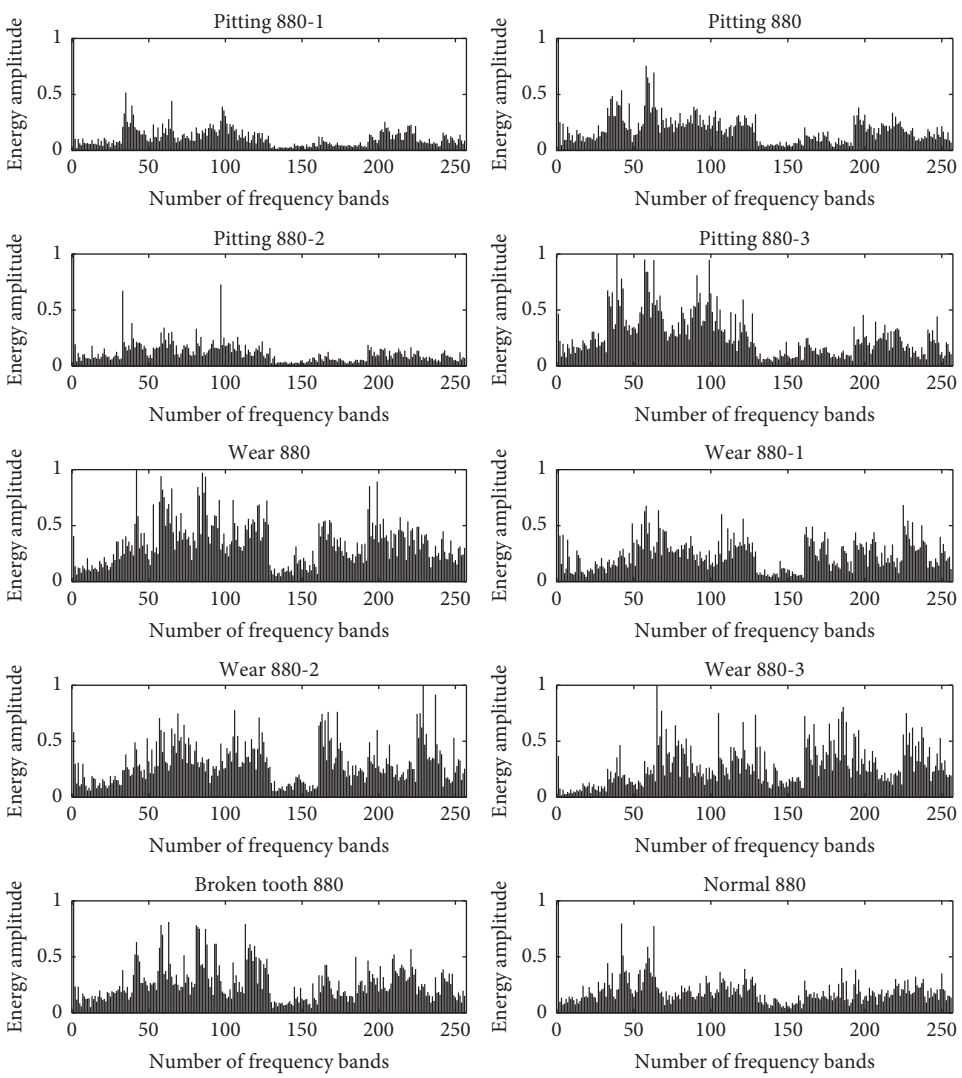


Figure 12: The relationship between the number of frequency bands of each fault and the normalized energy value.

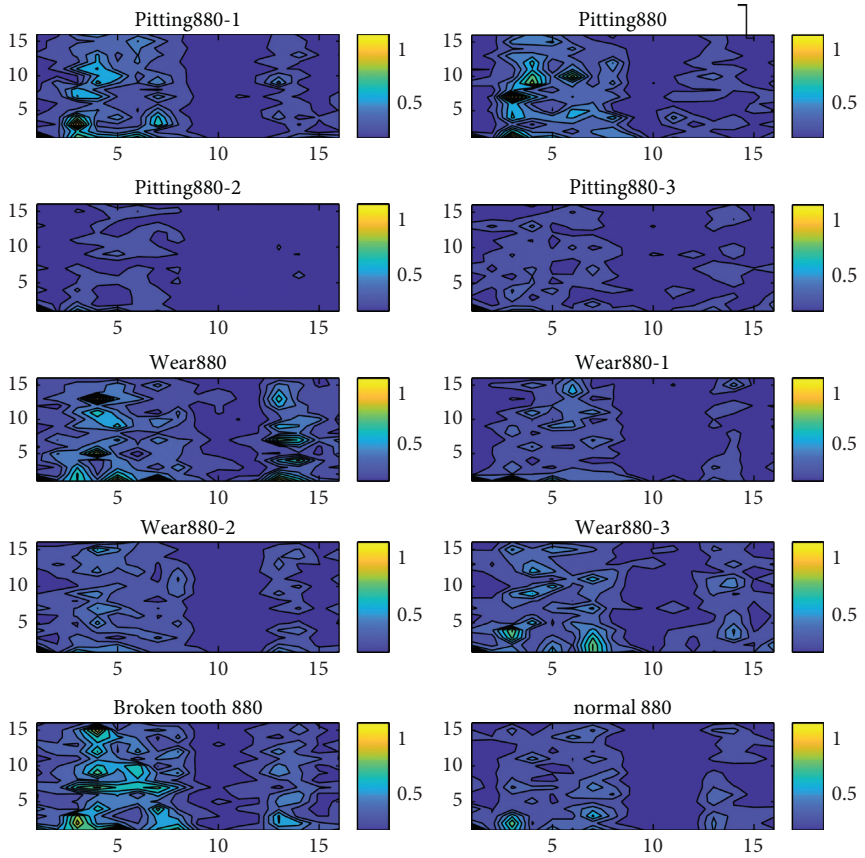


Figure 13: Feature distribution of some samples of the FRWT filtered signal.

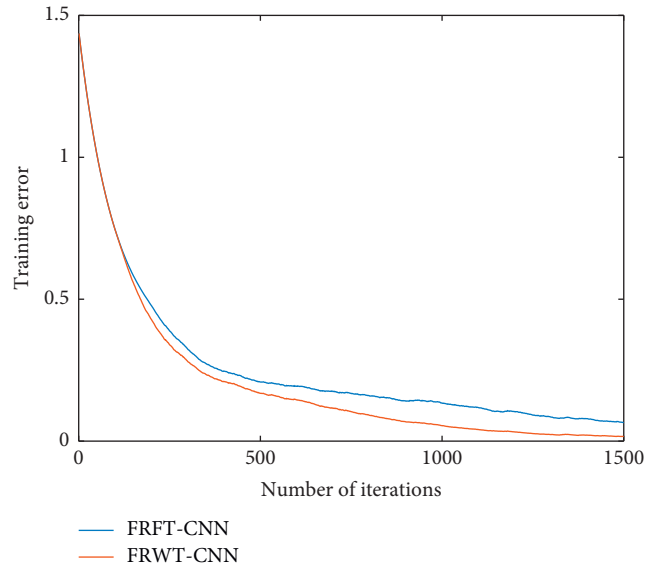


Figure 14: The relationship between model training error and training times.

predicted category label of the test set; the ordinate is the actual label category of the test set; the value of the diagonal position is the classification accuracy of each of the 10 types of faults; the position outside the diagonal is the type of fault.

Comparing Figures 15 and 16, it can be found that when FRWT+2D-CNN classifies and recognizes faults, only two samples are misclassified; that is, type 3 faults are misclassified as type 4 faults, and type 7 faults are wrong. The fault is classified as the 8th type of fault; when FRFT+2D-

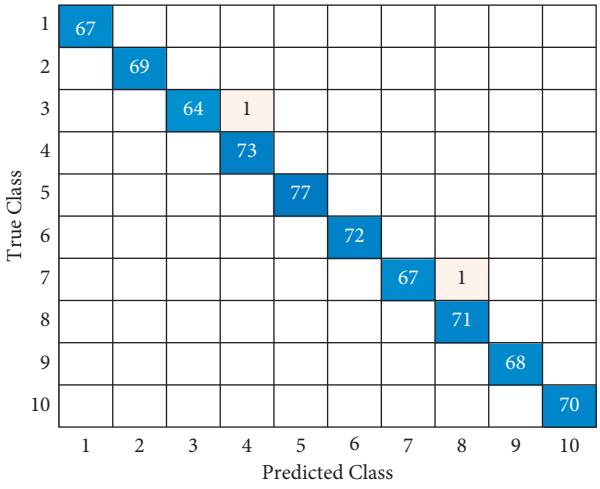


Figure 15: FRWT+2D-CNN test set classification label.

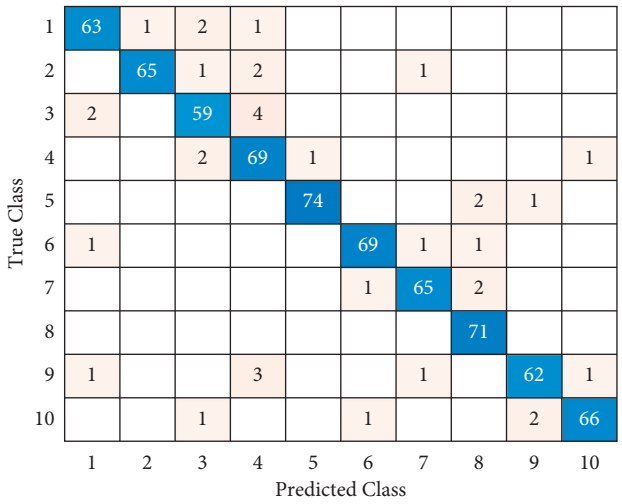


Figure 16: FRFT+2D-CNN test set classification label.

CNN classifies and recognizes the fault, except for the 5th type of fault, the other fault types are all misclassified, and the number of misclassified samples for each type of fault is greater than 1. Therefore, when classifying faults based on FRWT+2D-CNN, each type of fault can be accurately identified to the greatest extent possible.

In order to fully verify the stability and accuracy of the diagnosis method proposed in this article, this article randomly conducts 15 simulation tests on the two diagnosis methods (FRFT+2D-CNN, FRWT+2D-CNN), and the classification accuracy of each test is as shown in Figure 17. The average accuracy of the diagnosis models of the two classification methods is shown in Table 4.

By analyzing Figure 5, 20, it can be seen that the classification accuracy of the two diagnostic methods FRFT+2D-CNN and FRWT+2D-CNN basically remains stable, and the classification accuracies of FRFT+2D-CNN and FRWT+2D-CNN are both within 3%. With fluctuations up and down, from a macroperspective, the classification accuracy of

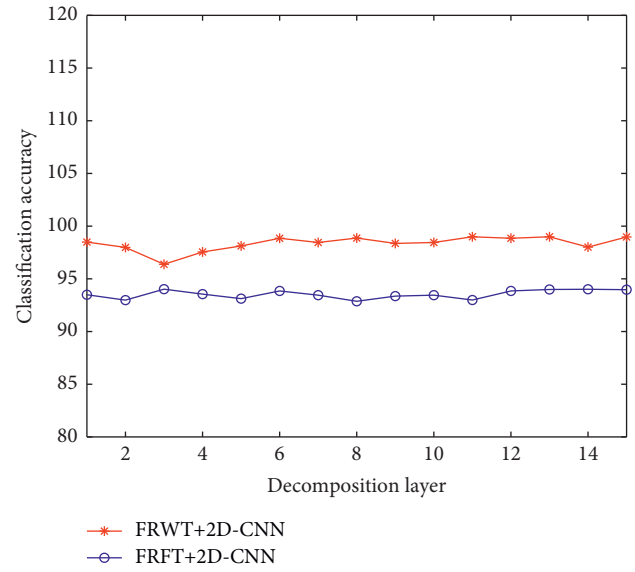


Figure 17: Test accuracy of the two methods.

Table 4: The average accuracy of classification by different methods.

Diagnosis method	Training samples	Test sample	Number of experiments	Average accuracy
FRFT+2D-CNN	1000	700	15	93.53
FRWT+2D-CNN	1000	700	15	98.36

FRWT+2D-CNN is higher than that of FRFT+2D-CNN. It can be seen from Table 4 that when the number of training samples, the number of test samples, and the number of trials are equal, the average accuracy of FRWT+2D-CNN classification is higher than the average accuracy of FRFT+2D-CNN classification. Therefore, in the fault diagnosis of planetary gears, the classification method of FRWT+2D-CNN is obviously better than the fault classification method of FRFT+2D-CNN.

5. Conclusion

(1) This paper adopts the FRWT-based planetary gear vibration signal filtering method. The simulation results show that both the fractional wavelet transform and the fractional Fourier transform can achieve the denoising effect of the signal; the denoising effect of the fractional wavelet transform is better than fractional Fourier transform: The energy-based fractional Fourier transform algorithm is better than the peak search-based fractional Fourier transform algorithm.

(2) This paper adopts a two-dimensional convolutional neural network model, and the signals after the fractional Fourier transform and the fractional wavelet transform are filtered, and the one-dimensional wavelet energy value is normalized and

converted into a two-dimensional feature matrix for diagnosis model training. The simulation results show that the two-dimensional convolutional neural network can effectively realize fault classification and recognition. In addition, the accuracy of planetary gear fault classification based on FRWT and 2D-CNN is better than the accuracy of planetary gear fault classification based on FRFT and 2D-CNN.

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

This work was supported by the National Natural Science Foundation of China (Grant no. 61973041) and the National Key Research and Development Program of China (Grant no. 2019YFB1705403).

References

[1] J. Yu, L. Gao, G. Yu, K. Liu, and Y. Guo, "Planetary gear fault identification based on SDAE and GRUNN," *Vibration and Shock*, vol. 40, no. 2, pp. 156–163, 2021.

[2] H. Gao and A. Fan, "Research on the fault of planetary gearbox based on CEEMD and CPSO-ELM," *Automation Technology and Application*, vol. 40, no. 1, pp. 95–100, 2021.

[3] Z. Wang, L. Yao, X. Qi, J. Zhang, and J. Zheng, "Fault diagnosis of planetary gearbox based on parameter optimization variational modal decomposition and multi-domain manifold learning," *Vibration and Shock*, vol. 40, no. 1, pp. 110–118, 2021.

[4] H. Li, Z. Qi, and J. Hu, "Research on intelligent judgment method of planetary gearbox tooth surface wear fault based on FFT-DBN," *Measurement & Control Technology*, vol. 39, no. 12, pp. 50–54, 2020.

[5] Y. Li, Z. Jiang, H. Liang, and X. Xu, "Fault diagnosis of planetary gearbox based on EEMD symmetric differential energy spectrum," *Modular Machine Tool and Automatic Manufacturing Technology*, vol. 47, no. 10, pp. 122–125, 2020.

[6] Q. Zhang, H. Li, and S. Li, "Application of Time-Frequency Features and PSO-SVM in Fault Diagnosis of Planetary Gearbox," in *Proceedings of the 2021*, Springer Science and Business Media Deutschland GmbH, vol. 1274, Xi'an, China, September 2021.

[7] Z. Wang, H. Huang, and Y. Wang, "Fault diagnosis of planetary gearbox using multi-criteria feature selection and heterogeneous ensemble learning classification," *Measurement: Journal of the International Measurement Confederation*, vol. 173, p. 173, 2021.

[8] X. Zhang, L. Lin, S. Liu, and J. Lei, "Empirical wavelet transform based on energy peak location and its application in weak bearing fault diagnosis," *Journal of Xi'an Jiaotong University*, vol. 12, no. 8, pp. 1–9, 2021.

[9] X. Li, Z. Ma, D. Kang, and X. Li, "Fault diagnosis for rolling bearing based on VMD-FRFT," *Measurement: Journal of the International Measurement Confederation*, vol. 155, 2020.

[10] W. Liu, "Digital watermarking algorithm based on fractional discrete wavelet transform," *Shandong Industrial Technology*, vol. 4, no. 9, p. 164, 2019.

[11] S. K. Upadhyay and K. Khatterwani, "Fractional wavelet transform through heat equation," *Journal of Thermal Stresses*, vol. 42, no. 11, pp. 1386–1414, 2019.

[12] Y. Ding, Y. He, B. Li, and J. Cui, "Inverter fault diagnosis based on wavelet packet and quantum neural network," *Journal of Chongqing University of Technology (Natural Science)*, vol. 35, no. 4, pp. 152–158, 2021.

[13] J. Xu, *Research on Bearing Fault Diagnosis Method of Wind Turbine Based on Convolutional Neural Network*, Zhejiang University, HangZhou, China, 2020.

[14] X. Xiong, J. Wang, Y. Zhang, Q. Guo, and S. Zong, "A two-dimensional convolutional neural network optimization method for bearing fault diagnosis," *Proceedings of the Chinese Society for Electrical Engineering*, vol. 39, no. 15, pp. 4558–4568, 2019.

[15] C. Huang, *Research on Optimal Method of Rolling Bearing Fault Diagnosis Combining Time-Frequency Analysis and Convolutional Neural Network*, Zhejiang University, HangZhou, China, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9966078, 11 pages
https://doi.org/10.1155/2021/9966078

Research Article

Weak Fault Feature Extraction of Rolling Bearing Based on SVMD and Improved MOMEDA

Xinyu Wang and Jie Ma

Mechanical Electrical Engineering School, Beijing Information Science & Technology University, Beijing 100192, China

Correspondence should be addressed to Jie Ma; mjbeijing@163.com

Received 26 June 2021; Accepted 21 October 2021; Published 28 December 2021

Academic Editor: Yong Chen

In order to solve the problem that it is very difficult to extract fault features directly from the weak impact component of early fault signal of rolling bearing, a method combining continuous variational mode decomposition (SVMD) with modified MOMEDA based on Teager energy operator is proposed. Firstly, the low resonance impulse component in the fault signal is separated from the harmonic component and noise by SVMD, and then the Teager energy operator is used to enhance the impulse feature in the low resonance component to ensure that the accurate fault period is selected by the MOMOEDA algorithm. After further noise reduction by MOMEDA, the envelope spectrum of the signal is analyzed, and finally the fault location is determined. The results of simulation and experimental data show that this method can accurately and effectively extract the characteristic frequency of rolling bearing weak fault.

1. Introduction

Rolling bearing is one of the important parts of rotating machinery, which ensures the working accuracy of the shaft [1]. When the bearing is partially damaged, the vibration signal will produce periodic impact characteristics, which will affect the normal operation of the equipment. Therefore, accurate and timely identification of bearing faults can effectively ensure the safety of equipment operation. However, early fault vibration signals often have nonlinear and nonstationary characteristics, the impact component of fault characteristics is easy to be submerged in strong background noise, and it is difficult to judge the fault type directly from time domain or frequency domain [2, 3]. Therefore, the early fault diagnosis of rolling bearing has always been a research difficulty and hot spot.

Empirical mode decomposition (EMD) can express nonlinear and nonstationary signals as the sum of physically averaged time-frequency components, which has been applied satisfactorily in many applications [4]. However, the results of EMD are highly influenced by the searching of extremum and interpolation methods. Different from EMD, variational mode decomposition (VMD) is a nonrecursive decomposition method, which compresses different frequency bands around different center frequencies [5]. However, both VMD and its improved algorithm are based on the assumption that the modal component is narrow band. In reference [6], an extension of VMD is given to decompose the signal containing broadband nonlinear frequency modulation component, that is, variational nonlinear frequency modulation mode decomposition (VNCMD). It is a method to transform the wideband Nonlinear FM component into narrow band signal for analysis by demodulation technology [6]. Recently, a new method, variational mode extraction (VME), is proposed to extract the intrinsic mode function by knowing the approximate central frequency of the intrinsic mode function [7]. Later, continuous variational mode decomposition (SVMD) algorithm is extended to VME, which is an efficient and fast adaptive signal variational decomposition method. This new decomposition method extracts all the modal components (IMF) in a continuous way, does not need to know the number of modes, and has low computational complexity [8].

In addition, the transmission process of the original fault signal of rolling bearing can be regarded as a linear convolution mixing process of the original signal and the channel, and the extraction of the original fault impact signal can be regarded as a deconvolution process. From this point of view, Wiggins [9] proposed minimum entropy deconvolution (MED) algorithm for the first time, which uses kurtosis maximization to recover the pulse like estimation of the original vibration signal. However, the results of MED cannot reflect the real situation of bearing failure. Later, Wiggins [10] and others improved MED and proposed a maximum correlation kurtosis deconvolution algorithm (MCKD). Based on correlation kurtosis, we can measure the impulsivity of signals correlated with a given period, but improper selection of parameters will lead to poor estimation of correlation kurtosis. An optimal minimum entropy adjusted deconvolution algorithm (MOMEDA) is proposed by Cabrelli [11], which proves that the output returned by OMEDA has a simpler structure. However, OMEDA only tends to deconvolute a single pulse effectively. Therefore, McDonald and Zhao [12] proposed an improved algorithm of OMEDA, which overcomes the limitations of MED, MCKD, and OMEDA by introducing the target vector to deconvolute the ideal periodic pulse sequence. However, when the early fault signal is directly denoised by MOMEDA algorithm, it is easy to be interfered by strong background noise.

Based on the above analysis, using Teager energy operator has strong antinoise ability and can enhance the property of transient impact [13–15]; using MOMEDA to highlight fault impact features from Teager energy operator can extract weak fault frequency more effectively. The SVMD algorithm is combined with the improved MOMEDA algorithm, and the simulation results are compared with other diagnosis methods to verify the effectiveness and practicability of the proposed method.

2. Fundamental Theory

2.1. SVMD. SVMD algorithm can be regarded as a continuous implementation of VMD and an extension of VME. In this method, some criteria are added to the VMD algorithm to ensure that the latest modal components are different from those previously found. The biggest advantage of SVMD algorithm is that it does not need to know the number of available modal components in the signal, but it is a key parameter for VMD algorithm.

In this algorithm, VME is applied to decompose the signal in turn, and some constraints are added to avoid converging to the previously extracted modal components. This process will continue until all modal components are extracted or the reconstruction error (the error between the sum of the input signal and the modulus) is less than the threshold. Mathematically, it is assumed that the input signal $f(t)$ is decomposed into two signals: the Lth modal component $u_L(t)$ and the residual signal $f_r(t)$, as follows:

$$f(t) = u_L(t) + f_r(t), \tag{1}$$

where $f_r(t)$ is an input signal other than $u_L(t)$ and contains two parts: the sum of the previously obtained moduli and the unprocessed part of the signal, namely,

$$f_r(t) = \sum_{i=1:L-1} u_i(t) + f_u(t). \tag{2}$$

Obviously, in order to find the first modal component, the first part of $f_r(t)$ (the sum of the modal components obtained previously) is zero. An iteration of SVMD can be completed by the following equation:

$$\widehat{u}_L^{n+1}(\omega) = \frac{\widehat{f}(\omega) + \alpha^2(\omega - \omega_L^n)^4 \widehat{u}_L^n(\omega) + \widehat{\lambda}(\omega)/2}{\left[1 + \alpha^2(\omega - \omega_L^n)^4\right]\left[1 + 2\alpha(\omega - \omega_L^n)^2 + \sum_{i=1}^{L-1} 1/\alpha^2(\omega - \omega_i)^4\right]}, \tag{3}$$

where α denotes the equilibrium parameters of data fidelity constraints, which can be solved by the Lagrange multiplier method. As explained in [5], α the value of α is usually very large. Therefore, the equation of updating ω. can be approximately expressed as

$$\omega_L^{n+1} = \frac{\int_0^\infty \omega \left|\widehat{u}_L^{n+1}(\omega)\right|^2 d\omega}{\int_0^\infty \left|\widehat{u}_L^{n+1}(\omega)\right|^2 d\omega}. \tag{4}$$

Finally, Lagrange multiplier is given λ. The modified equation of the method is obtained by the double rising method

$$\widehat{\lambda}^{n+1} = \widehat{\lambda}^n + \tau\left[\widehat{f}(\omega) - \left(\widehat{u}_L^{n+1}(\omega) + f_u^{n+1}(t) + \sum_{i=1}^{L-1} u_i^{n+1}(\omega)\right)\right], \tag{5}$$

where τ represents the update parameter.

$$\begin{cases} y(\omega) = \dfrac{\alpha^2\left(\omega - \omega_L^{n+1}\right)^4\left(\widehat{f}(\omega) - \widehat{u}_L^{n+1}(\omega) - \sum_{i=1}^{L-1}\widehat{u}_i(\omega) + \widehat{\lambda}(\omega)/2\right) - \sum_{i=1}^{L-1}\widehat{u}_i(\omega)}{1 + \alpha^2\left(\omega - \omega_L^{n+1}\right)^4}, \\ \widehat{\lambda}^{n+1} = \widehat{\lambda}^n + \tau\left[\widehat{f}(\omega) - \left(\widehat{u}_L^{n+1}(\omega) + y(\omega) + \sum_{i=1}^{L-1} u_i^{n+1}(\omega)\right)\right]. \end{cases} \tag{6}$$

The SVMD algorithm extracts modal components one by one until the reconstruction error reaches a certain threshold. The optimization problem in SVMD can be approximately regarded as a k-dimensional optimization problem at each frequency, while VMD is a k-dimensional optimization problem [8], which further shows that SVMD has lower computational complexity.

2.2. Teager Energy Operator. Teager energy operator is a kind of nonlinear differential operator, which can amplify the transient energy component of the signal by nonlinear combination of the instantaneous value and its differential of the vibration signal, which can more highlight the instantaneous characteristics of the impact [13].

For continuous time signals $x(t) = a(t)\cos[\Phi(t)]$, Teager energy operator is defined as

$$\psi[x(t)] = \left[\dot{x}(t)^2\right] - x(t)\ddot{x}(t), \tag{7}$$

where $\dot{x}(t)$ and$\ddot{x}(t)$ are the first and second derivatives of signal $x(t)$.

If $x(t)$ is a discrete signal, the differential is replaced by difference, and the Teager energy operator is approximately defined as

$$\psi[x(n)] = [x(n)]^2 - x(n+1)x(n-1). \tag{8}$$

It can be seen from equation (8) that for discrete-time signals, the Teager energy at the current time can be calculated as long as the samples at the current time and before and after the time are known.

The output of Teager energy operator is the product of the instantaneous amplitude and the square of the instantaneous frequency, which increases the product of the square of the frequency compared with the traditional energy definition [14]. Since the vibration frequency of transient shock is high, the output of Teager energy operator can effectively enhance the transient shock component.

Aiming at the problem of rolling bearing fault feature extraction under strong background noise, the instantaneous Teager energy operator can be used, which has the advantages of good time resolution and adaptive ability to the transient change of signal, highlighting the impact fault feature of bearing [14].

2.3. MOMEDA. The essence of MOMEDA is to recover the impulse characteristics of the original signal by finding an optimal filter f in a noniterative way, so as to minimize the impact of noise on the extracted impulse signal, which is a deconvolution process [16]. The process of shock signal transmission from signal source to sensor can be expressed as follows:

$$x = h * y + e, \tag{9}$$

where y expresses the bearing fault impact signal; h expresses the system transfer function; x expresses the original vibration signal collected by sensor; and e expresses the random noise.

The deconvolution process is as follows:

$$y = f * x = \sum_{k=1}^{N-L} f_k x_{k+L-1}, \tag{10}$$

where k expresses the total sampling points, $k = 1, 2, ..., n-1$, and L expresses the filter length.

For the characteristics of periodic pulse signals of rotating machinery, the method of MOMEDA defines multipoint D-norm based on D-norm:

$$\mathrm{MDN}(y,t) = \frac{1}{\|t\|}\frac{t^T y}{\|y\|}, \tag{11}$$

$$\mathrm{MOMEDA}(y,t) = \max_f \mathrm{MDN}(y,t) = \frac{1}{\|t\|}\frac{t^T y}{\|y\|}, \tag{12}$$

where y expresses the vibration signal vector; f expresses the filter vector bank; and t expresses the target vector, used to determine the pulse position and weight of deconvolution target.

When the fault impulse signal is completely coincident with the target vector t, the multipoint D-norm reaches the maximum, the corresponding filter f is the optimal filter, and the deconvolution effect is optimal.

The extremum of equation (12) is obtained by deriving the filter:

$$\begin{aligned} \frac{d}{df}\left(\frac{t^T y}{\|y\|}\right) &= \|y\|^{-1}\left(t_1 M_1 + t_2 M_2 + \cdots + t_k M_k\right) \\ &\quad - \|y\|^{-3} t^T y X_0 y = 0, \end{aligned} \tag{13}$$

where $f = f_1, f_2, \ldots, f_L; t = t_1, t_2, \ldots, t_{N-L}; M_k = \begin{bmatrix} x_{k+L-1} \\ x_{k+L-2} \\ \vdots \\ x_k \end{bmatrix}$.

By further simplifying equation (13), it can be concluded that

$$\frac{ty}{\|y\|^2}X_0 y = X_0 t, \tag{14}$$

where X_0 is the matrix form of pulse signal, $X_0 t = t_1 M_1 + t_2 M_2 + \ + t_k M_{k\circ}$.

Suppose that the inverse toplitz autocorrelation matrix $(X\,0X\,T\,0)^{-1}$ exists, and $y = X\,T\,0f$ is substituted into equation (14):

$$\frac{ty}{\|y\|^2} f = \left(X_0 X_0^T\right)^{-1} X_0 t. \tag{15}$$

From the above formula, we can take its special solution f as a group of optimal filters, which is the solution of MOMEDA:

$$f = \left(X_0 X_0^T\right)^{-1} X_0 t. \tag{16}$$

By substituting equation (16) into $y = X\,T\,0f$, the original shock signal can be recovered to the greatest extent.

2.4. Early Fault Feature Extraction of Rolling Bearing. The impact component of rolling bearing early signal is weak, so it is more difficult to extract fault features under strong background noise. Because the VMD method decomposes each modal component (IMF) at the same time, when the number of available modulus in the signal is high, the calculation time will increase significantly, and some of the decomposed modal components are regarded as interference or noise signals. SVMD is an algorithm for searching modal components step by step. This continuity helps to improve the convergence speed. At the same time, it does not extract unnecessary modal components and reduces the calculation time. Therefore, SVMD algorithm is better than VMD algorithm.

In addition, although MOMEDA algorithm has many advantages, it is easy to be interfered by strong background noise when extracting fault period directly from multipoint kurtosis spectrum of early fault signal, which leads to extracting noise period instead of fault period. Therefore, the Teager energy operator of the original signal is used to enhance the periodic fault impulse component in the signal, and then the fault period is extracted from the multipoint kurtosis spectrum of the Teager energy operator by using MOMEDA, which can achieve efficient noise reduction. The specific process is as follows:

(1) Reasonable parameters such as balance parameter, time step of double lifting, tolerance of convergence criterion, and type of stop criterion are set for SVMD algorithm, and a series of IMF components are obtained by decomposing fault signal with SVMD

(2) The sum of all components is used to reconstruct, and the Hilbert envelope demodulation analysis is used to determine whether the fault characteristic frequency can be extracted

(3) If the fault characteristic frequency cannot be extracted, the Teager energy operator of the reconstructed signal is calculated to highlight the continuous periodic impact component in the reconstructed fault signal

(4) Combined with the theoretical fault period, the reasonable period interval and filter length are set, and the noise reduction and feature enhancement of Teager energy operator are further carried out by using MOMEDA algorithm

(5) The Hilbert envelope is used to demodulate the above filtering signal, and the fault location is diagnosed by comparing with the theoretical fault characteristic frequency

The detailed fault diagnosis flow of the proposed method is shown in Figure 1.

Here, error rate η of the theoretical failure period δ and the actual failure cycle ω is introduced, as an index to measure the deviation between the actual cycle and the theoretical cycle. Error rate η is defined as

$$\eta = \frac{|\delta - \omega|}{\delta} \times 100\%. \tag{17}$$

3. Simulation Verification

In order to verify the noise reduction effect based on the combination of SVMD and improved MOMEDA algorithm, the following simulation signals are used to simulate the weak fault of rolling bearing:

$$\begin{cases} x(t) = s_0 e^{-2\pi\zeta f_n t} \sin\left(2\pi f_n \sqrt{1-\zeta^2}\, t\right), \\ y(t) = x(t) + n(t), \end{cases} \tag{18}$$

where s_0 expresses the displacement constant, $s_0 = 2$; ζ expresses the damping coefficient, $\zeta = 0.1$; FN expresses the natural frequency of bearing, $f_n = 2000$ Hz; $x(t)$ expresses the periodic shock component; $n(t)$ expresses the Gaussian white noise; the signal-to-noise ratio of the simulation signal is -13 dB; the known fault characteristic frequency $f = 100$ Hz; and the number of sampling points $N = 5120$.

The time-domain waveform of the signal shown in Figure 2 is obtained by simulation. Due to the heavy white Gaussian noise added to the impact signal, the original impact component has been completely submerged in the noise signal, and it is difficult to find obvious periodic impact characteristics.

It can be seen from the spectrum and envelope spectrum of the simulation signal in Figure 3 that the characteristic frequency and its frequency doubling of the impulse signal are affected by the frequency conversion modulation phenomenon and strong background noise, and the obvious fault characteristic frequency and its frequency doubling component cannot be found in the envelope spectrum.

From the above analysis, it can be seen that it is difficult to identify the rolling bearing fault simply by using envelope spectrum, so it is necessary to denoise the fault signal before

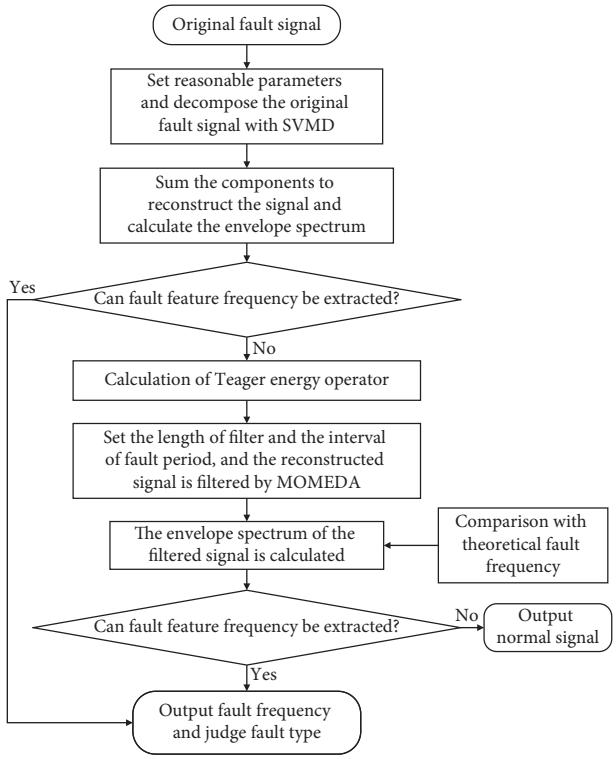


Figure 1: Fault diagnosis flow chart.

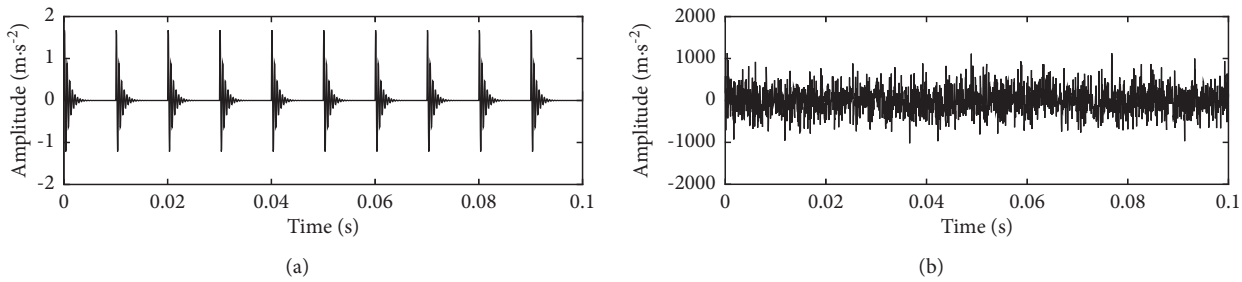


Figure 2: Time-domain waveform of simulation signal. (a) Signal without noise. (b) Signal with strong noise.

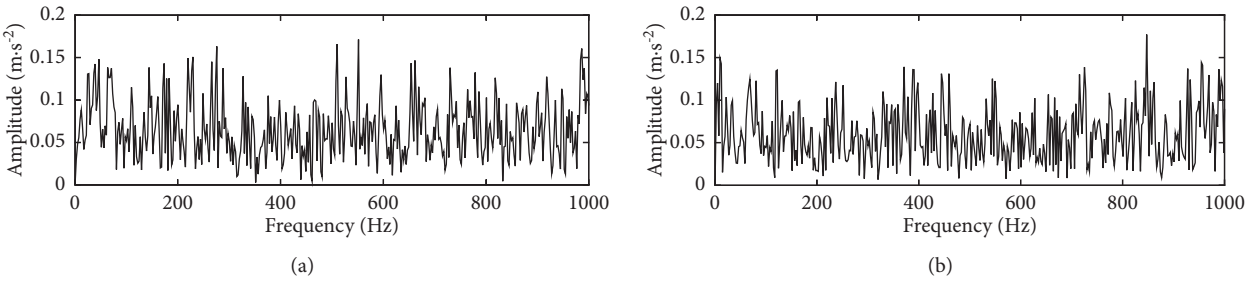


Figure 3: Frequency domain waveform of simulated signal. (a) Spectrum. (b) Envelope spectrum.

envelope analysis, so as to improve the signal-to-noise ratio and enhance the fault impact component. The simulation signal is decomposed by SVMD to separate weak fault features and noise interference frequency. First, set the maximum balance parameters $\alpha = 2000$. The obtained components (IMF1-IMF5) are shown in Figure 4. Since all the components extracted by SVMD are needed, the reconstructed signal and its envelope spectrum can be obtained by summation of each component signal directly, as shown in Figure 5.

The fault characteristic frequency is shown in Figure 5, which shows that SVMD decomposition has a certain effect.

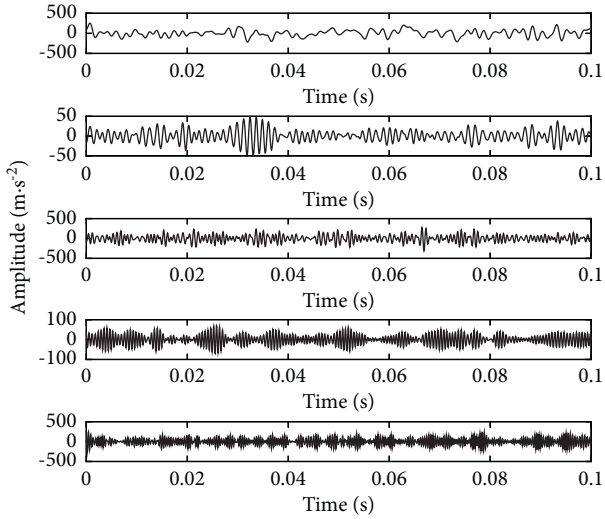


Figure 4: Decomposition process of SVMD.

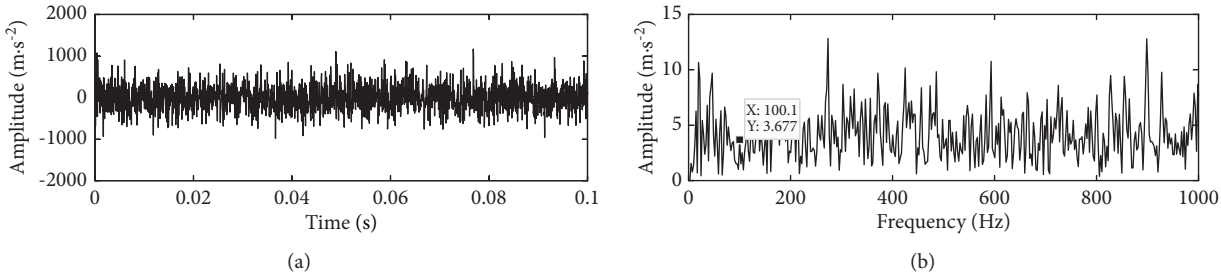


Figure 5: Reconstructed signal. (a) Time-domain waveform. (b) Envelope spectrum.

However, the amplitude of the interference frequency around the fault feature of the signal is also large, and its frequency doubling component cannot be found accurately, so it is necessary to further denoise the fault signal after decomposition and reconstruction.

The multipoint kurtosis spectrum and interval diagram of Teager energy operator of reconstructed signal are shown in Figure 6. It can be seen from the figure that the points with large multipoint kurtosis value can be observed in the multipoint kurtosis spectrum of Teager energy operator, and the corresponding periods are $T_1 = 101$, $T_2 = 201$, and $T_3 = 301$, respectively, corresponding to 0.5, 1, and 1.5 times of the theoretical fault period. The period corresponding to the maximum kurtosis value ($T_3 = 201$) can be observed clearly in the period interval [150, 250], and the periodic error rate at this time $\eta = 5\%$, which is basically consistent with the theoretical failure period, that is, $T_2 = 201$ as the accurate failure period.

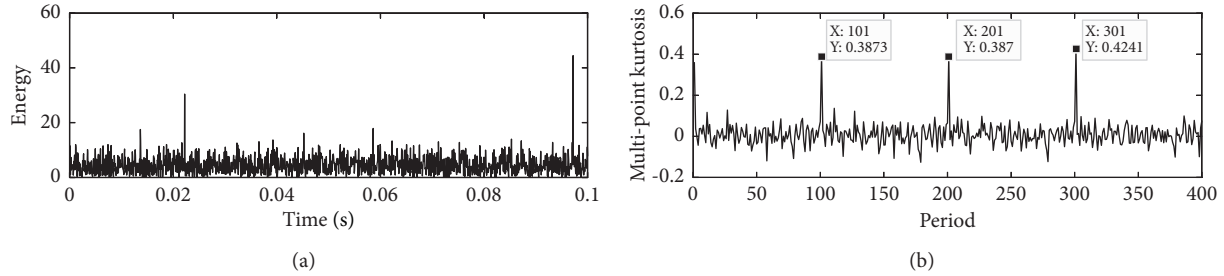
The time-domain waveform of Teager energy operator is extracted by using MOMEDA algorithm and its envelope spectrum is calculated. Set the filter length of MOMEDA to 1000 and the fault period interval to [195, 205]. The results are shown in Figure 7. In Figure 7(b), a number of impact components with regular attenuation can be observed, corresponding to the frequency doubling $f \sim 7f$ of the fault characteristic frequency of the simulation signal, and the fault type can be judged according to this.

Compared with Figures 5 and 7, it is found that the periodic pulse component of the time-domain waveform obtained by the proposed method is more obvious, and the fault characteristic frequency and its frequency doubling component of the envelope spectrum have also been accurately extracted. In order to further verify the effectiveness of the proposed method, it is compared with the method based on SVMD and MOMEDA. The noise reduction signal and its envelope spectrum are obtained by MOMEDA filtering on the reconstructed SVMD signal (the parameter settings are consistent), as shown in Figure 8.

Compared with Figures 7 and 8, it is found that the method of SVMD and improved MOMEDA is obviously better than the comparison method. Firstly, the periodic impulse characteristics of the signal obtained by the proposed method are more prominent in the time-domain diagram, and the peak values of the fault characteristic frequency and its frequency doubling components extracted from the envelope spectrum are more obvious, which further verifies the effectiveness of the proposed method.

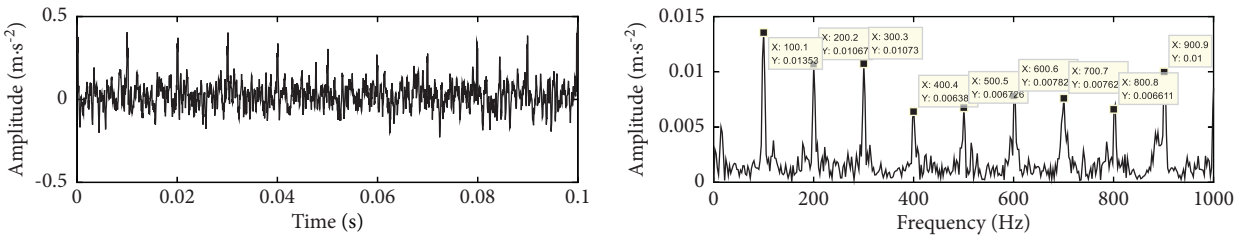
4. Experimental Verification

The method proposed in this paper is verified by using the bearing open data set of Case Western Reserve University and



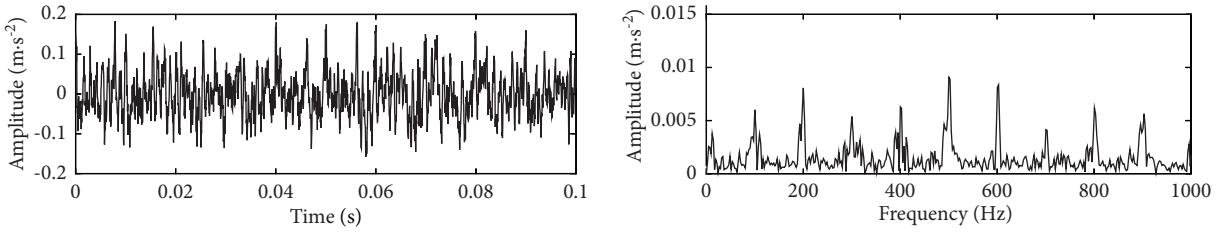
(a) (b)

Figure 6: Teager energy operator and its multipoint kurtosis spectrum of reconstructed signal. (a) Teager energy operator. (b) Multipoint kurtosis spectrum.



(a) (b)

Figure 7: Signal based on SVMD and improved MOMEDA. (a) Time-domain waveform. (b) Envelope spectrum.



(a) (b)

Figure 8: Signal based on SVMD and MOMEDA. (a) Time-domain waveform. (b) Envelope spectrum.

compared with other methods. In addition, the data come from the official website of bearing data of Western Reserve University: https://csegroups.case.edu/bearingdatacenter/home.

4.1. Early Fault Data Description. In the experiment, 6203-2RS JEM SKF deep groove ball bearing was selected, and the specific parameters are shown in Table 1.

The motor speed is 1750 r/min, the rotation frequency is 29.17 Hz, and the sensor sampling frequency is 12 KHz. For the convenience of experimental calculation, 24000 points of inner ring fault data are selected. According to the following formula of fault frequency, the failure frequency of inner ring of rolling bearing can be calculated as 144.3 Hz.

$$f_i = 0.5\mathrm{rn}\left(1 + \frac{d}{D}\cos\alpha\right), \tag{19}$$

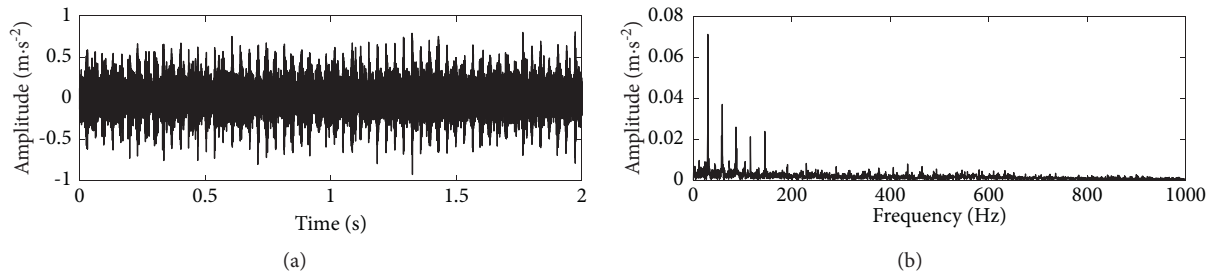
where d expresses the rolling body diameter; D expresses the bearing feature diameter; α expresses the bearing contact angle; and n expresses the number of rolling bodies, and r represents the transition. Early fault time-domain signals and their envelope are shown in Figure 9.

Observing Figure 9, it is difficult to observe obvious and regular periodic impact characteristics in time-domain waveform because the impact component in original fault signal is submerged by strong background noise. Moreover, only the motor rotation frequency fr and its frequency doubling can be observed in the envelope spectrum, and the rolling bearing fault characteristic frequency and its frequency doubling component are difficult to be effectively extracted due to the occurrence of frequency component modulation such as noise. Therefore, it is necessary to use the proposed method to denoise the original fault signal.

4.2. Early Fault Analysis Based on SVMD. Because the envelope spectrum of the original signal cannot directly reflect the fault state information, a new signal decomposition method, SVMD decomposition, is used for preliminary noise reduction. First, set the maximum balance parameters $\alpha = 12000$. The time step of double promotion $tau = 0$. The tolerance of convergence criterion $tol =$ 1e-6. The stop criterion type $stopc = 4$ (converges to the energy of the last modal component), and then SVMD decomposition is

Table 1: Structural parameters of 6203-2RS JEM SKF deep groove ball bearing.

Bearing type	Diameter of rolling element (mm)	Bearing pitch diameter (mm)	Bearing contact angle	Number of rolling
SKF 6203	6.75	28.5	0°	8



(a) (b)

Figure 9: Early fault signal. (a) Time-domain waveform. (b) Envelope spectrum.

performed to obtain 11 modal components (IMF1-IMF11), the first five of which are shown in Figure 10. Finally, according to the theory that all components extracted by SVMD algorithm are valid, the SVMD reconstructed signal and its envelope spectrum are obtained by summing all components, as shown in Figure 11.

It can be seen from Figure 10 that the impact component of the signal decomposed and reconstructed by SVMD is more prominent in the time-domain diagram, and the fault characteristic frequency and its frequency doubling component of the signal can be preliminarily found in the envelope spectrum, which shows that the decomposition method has a certain effect on the noise reduction of the signal. However, the peak value of fault characteristic frequency and its frequency doubling is not prominent, especially the weak signal can only be found at the frequency doubling, which is easily submerged by noise. Therefore, this method cannot accurately and effectively extract the early fault characteristic frequency of rolling bearing. It is necessary to take further noise reduction methods to extract the fault characteristic frequency more accurately.

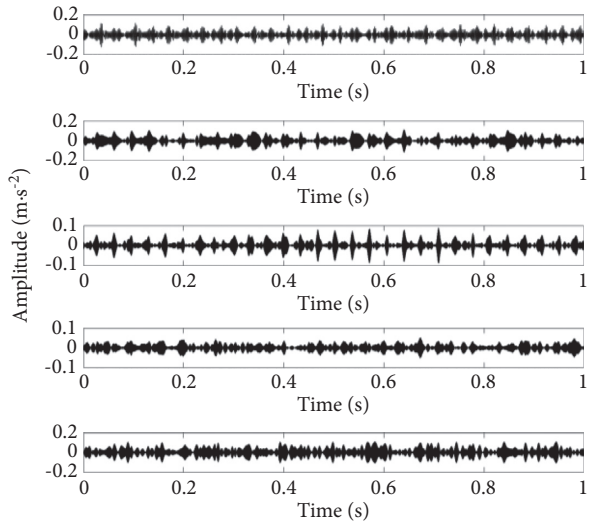


Figure 10: SVMD decomposition process.

4.3. Early Fault Analysis Based on SVMD and Improved MOMEDA. In order to highlight the characteristic frequency and frequency doubling component of the early fault signal of rolling bearing, the improved MOMEDA method is adopted to further enhance the signal characteristics after SVMD decomposition and reconstruction. Firstly, the Teager energy operator and its multipoint kurtosis spectrum of SVMD reconstructed signal are calculated, and the results are shown in Figure 12.

It can be seen from Figure 12 that the significantly enhanced impact component can be observed in the signal Teager energy operator diagram. In the multipoint kurtosis spectrum, the peak values of multipoint kurtosis with corresponding periods of 40, 84, and 167 are 0.5, 1, and 2 times of the theoretical fault period, respectively. In the period interval [80, 86], it is obvious that the period corresponding to the maximum kurtosis value is 84, and the periodic error rate is lower $\eta = 1.01\%$, which is basically consistent with the theoretical failure cycle. It is judged that the selected period is the accurate failure period.

The time-domain diagram and envelope spectrum of Teager energy operator extracted by MOMEDA filtering are shown in Figure 13. Obvious periodic impact component can be observed from the time-domain diagram, and the noise interference in the envelope spectrum can be significantly reduced. Obvious impact component (145 Hz) can be observed, which basically corresponds to the theoretical fault frequency 144.3 Hz of inner ring and its multiple frequency multiplication. According to this, it can be judged that there is inner ring fault in this rolling bearing.

In order to further prove the superiority of the proposed method, it is compared with other fault feature extraction methods. The time-domain signal and envelope spectrum are obtained by using the method based on SVMD and MOMEDA, as shown in Figure 14. Compared with the proposed method, it can be seen from the envelope spectrum that this method can only extract a single frequency of bearing early fault, and its frequency doubling component is easily submerged by noise, so it cannot be effectively

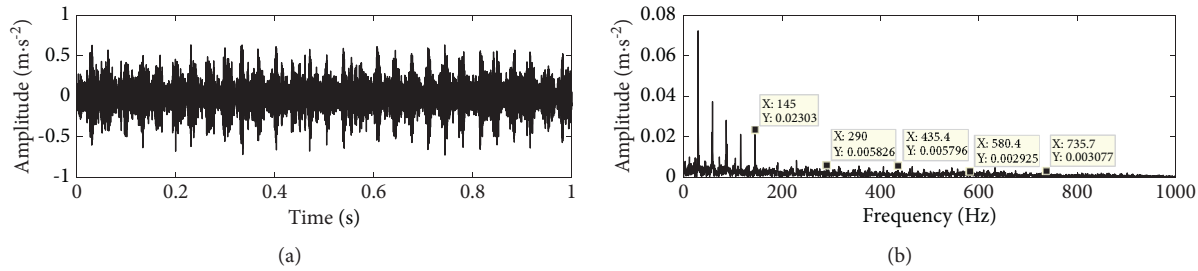


Figure 11: Signal reconstructed after SVMD decomposition. (a) Time-domain waveform. (b) Envelope spectrum.

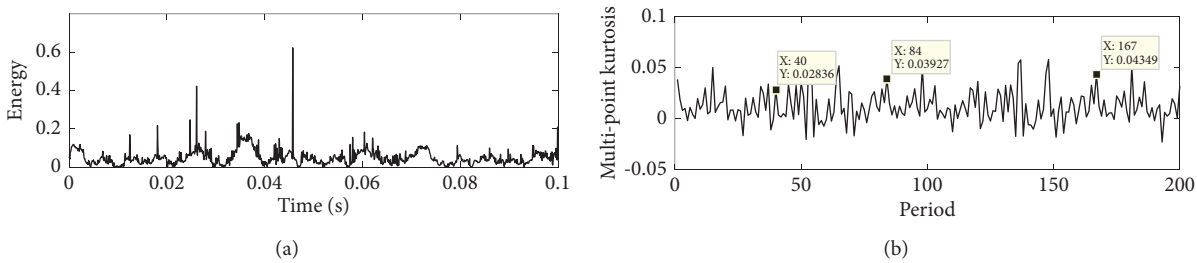


Figure 12: Teager energy operator and its multipoint kurtosis spectrum of reconstructed signal. (a) Teager energy operator. (b) Multipoint kurtosis spectrum.

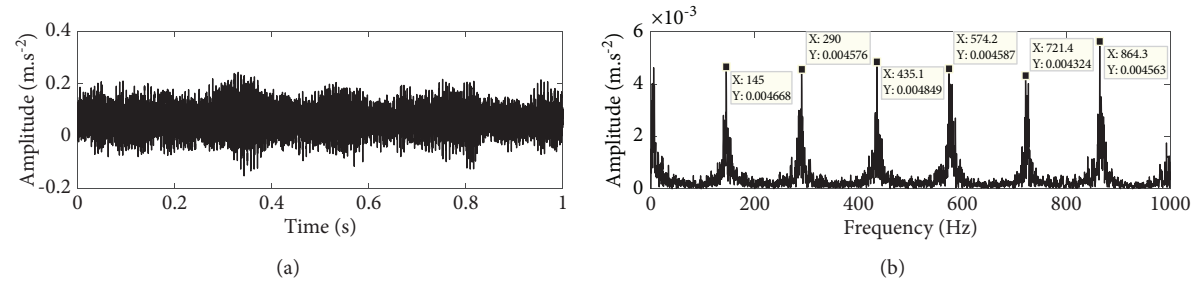


Figure 13: Signal based on SVMD and improved MOMEDA. (a) Time-domain waveform. (b) Envelope spectrum.

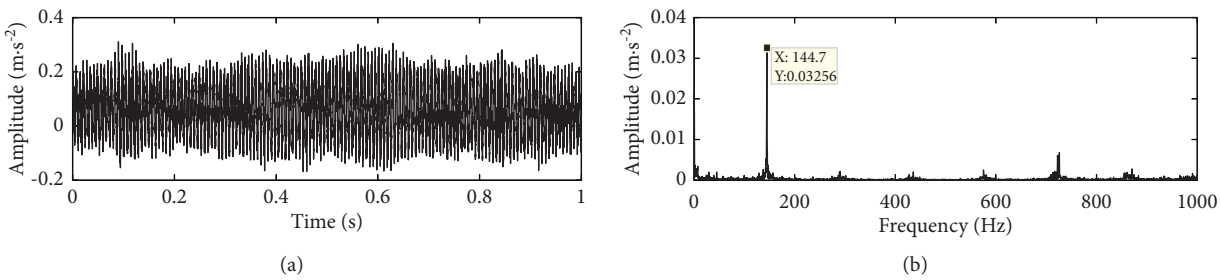


Figure 14: Signal based on SVMD and MOMEDA. (a) Time-domain waveform. (b) Envelope spectrum.

extracted. Then, the method based on SVMD and MCKD is used to extract the characteristic frequency of the original fault signal (set the filter length as 100 and the number of iterations as 200), and the time-domain signal and envelope spectrum are obtained, as shown in Figure 15.

It can be seen from the time-domain waveform in Figure 15 that only a small number of periodic impact components appear in the deconvolution reconstruction signal of MCKD, which indicates that the deconvolution effect is not ideal. The fault characteristic frequency can be found in the envelope spectrum, but its amplitude is very small, and its frequency doubling cannot be accurately found, and the characteristic frequency is easily interfered by noise and other modulation frequencies. The effectiveness of the improved MOMEDA method is further verified.

In addition, the time-domain waveform and envelope spectrum of the signal are obtained by using the method proposed in reference [17], as shown in Figure 16. Although

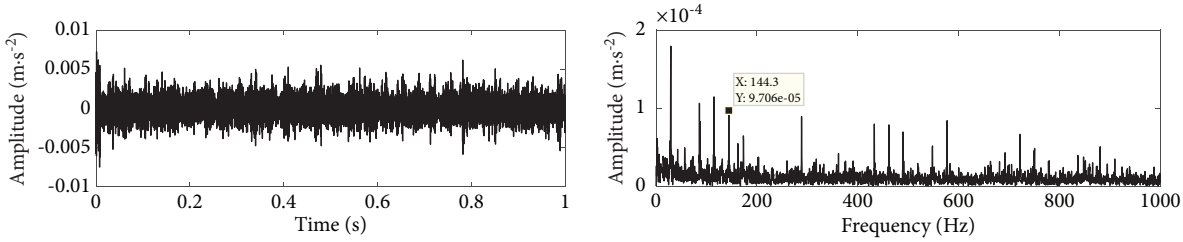


Figure 15: Signal based on SVMD and MCKD. (a) Time-domain waveform. (b) Envelope spectrum.

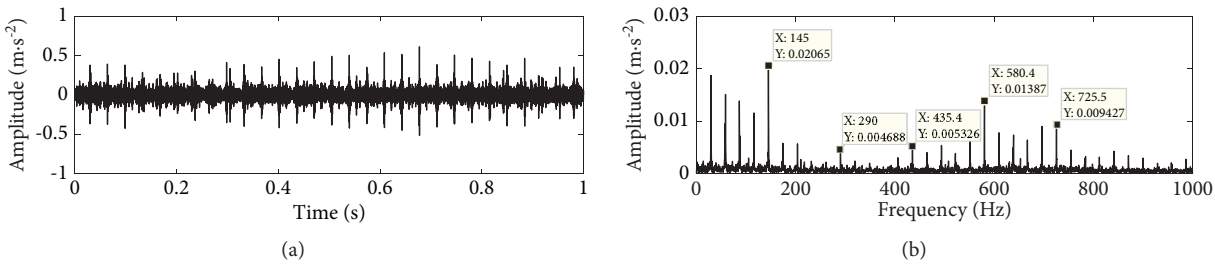


Figure 16: Signal based on adaptive MCKD and CEEMDAN. (a) Time-domain waveform. (b) Envelope spectrum.

the fault impulse component can be seen in the time-domain diagram, the fault characteristic frequency and its frequency doubling component obtained by the proposed method are weaker than those obtained by the proposed method in the envelope spectrum, especially the frequency doubling component is easily interfered by other frequency components, which further illustrates the effectiveness of the proposed method.

By comparing the proposed method with the other three feature extraction methods, it is found that the proposed method is the least disturbed by strong background noise, and the peak value is the most prominent at the fault feature frequency and its frequency doubling, which can accurately extract the early fault feature frequency of rolling bearing. The effectiveness and practicability of the method based on SVMD and improved MOMEDA are fully proved.

5. Conclusion

This paper presents a bearing fault diagnosis method based on SVMD and improved MOMEDA. Firstly, SVMD technology is used to decompose the vibration signal of the bearing, and several IMF components are obtained. Considering that the components extracted by SVMD method are all useful, the fault impulse signal is reconstructed directly by summing the components. Then, the improved MOMEDA is used to denoise the reconstructed signal. Finally, the denoised signal is demodulated by Hilbert envelope.

The simulation and experimental results show that compared with VMD, the most advantage of SVMD is that it does not need to know the number of available modes in the signal, and it has lower computational complexity and can successfully converge to the real modal components. In addition, compared with MOMEDA and other deconvolution algorithms, the improved MOMEDA algorithm can highlight the fault characteristic frequency when dealing with periodic pulse signals with strong background noise. The interference components near the peak are few, and the characteristic frequency and frequency doubling characteristics are greatly enhanced, which is more convenient for fault analysis.

Through the analysis of simulation signals and experimental data, compared with other rolling bearing fault feature extraction methods, the proposed method can extract fault feature frequency more accurately under strong noise interference, and the time cost is lower, which verifies the effectiveness and practicability of the method.

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

This work was supported by the National Natural Science Foundation of China (grant no. 61973041) and the National Key R&D Program (grant no. 2019YFB1705403).

References

[1] T. Zan, Z.-L. Pang, M. Wang, and X.-S. Gao, "early fault diagnosis method of rolling bearings based on VMD," *Journal of Beijing University of Technology*, vol. 45, no. 2, pp. 103–110, 2019.

[2] X.-Y. Zhu and Y.-J. Wang, "A method of incipient fault diagnosis of bearings based on autocorrelation analysis and MCKD," *Journal of Vibration and Shock*, vol. 38, no. 24, pp. 183–188, 2019.
[3] J. Ma and J. Xu, "Fault prediction algorithm for multiple mode process based on reconstruction technique," *Mathematical Problems in Engineering*, vol. 2015, Article ID 348729, 8 pages, 2015.
[4] R.-X. Zhuo and J.-F. Xiao, "Research on fault diagnosis method of motor bearing based on improved EEMD and SVM," *Machine Building & Automation*, vol. 48, no. 1, pp. 36–39, 2019.
[5] G.-J. Tang and X.-L. Wang, "VMD and its application in early fault diagnosis of rolling bearing," *Journal of Vibration Engineering*, vol. 29, no. 4, pp. 638–648, 2016.
[6] S. Chen, X. Dong, Z. Peng, W. Zhang, and G. Meng, "Nonlinear chirp mode decomposition: a variational method," *IEEE Transactions on Signal Processing*, vol. 65, no. 22, pp. 6024–6037, 2017.
[7] Q.-Y. Lin, L.-Y. Wei, J.-K. Yae, and C.-C. Yi, "Nonlinear FM mode decomposition and application in mechanical equipment troubleshooting," *Mechanical Design and Manufacturing*, vol. 5, pp. 77–81, 2021.
[8] N. Mojtaba and S. S. Mahmoud, "Variational mode extraction: a new efficient method to derive respiratory signals from ECG," *IEEE journal of biomedical and health informatics*, vol. 22, 2017.
[9] N. Mojtaba and S. S. Mahmoud, "Successive variational mode decomposition," *Signal Processing*, vol. 174, 2020.
[10] R. A. Wiggins, "Minimumentropy deconvolution," *Geoexploration*, vol. 16, no. 1/2, pp. 21–35, 1978.
[11] M. Buzzoni, J. Antoni, and G. D'Elia, "Blind deconvolution based on cyclostationarity maximization and its application to fault identification," *Journal of Sound and Vibration*, vol. 432, pp. 569–601, 2018.
[12] C. A. Cabrelli, "Minimum entropy deconvolution and simplicity: a noniterative algorithm," *Geophysics*, vol. 50, no. 3, pp. 394–413, 1985.
[13] G. L. McDonald and Q. Zhao, "Multipoint optimal minimum entropy deconvolution and convolution fix: application to vibration fault detection," *Mechanical Systems and Signal Processing*, vol. 82, pp. 461–477, 2017.
[14] E. Ma, Y.-C. Li, and Z. Liu, "Extraction of rolling bearing fault characteristics based on perverted mode decomposition and Teager energy operator," *Vibration and Impact*, vol. 35, no. 13, pp. 134–139, 2016.
[15] D. Yan, J.-H. Yue, and J. Jia, "Extraction of rolling fault characteristics based on autocorrelation and energy operator enhancement," *Vibration and Impact*, vol. 40, no. 11, pp. 101–108+123, 2021.
[16] Z.-C. Qiao, Y.-Q. Liu, and Y.-Y. Liao, "Application of the improved MOMEDA method based on Teager energy operators in the troubleshooting of bearings on railway wheels," *Bearings*, vol. 4, pp. 43–50, 2020.
[17] H.-M. Zhang and J.-H. Zou, "Feature extraction of rolling bearing weak fault based on adaptive MCKD and ceemdan," *Journal of Electronic Measurement and Instrument*, vol. 33, no. 4, pp. 79–86, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 2516118, 13 pages
https://doi.org/10.1155/2021/2516118

Research Article

A New Fault Diagnosis Model for Circuits in Railway Vehicle Based on the Principal Component Analysis and the Belief Rule Base

Hao Wu,[1] Bangcheng Zhang,[2] Zhi Gao,[3,4] Siyu Chen,[2] and Qianying Bu[5]

[1]*China Oil and Foodstuffs Corporation, Information Management Center, Beijing 100020, China*
[2]*School of Mechanical and Electrical Engineering, Changchun University of Technology, Changchun 130012, China*
[3]*School of Mechatronic Engineering, Changchun University of Technology, Changchun 130012, China*
[4]*School of Mechanical and Electrical Engineering, Changchun University of Science and Technology, Changchun 130022, China*
[5]*Changchun Fuwei an Daotuo Automobile Trim System Co. Ltd., Changchun 130033, China*

Correspondence should be addressed to Zhi Gao; gaozhi@ccut.edu.cn

Received 11 August 2021; Revised 27 October 2021; Accepted 11 November 2021; Published 28 December 2021

Academic Editor: Chuan Li

Circuits are considered an important part of railway vehicles, and circuit fault diagnosis in the railway vehicle is also a research hotspot. In view of the nonlinearity and diversity of track circuit components, as well as the diversity and similarity of fault phenomena, in this paper, a new fault diagnosis model for circuits based on the principal component analysis (PCA) and the belief rule base (BRB) is proposed, which overcomes the shortcomings of the circuit fault diagnosis method based on data, model, and knowledge. In the proposed model, to simplify the model and improve the accuracy, PCA is used to reduce the dimension of the key fault features, and varimax rotation is used to deduce the fault features. BRB is used to combine qualitative knowledge and quantitative data effectively, and evidential reasoning (ER) algorithm is used to carry out the inference of knowledge. The initial parameters of the model are optimized, and the optimal precondition attributes, rule weights, and belief degree parameters are obtained to improve the accuracy. Through the training and testing of the model, the experimental results show that the method can accurately diagnose the fault of the driver controller potentiometer in the railway vehicle. Compared with other methods, the model shows high accuracy.

1. Introduction

The circuits are important parts of the railway vehicle [1]. Under the adverse working conditions of high temperature and vibration environment of the railway vehicle, the failure probability is very high, and the diagnosis is difficult. Any damage to the circuits of a railway vehicle may result in security issues. In order to enhance the security of railway vehicle, the circuit fault diagnosis of railway vehicles is a very important way to find the fault in time. However, the development of circuit fault diagnosis is slow; due to the complexity and nonlinearity of the electronic system, the tolerance of the components, and the diversity of the faults, even some of the faults are slowly changed, and whether the circuit is faulty cannot be determined unilaterally [2].

Based on the particularity of the circuit system, it can be known that circuit failure is not a failure mode in the traditional sense. It is necessary to timely feed back the failure and future failure information to the driver. Therefore, the fault detection method cannot use the traditional fault detection method. It is necessary to develop a suitable fault diagnosis method according to the characteristics of the electronic system.

At present, fault diagnosis methods are mainly divided into two categories: quantitative methods and qualitative methods. Quantitative methods are mainly based on

data-driven methods, using large amounts of data to train the system, which are accurate. However, they cannot add expert knowledge and output the fault samples corresponding to the fault category label, which belong to the black box and cannot output other diagnostic information, such as the probability of the fault sample belonging to each fault category. There are mainly backpropagation neural network (BPNN) [3], support vector machine (SVM) [4], and DS evidence theory fusion method [5]. Qualitative methods can build a knowledge base or use the empirical knowledge of experts. The more the knowledge base, the higher the diagnostic accuracy, but the data cannot be used for learning and updating. For example, Huang et al. [6] proposed an improved hidden Markov model for rail transit motor faults. Li [7] proposed an expert system-based fault diagnosis method for locomotive electric drive circuits. Yang et al. and Zhang et al. [8, 9] proposed reasoning methods based on DS theory, decision theory, and rule base. This method adds rule parameters on the basis of traditional rules and establishes the learning model of the RIMER expert system.

Combining qualitative and quantitative methods can effectively solve the problem of fault classification and use data for learning and updating. Liu and Tong and Zhang [10, 11] put forward the method of combining neural network and expert system to study the fault of locomotive wheel pair, which effectively solves the misdiagnosis rate and real-time problems in traditional faults. However, the problem of the probability of failure has not been resolved. Therefore, this paper proposes a new PCA-BRB fault diagnosis model. At first, when the dimensionality of the data was too high, PCA was used to reduce the dimensionality of the feature quantity. In order to extract more representative and physically meaningful fault feature quantities, the factor rotation method, that is, the maximum variance orthogonal rotation, is used to reduce the dimension of the load matrix through factor analysis to obtain the physical meaning of the fault feature quantities. Then, the reasoning method of the belief rule base using the evidence reasoning method (RIMER) is used to diagnose the fault. The combination of principal component analysis and factor analysis can effectively reduce the number of input confidence rules in the confidence rule base (BRB) and improve learning efficiency and fault diagnosis accuracy. This paper mainly solves the problems of fault in railway vehicle circuits. The RIMER expert system can be used to classify the faults effectively; meanwhile, calculated sample belongs to the probability of each failure categories. Therefore, it can be utilized for the circuits of engineering practice. The main work in this paper includes the following:

(1) Due to the high complexity of the rail vehicle line model, the variety of data, and the difficulty of data monitoring, it is easy to cause a combination explosion, especially in a large-scale system, and it is difficult to monitor a large number of variables in each subsystem. This results in a need for fault diagnosis methods that can work with a limited set of monitoring signals. In order to simplify the model and improve the accuracy of using principal component analysis (PCA) to construct the factor model, [12], use maximum variance rotation to invert the relationship between principal components and feature quantities, and select key fault features through dimensionality reduction.

(2) In order to classify the fault samples of the circuits of rail vehicles, the RIMER method based on evidential reasoning is used to diagnose the fault. Knowledge is expressed by BRB and reasoned by ER. The input is a prerequisite attribute sample after dimensionality reduction, the basic probability mass of each confidence rule is established, and then the rules are combined by ER, and the output is the probability of each fault; according to the corresponding fault probability and the fault type semantic value, the fault type semantic value of the final output is calculated.

(3) Fmincon's active-set algorithm is employed as the optimization method to train and optimize BRB parameters. The objective function is the minimum variance of the real fault semantic value and the initial BRB output. As a result, the optimized BRB can reflect system's behavior accurately.

(4) The new PCA-BRB-based model was proposed and applied to an actual engineering system to verify the validity of the new model. Compared with the other approaches, the proposed model has shown higher accuracy.

The rest of this paper is organized as follows. In Section 2, a new fault diagnosis model for circuits in railway vehicle based on the principal component analysis and the belief rule base is proposed. In Section 3, fault diagnosis steps are put forward. In Section 4, the potentiometer is chosen as a numerical example to validate the efficiency of the new proposed model. In Section 5, conclusions are provided.

2. Fault Diagnosis Theory Based on PCA-BRB

2.1. Feature Dimension Reduction Method for PCA. The railway vehicle circuits have many electronic components and many kinds of fault characteristics, and the characteristics of faults are complicated, but it is difficult to monitor a large number of variables. It is especially important to effectively reduce the number of features and find smaller dimensions and more representative features without changing the qualitative knowledge contained in the data itself and to simplify the fault diagnosis model at the same time.

Principal component analysis (PCA) is an effective method for statistical analysis of data, which is based on the Karhunen–Loeve decomposition. Its purpose is to find a set of vectors in the data space to explain the variance of the data as much as possible. Through a special vector matrix, the data are mapped from the original high-dimensional space to the low-dimensional vector space. After the dimensionality is reduced, the vector retains the main information of the original data, making the data easier to process [13]. But

the result of PCA is to obtain the principal component which contains the comprehensive index. The main component is reconstructed, and it will lose some qualitative knowledge and does not have the actual physical meaning, which is not conducive to the construction of the BRB model. The factor analysis, through the study of the internal relationship between the correlation matrix and covariance matrix of many variables, finds out a few random variables (factors) that can integrate the main information of all variables.

Therefore, a method of combining principal component analysis with varimax rotation is proposed to reduce the dimension of fault feature. Firstly, PCA is used to extract the principal component, and then the principal component is used to carry out the reverse reasoning [14], using the method of maximum variance rotation in the factor analysis method to solve the load matrix. The purpose is to make the absolute value or square value of each element of the load matrix as much as possible two-level differentiation, that is, the absolute value or square value of a few elements is as large as possible, while other elements close to zero. The fault characteristic quantity with the absolute value of more than 0.9 is selected as the key fault characteristic of the principal component. Specific steps are as follows.

2.1.1. Standardization of Raw Data. The original dataset of the fault characteristic value of electronic circuits is $(x_{ij})^{T\times P}$ (where T is the number of samples and P is the number of fault features). In order to eliminate the influence of the different dimensions of original variables and large numerical difference, it is necessary to standardize the original variables, that is:

$$x_{ij}^{*} = \frac{x_{ij} - x_j}{\sqrt{S_{ij}}}, \quad i = 1, 2, \ldots T; j = 1, 2, \ldots p, \tag{1}$$

where

$$x_j = \frac{1}{T}\sum_{i=1}^{T} x_{ij}, \tag{2}$$

$$S_{ij} = \frac{1}{T-1}\sum_{i=1}^{T}\left(x_{ij} - x_j\right)^2. \tag{3}$$

2.1.2. Establishing Correlation Matrix and Calculating the Matrix Eigenvalue and Eigenvector

$$R = \frac{1}{T-1}X^{*}X^{T}, \tag{4}$$

where X^{*} is data matrix of standardization. Eigenvalue λ_j and eigenvector e_j are calculated by R, for $j = 1, 2, \ldots p$.

2.1.3. Selection of Principal Component. The contribution rate of the jth principal component to the total variance is calculated, that is, the variance contribution rate is calculated:

$$\text{per} = \frac{\lambda_j}{\sum_{j=1}^{P}\lambda_j \times 100\%}. \tag{5}$$

The P principal components are sorted from large to small order according to the contribution rate; usually, the principal component number n whose cumulative contribution rate is greater than 95% is selected.

2.1.4. Establishing a Factor Model

$$y_n = Ex_p. \tag{6}$$

In general, the PCA transforms P vectors $(x_1, x_2, x_3, \ldots x_p)$ to n vectors $(y_1, y_2, y_3, \ldots y_n)$. where E represent the load matrix, $E = \text{dig}\sqrt{\lambda_n}e_{np}$.

2.1.5. Orthogonal Rotation of Load Matrix

$$\begin{aligned} B &= EQ \\ &= \left(b_{ij}\right). \end{aligned} \tag{7}$$

B is the rotation factor load matrix. Let matrix B's variance of sum be max, and thus orthogonal matrices Q and B are obtained. The critical fault characteristics of $b_{ij} > 0.9$ are selected.

Through principal component analysis, the input of the expert system of the rule base is reduced, and the number of rules and the structure of the model are simplified, which can greatly improve the learning speed.

2.2. RIMER Method. The belief rule base inference methodology using the evidential reasoning approach (RIMER) is mainly composed of two parts, which are the regular representation of known information and inference of decision process rules. At present, this method is effective in solving classification problems, pattern recognition, fault diagnosis, and so on. Because of the nonlinear characteristics of the circuits, the diversity of the components, and the complex circuit connection, the mathematical model is difficult to be established when the fault of an electronic component or a branch is not proportional to the total output. However, the rule base expert system does not need to establish mathematical model through historical data and expert knowledge. It is only necessary to carry out the knowledge representation by means of the rule base, and the inference of knowledge is realized by ER inference. It belongs to white box diagnosis, which has clear observation of the reasoning process and a clear explanation of the diagnostic results. This paper presents a method of fault diagnosis for railway vehicle electronic circuits based on the belief rule base inference methodology using the evidential reasoning approach (RIMER).

2.2.1. The Knowledge Representation of the BRB. Its main purpose is to calculate the circuit input data and the rules of the rule base operations. After regularization, the known input contains the fault information of each rule, and the

fault information is an equivalent to known input weights assigned to each rule. The regularization process is not a simple IF-THEN rule, and it is a more comprehensive introduction of fault feature weight and rule weight. Sometimes, the known data are highly similar to a rule. But this rule is not important, or the premise attribute is not important, and the weight after regularization is not necessarily very high [15]. The kth rule of the belief rule base is represented as

$$\begin{aligned} &\text{If } x_1 \text{ is } A_1^k \wedge x_2 \text{ is } A_2^k \text{ is } A_M^k, \\ &\text{Then}\left\{\left(D_1, \beta_{1,k}\right), \ldots, \left(D_N, \beta_{N,K}\right)\right\}. \\ &\text{With a rule weight } \theta_k \text{ and} \\ &\text{fault feature weight } \delta_{1,k}, \ldots, \delta_{M_k,k}, \end{aligned} \tag{8}$$

where X is input vector; A^k represents a collection of entered reference values by the rule k, $A^k = \{A_1^k, A_2^k, \ldots, A_M^k\}$, $k = 1, 2, \ldots, L$, where L is the number of all rules; D is the result vector, and $D = [D_0, D_1, D_2, \ldots, D_N]$, where N is the system states number including the normal state; β^k is the vector belief degree, which is the possibility that the result of fault diagnosis results in various states; θ^k is the rule weight, which represents the importance of the corresponding rules; and δ_j is the weight of fault feature, which indicates the importance of the jth precondition (fault feature).

The BRB expert system can effectively use various types of information, and a nonlinear model is established by the rail vehicle electronic circuits characteristic input and fault class output. Compared with the traditional IF-THEN rules, it provides a way to contain more information, more close to the actual knowledge expression.

2.2.2. ER Algorithm. The BRB is the expression of frame of knowledge, but after a group of fault feature input, how to combine the L rules in the rule base is important, so that the results can be achieved in the fault detection and diagnosis. Specific steps have to be achieved by reasoning algorithm. The disadvantage of the traditional Dempster evidence theory is that there is no fusion of conflict events [16]. To solve this problem, there are two kinds of solutions, one is that the Dempster rule is flawed and needs to be modified; the other is that it is not related to the Dempster rule, and the reason lies in the modelling not being accurate, and it needs to preprocess the evidence and then make a combination [17]. However, most of the modifications to the rules have destroyed the combination rule of the Dempster rules, and the pretreatment of the evidence changed the specificity of the evidence and did not take into account the reliability and importance of the evidence. Li et al. [18] put forward a new evidence combination rule of evidence reasoning (ER) algorithm; after preprocessing the evidence, the Dempster rule is used to synthesize the algorithm, which not only satisfies the commutative law and associative law but also keeps the specificity of the evidence in the preprocessing. Specific steps are as follows [15]:

(1) Calculating the activation weights of belief rules: if x_i is assumed to be the ith input, the activation weights in the kth rule can be expressed as

$$\omega_k = \frac{\theta_k \prod_{i=1}^{M} \left(\alpha_i^k\right)^{\overline{\delta}_i}}{\sum_{l=1}^{L} \theta_l \prod_{i=1}^{M} \left(\alpha_i^k\right)^{\overline{\delta}_i}}, \tag{9}$$

$$\omega_k = \theta_k \prod_{i=1}^{M} \left(\alpha_i^k\right)^{\overline{\delta}_i}. \tag{10}$$

In this paper, two methods for calculating RIMER activation weights of circuit fault diagnosis are put forward. The activation weights in equation (9) are required to be normalized, that is, $\sum \omega_k = 1$; equation (10) is nonnormalization; when there are many rules of circuit belief rules, the normalized weight will result in the small weight of all rules, so that the fusion result is very close to the mean value. The equation (10) is used to calculate when the conflicts between the rules are not very large or the weight of the conflicting evidence is small. Fusion after normalization is used when there is a large conflict in the rules and the weight of conflict evidence is large.

$$\alpha_{ij}\left(x_i^*\right) = \begin{cases} \dfrac{x_{i(k+1)} - x_i^*}{x_{i(k+1)} - x_{ik}}, & j = k\left(x_{ik} \le x_i \le x_{i(k+1)}\right), \\ \dfrac{x_i^* - x_{ik}}{x_{i(k+1)} - x_{ik}}, & j = k+1, \\ 0, & j = 1, 2, \ldots \left|x_i\right|,\ j \ne k,\ k+1, \end{cases} \tag{11}$$

where $\omega_k \in [0,1]$, $k = 1, 2, \ldots, L$; $a_{i,j}$ is the matching degree of the input information which is based on rules or utility. The input information is the matching of i attribute of j rules.

(2) ER algorithm: the final output of BRB $S(x)$ can be realized by the combination of all the rules in the ER algorithm:

$$S(x) = \left\{\left(D_n, \widehat{\beta}_n\right)\right\}, n = 1, \cdots, N. \tag{12}$$

The confidence level of D_n is shown in equation (8):

$$\widehat{\beta}_j = \frac{\mu \times \left[\prod_{k=1}^{L}\left(\omega_k \beta_{n,k} + 1 - \omega_k \sum_{n=1}^{N} \beta_{n,k}\right) - \prod_{k=1}^{L}\left(1 - \omega_k \sum_{n=1}^{N} \beta_{n,k}\right)\right]}{1 - \mu \times \left[\prod_{k=1}^{L} (1 - \omega_k)\right],} \tag{13}$$

$$\mu = \left[\sum_{n=1}^{N}\prod_{k=1}^{L}\left(\omega_k \beta_{n,k} + 1 - \omega_k \sum_{n=1}^{N} \beta_{n,k}\right) - (N-1)\prod_{k=1}^{L}\left(1 - \omega_k \sum_{n=1}^{N} \beta_{n,k}\right)\right]^{-1}. \tag{14}$$

ω_k in equations (13) and (14) can be obtained by the activation weight equation (9) or (10). Therefore, it can be seen that $\hat{\beta}_n$ is described by the function of confidence $\beta_{n,k}(n = 1, \ldots, N, k = 1, \ldots, L)$, rule weight $\theta_k (k = 1, \ldots, L)$, and fault feature weight δ_j.

Finally, the discriminant result of output can be expressed as

$$y_i = \sum_{1}^{N} D_n \beta_n, \quad n = 1, 2, \ldots, N. \tag{15}$$

2.2.3. Optimization of the BRB Parameters. The initial BRB parameters are usually given by experts based on prior knowledge and historical information. It is difficult to determine the precise values of these parameters. Especially in the rail passenger car driver's console, there are many types of faults and the diagnosis accuracy is required. In the case of high demand, the driver's console fault diagnosis result output by the initial BRB will deviate from the real result. The accuracy of the evaluation is reduced. Therefore, in order to improve the evaluation accuracy, the initial BRB needs to be optimized, and the purpose of optimization is to minimize and optimize the error between the actual output result and the initial BRB output result. So, it is necessary to optimize the initial BRB, and fmincon's active-set algorithm is used in this paper [19].

The BRB parameter vector is $V = [\theta^1 \cdots \theta^K \cdots \delta^1 \cdots \delta^m \cdots \beta_1^k \cdots \beta_n^k]^T$, and its constraints are

$$\min \mathrm{MSE}\left(y\left(\theta^k \delta^m, \beta_n^k\right)\right), \tag{16}$$

$$0 \le \theta_k \le 1, \tag{17}$$

$$0 \le \beta_n^k \le 1, \tag{18}$$

$$\sum_{n=1}^{N} \beta_n^k, \quad k = 1, 2, \ldots L. \tag{19}$$

The objective function [20] is

$$\mathrm{MSE} = \frac{1}{T}\sum_{i=1}^{T} (y_i - y_{ir})^2, \tag{20}$$

where y_{ir} is the real referenced value of the data.

3. PCA-BRB Fault Diagnosis Model for Circuits in Railway Vehicle

3.1. Circuit Fault Diagnosis Flowchart. The flowchart of the PCA-BRB fault diagnosis model for circuits is shown in Figure 1.

3.2. Circuit Fault Diagnosis Procedure. The steps of the PCA-BRB fault diagnosis model for circuits are shown in Table 1.

4. Case Study

The driver controller is the key electrical equipment on the rail vehicle, and it is the main control device used by the driver to operate the locomotive. All instructions for the traction, braking, and speed regulation of the motor train unit are realized by the driver controller. It mainly uses the low-voltage electrical apparatus of the control circuit to indirectly control the electrical equipment of the main circuit; whether the action is well will directly affect the smooth operation of the locomotive and the realization of various conditions. The small fault of the driver controller will bring a lot of security risks to the locomotive, for example, "motor vehicle suddenly channels endplay," no voltage and no electricity, the electronic circuit receives the wrong instruction to produce the disoperation, and so on, and these will bring very big threats to the equipment safety and the driving safety. So, using numerical observations accurately to diagnose for hidden failures is particularly important.

The driver controller is prone to electrical failure, and the potentiometer circuit fault of it directly leads to serious problems of the locomotive, for example, the internal open circuit of the potentiometer could lead to no voltage and no electricity of the railway vehicle, element parameters change, and the external load of the potentiometer affects

Table 1: PCA-BRB fault diagnosis model for circuits.

Step 1: fault feature selection by the PCA method.
Step 1.1: constructing the standardized data x_{ij}^{*} by equations (1)–(3).
Step 1.2: calculating the correlation matrix R by equation (4) and calculating the eigenvalue λ_j and characteristic vector e_i of R matrix.
Step 1.3: calculating the principal component of eigenvalue by equations (6) and (7).
Step 2: the establishment of the initial BRB.
Step 2.1: defining semantic values of system input and output.
Step 2.2: setting the initial parameters of the BRB.
Step 2.3: calculating the initial BRB output semantic value by equations (8)–(14).
Step 3: optimization of the BRB parameters.
Step 3.1: calculating the initial BRB and real output variance MSE, as the target function by equation (20).
Step 3.2: establishing the constraint by equations (16)–(19).
Step 3.3: selecting the best parameters in the current generation.
Step 4: electronic circuit fault diagnosis.
Step 4.1: the testing data are transformed to the belief degrees by equation (11).
Step 4.2: calculating the activated weights of rules in the optimized BRB by equation (9) or (10).
Step 4.3: deriving the distributed output of the system's state by equations (13) and (14).
Step 4.4: calculating the expected utility of the fault diagnosis by equation (15).

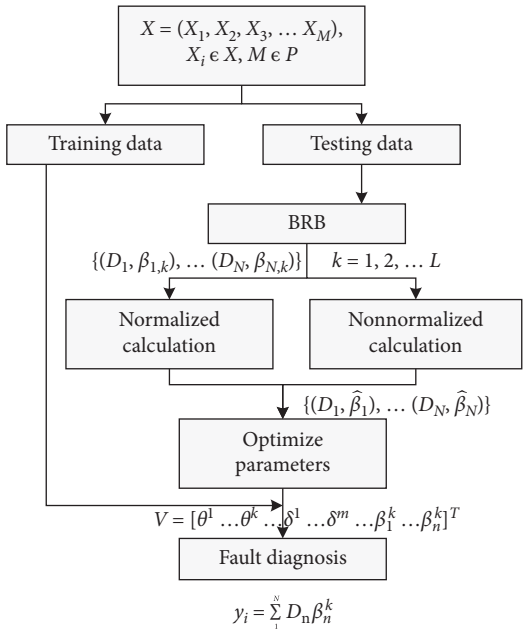


Figure 1: The flowchart of the PCA-BRB fault diagnosis model for circuits.

the voltage value of the traction system's acquisition accuracy. The purpose of this section is to solve the problem of fault diagnosis of the driver controller and take a large number of fault diagnosis problems in the circuit of the potentiometer as an example. The method is used to establish the rule base of the driver controller, and the corresponding learning model is established. The PCA is used to select the fault feature, so as to simplify the learning model. In order to verify the proposed fault diagnosis model for rail vehicles, the most common faults in working process of the driver controller potentiometer are diagnosed, and these five states are normal work, the forward bias of the component parameters, the negative deviation of the component parameters, the external load, and internal open circuit.

4.1. Experimental Design. In this paper, the fault diagnosis of the implementation of the framework is from the literature [21], as shown in Figure 2. The framework is divided into 5 stages: fault circuit generation, fault data simulation, fault feature extraction, fault diagnosis, and parameter optimization [22, 23].

4.1.1. Generating the Potentiometer Fault Circuit. Potentiometer faults can be divided into hard and soft faults. Hard faults, such as open circuit or short circuit, will cause the topology change of the potentiometer circuit, and the circuit function will change greatly.

The soft fault is the actual value of the component parameter beyond the tolerance range, leading to the reduction of the working performance of the potentiometer, the speed command error, and so on.

In the generation phase of the potentiometer fault circuit, a variety of faults are injected into the potentiometer simulation circuit by the mutation operation, and a series of circuit under test (CUT) variants with different fault modes are obtained. Fault generation is shown in Table 2. Among them, the forward bias of the component parameters and the negative deviation of the component parameters are the PCH mutation operations, and they are uniformly distributed ($U[0.1X_n, X_n - 2t]$ and $U[X_n + 2t, 2X_n]$) (X_N is the nominal value of component and t is tolerance value of the component); potentiometer external load fault is GRB mutation operation which is uniform distribution ($U[10\Omega, 1K\Omega]$); potentiometer internal open circuit is ROP mutation operation which is uniform distribution ($U[100k\Omega, 100M\Omega]$).

In this paper, potentiometer schematic diagram is adapted from the literature [23]. Figure 3 shows the simulation of the driver controller potentiometer in Proteus. Train management system (TMS) provides $15V$ analog power supply for the controller, where $RV1$ and $RV2$ are "traction" and "brake" sliding rheostats. When the speed control handle A of the master controller is manipulated and rotated in the $RV1$ or$RV2$ area, the output voltage of the

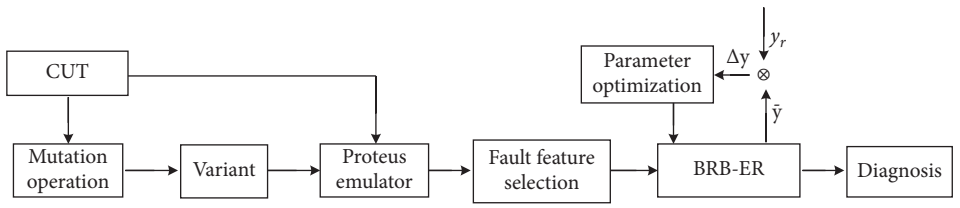


Figure 2: The structure of the PCA-BRB model.

Table 2: Mutation operators.

Operators	Name	Description
PCH	Parameter changes	Specified parameters of the component diverge the tolerance range
GRB	Global resistive bridging	A resistor with a very low resistance is connected between the two nodes of different components
ROP	Resistance open	A resistor with a high resistance is connected between the ports of the component to represent the open circuit

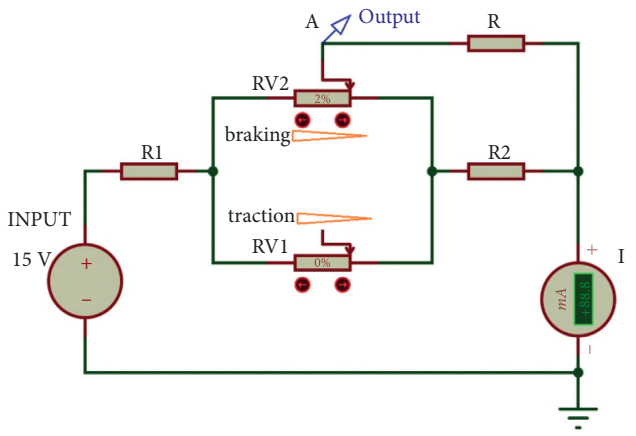


Figure 3: The circuit diagram of motorman controller potentiometer.

potentiometer is changed, thereby issuing instructions to adjust the speed of the locomotive. In this paper, 101 kinds of data in the potentiometer from the "0" slide to the braking process are randomly injected into each fault.

Fault description is shown in Table 3.

In the Proteus simulation software, each class of 101 sets of fault samples (505 groups) is obtained. The fault samples are divided into 2 categories according to the category. The first one is the training sample (group 340), and the second one is the test sample (group 165).

4.2. Feature Reduction. The driver controller potentiometer has a lot of fault types and fault element prerequisite attributes, and the circuit has some branches and trunk roads, such as voltage, current, resistance of the branches and trunk roads, sliding ratio of sliding rheostat in braking process $RV2$, and so on. The correct choice of characteristic quantity is not only easy to diagnose but also simplifies the learning model. Because the potentiometer is mainly operated by regulating the resistance variation of the change of voltage to control the speed of railway vehicles, this paper mainly

Table 3: Fault description of potentiometer.

Number	Fault description	Nominal value	Fault value
D_0	Fault-free		
D_1	RV2↑	1043Ω	[1460.2, 2086]Ω
D_2	RV2↓	1043Ω	[104.3, 625.8]Ω
D_3	GRB($R5$)		[10, 1000]Ω
D_4	ROP(RV2)	1043Ω	[100, 100000]KΩ

assumes the circuit output voltage, main circuit current, and output resistance as characteristic quantities. For more complex circuit or circuit having a plurality of output terminals, more characteristic quantities can be chosen.

Using the training data, the number of samples is $T = 340$ and the fault feature number is $P = 3$. It is necessary to perform standardized treatment of original variables, that is, the original variables need to be standardized.

The mean is $x_j = [\,5.89 \;\; 10.48 \;\; 0.63\,]$; the standard deviation is $S_{ij} = [\,2.68 \;\; 2.23 \;\; 0.40\,]$; and the characteristic quantity is $\lambda_j = [\,2.672 \;\; 0.306 \;\; 0.022\,]$. Calculate the contribution to the total variance of the principal components in the ith rate, and the variance contribution rate is

$$\text{per}_i = \frac{\lambda_j}{\sum_{j=1}^{P} \lambda_j \times 100\% = [\,89.07\% \;\; 10.19\% \;\; 0.74\%\,]}. \quad (21)$$

It shows that there are 2 total contribution rates of more than 95% principal components (because the characteristic of greater than 1 is only one factor by using SPSS analysis, it cannot be rotated, and two principal components are extracted by the gravel method); according to varimax rotation, the load matrix is obtained, and the correlation between the principal components and key fault feature is shown in Table 4.

The main elements usually take the total contribution rate of more than 95%, which contain most of the original variable information. Through principal component analysis, two principal components of total contribution rate are more than 95%, and output voltage coefficient of component

Table 4: The correlation between the two principal components and the fault characteristics.

Feature	Component 1	Component 2
Output voltage	0.924	0.379
Electric current	−0.414	0.910
Resistance	0.855	0.495

1 which is obtained by maximum variance rotation method is greater than 0.9, the main current coefficient (0.910) of component 2 is greater than 0.9, and the two features are selected as an input of the BRB expert system. The number of rules and model structure are simplified, which can greatly improve the learning rate.

4.3. RIMER Algorithm. The output voltage is selected by 6 reference values: zero (Z), very small (VS), small (S), medium (M), large (L), and very large (VL), that is, $A_1 \in \{Z, VS, S, M, L, VL\}$. The electric current is selected by 7 reference values: very small (VS), medium small (MS), small (S), medium (M), large (L), medium large (ML), and very large (VL), $A_2 \in \{VS, MS, S, M, L, ML, VL\}$. The semantic value and referenced values of features are shown in Tables 5 and 6.

The potentiometer has five fault states which are normal work (D0), element parameter positive deviation (D1), element parameter negative deviation (D2), external load (D3), and potentiometer open circuit (D4), D = [D0, D1, D2, D3, D4]. Above the semantic value, it needs to be quantified, and the referenced values of the fault states are shown in Table 7.

4.3.1. The Initial BRB Potentiometer Fault Diagnosis. Because the output voltage has 6 reference values and the current has 7 reference values, there are 42 belief rules in belief rule base (BRB). Divide the experimental data into 152 sets of training data and 238 sets of test data $(x_{152,1}\ x_{152,2}) = (4.27253\ 11.4)$, $(x_{283,1}\ x_{283,2}) = (3.43793\ 11.9)$. The output voltage and electric current reference values of the two sets data are very similar, and the semantic value A_1 is between S and M. The semantic value A_2 is between M and L. However, 152 groups of fault data are D2 (component parameters $RV2\downarrow$), and the reference value is 3; 238 groups of fault data are D_3 (external load failure), and the reference value is 4, which is diagnosed by the initial confidence rule base.

(1) The initial BRB diagnosis processes of 152 data $(x_{152,1}\ x_{152,2}) = (4.27523\ 11.4)$ are shown in Tables 8–10.

$$\begin{aligned}\{(D_0, \beta_{1,k}), \ldots (D_N, \beta_{N,k})\} = \{&(D_0, 0.0320),\\ &(D_1, 0), (D_2, 0.650),\\ &(D_3, 0.3180), (D_4, 0)\}.\end{aligned} \tag{22}$$

That is, the probability output of the fault sample belongs to the fault types; the probability of the third fault is the largest.

Table 5: The referenced values of A_1.

Semantic values	Z	VS	S	M	L	VL
Referenced values	0	2.4	3	6	8	10.1

Table 6: The referenced values of A_2.

Semantic values	VS	MS	S	M	L	ML	VL
Referenced values	7.5	9	10	11.5	13.5	14.5	21.5

Table 7: The referenced values of the fault states.

Semantic values	D0	D1	D2	D3	D4
Referenced values	1	2	3	4	5

Table 8: Relative membership degree of A_1.

Semantic values	Z	VS	S	M	L	VL
Membership	0	0	0.575823	0.424177	0	0

Confidence relative evaluation results of ER algorithm:

$$\begin{aligned}y_i &= \sum_{1}^{N} D_n \beta_n^k\\ &= 3.254.\end{aligned} \tag{23}$$

Fault status belongs to D3, component parameters $RV2\downarrow$. The diagnosis is correct.

(2) The initial BRB diagnosis processes of 238 data $(x_{238,1}\ x_{238,2}) = (3.43793\ 11.9)$ are shown in Tables 11–13.

$$\begin{aligned}\{(D_0, \beta_{1,k}), \ldots (D_N, \beta_{N,k})\} = \{&(D_0, 0), (D_1, 0),\\ &(D_2, 0.5558), (D_3, 0.4442), (D_4, 0)\}.\end{aligned} \tag{24}$$

That is, the probability output of the fault sample belongs to the fault types; the probability of the third fault is the largest, and it is inconsistent with the real situation.

Confidence relative evaluation results of ER algorithm:

$$\begin{aligned}y_i &= \sum_{1}^{N} D_n \beta_n^k\\ &= 3.4442.\end{aligned} \tag{25}$$

Fault status belongs to D3, component parameters $RV2\downarrow$. The diagnosis is wrong. In the next section, the initial BRB parameters are optimized to improve the accuracy.

4.4. The Fault Diagnosis Using the Optimized BRB. The BRB parameters $V = [\theta^1 \cdots \theta^K \cdots \delta^1 \cdots \delta^m \cdots \beta_1^k \cdots \beta_n^k]^T$ are optimized by establishing the true reference value of the data category and the variance of the actual output value. It is diagnosed by the belief rule base after optimization.

Table 9: Relative membership degree of A_2.

Semantic values	Z	VS	S	M	L	VL	VL
Membership	0	0	0	0.067	0.933	0	0

Table 10: The activation weight calculation section of the kth rule.

Article k these rules	17	18	24	25	Other rules
ω_k	0.03839	0.53744	0.02828	0.39589	0

ER algorithm is used to calculate the relative evaluation results of confidence.

Table 11: Relative membership degree of A_1.

Semantic values	Z	VS	S	M	L	VL
Membership	0	0	0.8540233	0.1459767	0	0

Table 12: Relative membership degree of A_2.

Semantic values	VS	MS	S	M	L	ML	VL
Membership	0	0	0	0.8	0.2	0	0

Table 13: Kth rule activation weight calculation.

Article k these rules	18	19	25	26	Other rules
ω_k	0.6832	0.5723	0.1168	0.0292	0

ER algorithm is used to calculate the relative evaluation results of confidence.

(1) The optimized BRB diagnosis process of 152 data.

ER algorithm is used to calculate the relative evaluation results of evaluation.

$$\begin{aligned}\left\{\left(D_0,\beta_{1,k}\right),\ldots\left(D_N,\beta_{N,k}\right)\right\} = \{&(D_0,0.0739),\ (D_1,0.2242),\\ &(D_2,0.4886),\ (D_3,0.0195),\\ &(D_4,0.1889)\}.\end{aligned} \tag{26}$$

That is, the probability output of the fault sample belongs to the fault types, and the probability of the third fault is the largest.

The evaluation results of semantic value.

$$\begin{aligned}y_i &= \sum_{1}^{N} D_n\beta_n^k\\ &= 3.0106.\end{aligned} \tag{27}$$

Fault status belongs to D3, component parameters $RV2\downarrow$. The diagnosis is correct and more accurate than the original BRB.

(2) The optimized BRB diagnosis process of 238 data.

ER algorithm is used to calculate the relative evaluation results of evaluation.

$$\begin{aligned}\left\{\left(D_0,\beta_{1,k}\right),\ldots\left(D_N,\beta_{N,k}\right)\right\} = \{&(D_0,0.0055),\ (D_1,0.0263),\\ &(D_2,0.4440),\ (D_3,0.1365),\\ &(D_4,0.3761)\}.\end{aligned} \tag{28}$$

That is, the probability of the output result sample belongs to the fault category, and the probability of third kinds of faults is the largest, which is not consistent with the real situation.

The evaluation results of semantic value.

$$\begin{aligned}y_i &= \sum_{1}^{N} D_n\beta_n^k\\ &= 3.8166.\end{aligned} \tag{29}$$

It is worth noting here that it belongs to D4 fault when $3.6 \le y_i < 4.4$.

Fault state belongs to D4, external load fault. The diagnosis is correct. Through the comparison of the above two methods, in terms of precision or accuracy, the updated BRB diagnosis effect is much better than the initial BRB.

The fault category of the testing data can be calculated by the same way. Figure 4 shows the output of the

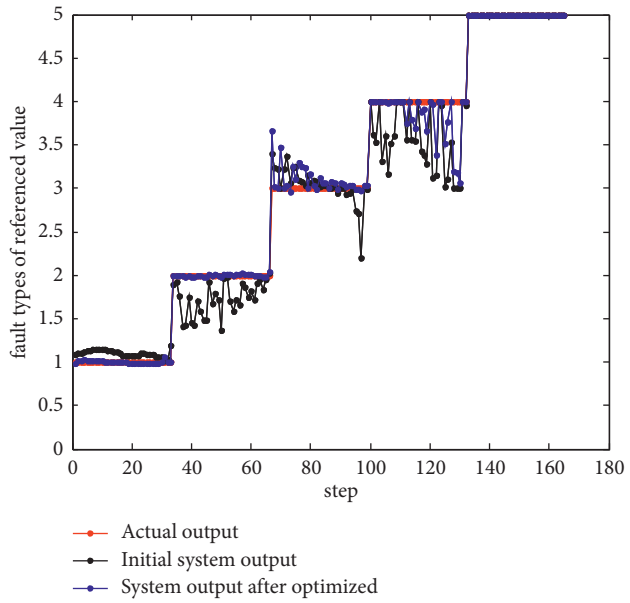


Figure 4: Comparison chart of optimized system fault diagnosis results.

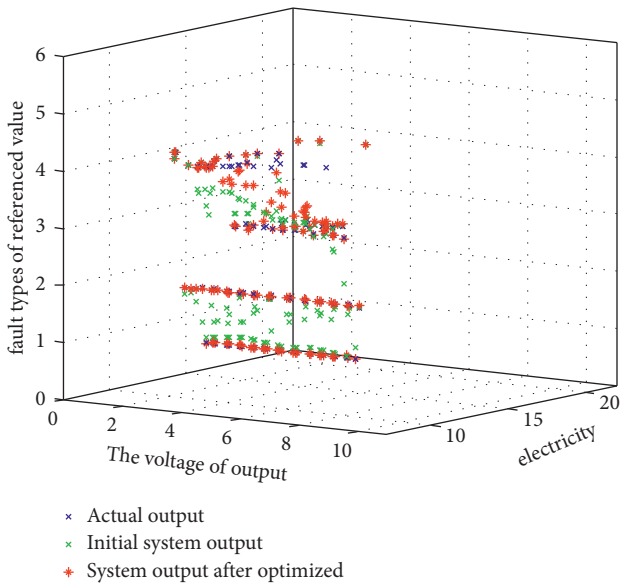


Figure 5: Three-dimensional diagram of optimized system fault diagnosis results comparison.

potentiometer's fault state using initial and optimized BRB. Figure 5 shows the fault feature and output of the potentiometer's fault state using initial and optimized BRB.

As can be seen from Figure 5, the fault type coordinates of the optimized BRB diagnosis system are more suitable to the actual reference value; as can be seen from Figure 4, the reference value of the optimized BRB fault diagnosis system is similar to the actual reference value. The blue line in the figure is almost entirely within the range of the true reference value; therefore, the updated diagnostic results are more accurate and the diagnostic accuracy is higher.

4.5. Fault Diagnosis of Optimized BRB Potentiometer Based on Unnormalized Activation Weight. The comparative study is also done by the ER algorithm of nonstandardized ω_k using the same training and testing data. Figure 6 shows the comparison chart of output value of unnormalized activated weights fault diagnosis and true reference value.

ER algorithm of unnormalized is

$$\omega_k = \theta_k \prod_{i=1}^{M} \left(\alpha_i^k\right)\overline{\delta}_i. \tag{30}$$

The accuracy by using unnormalized activated weights is more than that using initial BRB. However, the accuracy is less than that of optimized BRB of normalized activated weights. It proves that when there is a conflict of rules and the weight of conflict evidence is very large, it is suitable for fusion after normalization.

4.6. The Fault Diagnosis Using the PCA to Reconstruct Fault Feature in the Optimized BRB. The principal component of reconstruction by PCA is used as input feature of the BRB model, not the varimax rotation back stepping features. This

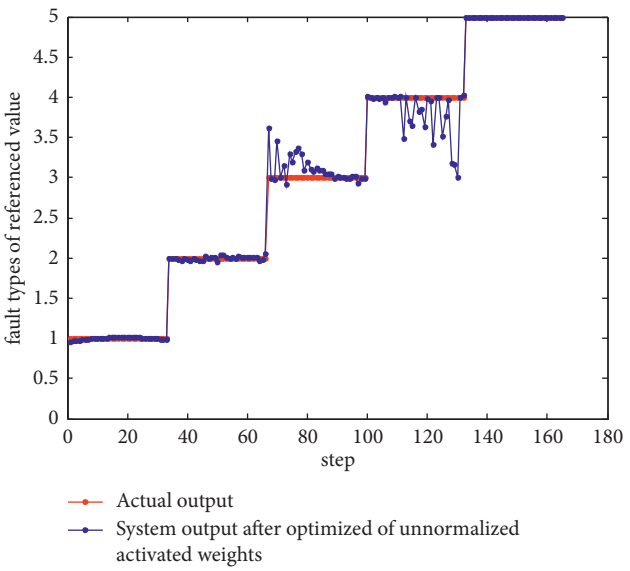


Figure 6: Comparison of the diagnostic results of the unnormalized activation weight potentiometer and the true value of the test data.

paper also conducts a comparative study of this method, as shown in Figure 7. The accuracy and error are less than those of optimized BRB of varimax rotation. The data diagnosis process without physical meaning and qualitative knowledge proved to be difficult.

4.7. The Fault Diagnosis Using the Neural Net Algorithm. In this section, the potentiometer is diagnosed by using BP neural network. The parameter settings are as follows: net.train Param.epochs = 340 and net.train Param.goal = $le-5$. Others are the default values. The training data and testing data are the same text. Figure 8

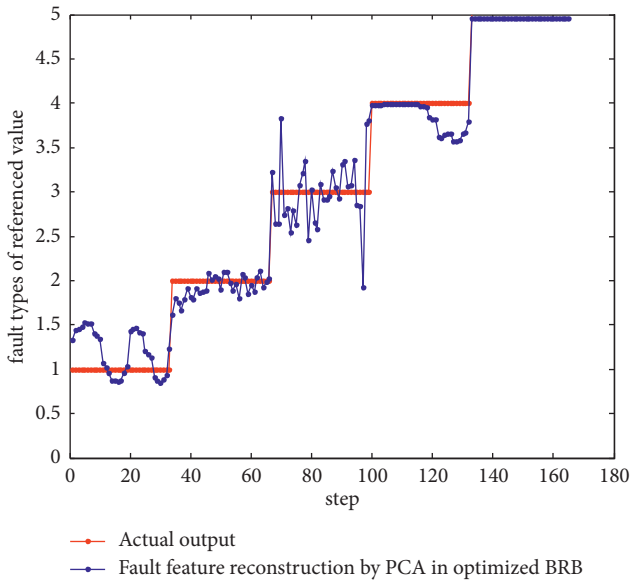


Figure 7: PCA reconstructed data are used to optimize the BRB potentiometer fault diagnosis result and the true value comparison chart of the test data.

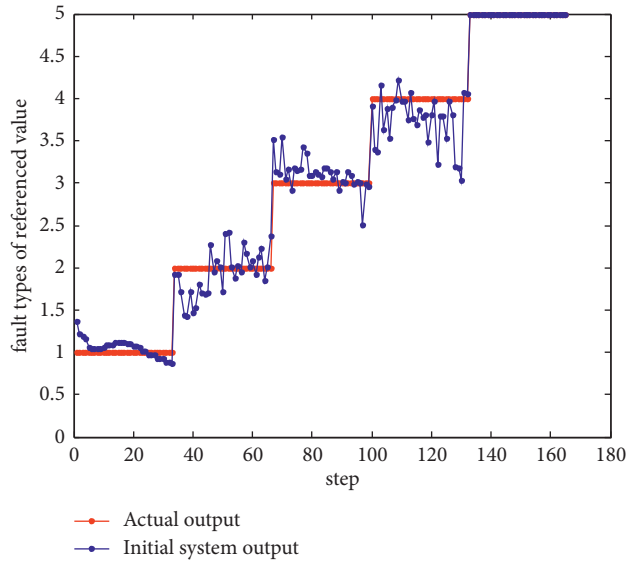


Figure 8: Comparison chart of actual output and initial system output.

Table 14: Comparison of the results of different methods.

Method	Accuracy (%)	MSE
Optimized BRB	95.8	0.1603
Initial BRB	83.6	0.3091
ER algorithm of unnormalized ω_k in optimized BRB	95.2	0.1722
Fault feature reconstruction by PCA in optimized BRB	87.9	0.2545
BP neural network	89.1	0.2396

shows the comparison chart of output value of BP neural network fault diagnosis and true reference value.

Moreover, BP neural network belongs to the "black box" method, so it is difficult to explain the complicated relationship of complex system between the input and output. Only the category label corresponding to the fault sample can be output, and other diagnostic information, such as the probability of each fault category in the fault sample, cannot be output.

4.8. Comparison of the Results of Different Methods. Two methods of true reference value and output reference value

are used to evaluate the fault diagnosis capability, as shown in Table 14. The results prove that the optimized BRB results are better than the other two methods, with better accuracy and lower error.

5. Conclusion

Taking the controller potentiometer as an example, a fault diagnosis model is established in this paper. This paper focuses on the fault diagnosis methods of railway passenger cars, which can also be developed into a fault diagnosis system of railway passenger cars. The feasibility study of applying the PCA-BRB fault diagnosis method to electronic circuit of the railway vehicle is introduced. Firstly, the limited fault feature acquisition and simplified circuit model were solved by the principal component analysis and maximum variance rotation method. In this method, qualitative knowledge and quantitative information are combined effectively, and two fusion methods of normalized activation weights and nonnormalized activation weights are proposed according to different actual situations. Finally, the initial BRB parameters are optimized. Compared with initial BRB, unnormalized activation weights in optimized BRB, PCA, and neural network for reconstruction of fault features in optimized BRB, the proposed method has higher diagnostic accuracy and minimum error. The fault diagnosis method proposed in this paper can find the faults of railway vehicles and ensure the reliability of railway operation.

Further efforts should be made to further study the following aspects in the future:

(1) Higher precision optimization algorithm could be explored to improve the optimization ability of confidence rule base model.

(2) Although this paper only takes the common circuit faults as an example, more emphasis will be placed on the study of fault diagnosis methods for rail cars in the future, which can be developed into the fault diagnosis system for rail cars.

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 they have no conflicts of interest.

Acknowledgments

This research was supported by the Science and Technology Project of the 13th Five-Year Plan of Education Department of Jilin Province (no. JJKH20200655KJ) and the Science and Technology Development Project of Jilin Province (no. 20200301038RQ).

References

[1] B. S. Kariyappa, "A survey on fault diagnosis of analog circuits: taxonomy and state of the art," *AEUE-International Journal of Electronics and Communications*, vol. 73, pp. 68–83, 2017.

[2] H. Yuan, G. Chen, and Y. Xie, "Feature evaluation and extraction based on neural network in analog circuit fault diagnosis," *Journal of Systems Engineering and Electronics*, vol. 18, no. 2, pp. 434–436, 2007.

[3] G. Song, S. Jiang, and H. Wang, "Analog circuit fault diagnosis approach using optimized SVMs based on MST algorithm," *IEEE*, vol. 4, pp. 236–240, 2011.

[4] A. Zhang, C. Chen, and B. Jiang, "Analog circuit fault diagnosis based UCISVM," *Neurocomputing*, vol. 173, pp. 1752–1760, 2016.

[5] L. Oukhellou, A. Debiolles, and T. Denœux, "Fault diagnosis in railway track circuits using Dempster–Shafer classifier fusion," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 1, pp. 117–128, 2010.

[6] D. R. Huang, L. Y. Ke, X. Y. Chu, L. Zhao, and B. Mi, "Fault diagnosis for the motor drive system of urban transit based on improved Hidden Markov Model," *Microeletronics Reliability*, vol. 82, pp. 179–189, 2018.

[7] X. T. Li, *Research on Fault Diagnosis System of Locomotive Electric Drive Circuit*, Northwestern Polytechnical University, Xi'an, China, 2003.

[8] J. B. Yang, J. Liu, and J. Wang, "Belief rule-base inference methodology using the evidential reasoning approach-RIMER," *IEEE Transactions on Systems*, vol. 36, no. 2, pp. 266–285, 2006.

[9] W. Zhang, C. Shi, C. H. Hu, and Z. Z. Zhou, *Fault Diagnosis of Engine Based on RIMER Expert System*, System Simulation Technology, China, 2011.

[10] Z. W. Liu, "Research on fault diagnosis method of locomotive wheelset based on neural network," *Application of Automation*, vol. 8, pp. 17-18, 2019.

[11] Z. Tong and Y. P. Zhang, "Research on locomotive wheelset fault diagnosis system based on BP neural network," *Railway Technology Innovation*, vol. 109, no. 1, pp. 21–23, 2009.

[12] L. L. Chang, L. I. Meng-Jun, and L. U. Yan-Jing, "Structure learning for belief rule base using principal component analysis," *Systems Engineering-Theory & Practice*, vol. 34, no. 5, pp. 1297–1304, 2014.

[13] Z. J. Zhou, *Belief Rule Base of Expert System and Complex System Modeling*, Science Press, China, 2011.

[14] G. Shafer, "A mathematical theory of evidence," *Technometrics*, vol. 20, no. 1, p. 242, 1976.

[15] Z. Zhou, X. B. Xu, and C. L. Wen, "An optimal method for combining conflicting evidences," *Acta Automatica Sinica*, vol. 38, no. 6, p. 976, 2012.

[16] J. B. Yang and D. L. Xu, "Evidential reasoning rule for evidence combination," *Artificial Intelligence*, vol. 205, no. 205, pp. 1–29, 2013.

[17] J. M. Bradley and E. M. Atkins, "A cyber-physical optimization approach to mission success for unmanned aircraft systems," *Journal of Aerospace Computing, Information, and Communication*, vol. 11, no. 1, pp. 48–60, 2014.

[18] G. Li, Z. Zhou, and C. Hu, "A new safety assessment model for complex system based on the conditional generalized minimum variance and the belief rule base," *Safety Science*, vol. 93, pp. 108–120, 2017.

[19] X. Tang and A. Xu, "Practical analog circuit diagnosis based on fault features with minimum ambiguities," *Journal of Electronic Testing*, vol. 32, no. 1, pp. 83–95, 2016.
[20] X. Tang, X. U. Aiqiang, and L. I. Wenhai, *A Stochastic Mutant Generation Method for Multi-Level Analog Circuits*, China Measurement & Test, Shanghai, China, 2016.
[21] L. Wang, *The Research of Master Controller for Metro*, Dalian University of Technology, Dalian, China, 2015.
[22] W. S. Pritchard, "PCAVR: a portable laboratory program for performing varimax-rotated principal components analysis of event-related potentials," *Brain Research Bulletin*, vol. 13, no. 3, pp. 465–473, 1984.
[23] Y. Xiao and L. Feng, "A novel linear ridgelet network approach for analog fault diagnosis using wavelet-based fractal analysis and kernel PCA as preprocessors," *Measurement*, vol. 45, no. 3, pp. 297–310, 2021.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 2605095, 16 pages
https://doi.org/10.1155/2021/2605095

Research Article

A Design Method for the Roadside Clear Zone Based on Accident Simulation Analysis

Rui Cheng, Ye Pan, and Tao Wang

School of Architecture and Transportation, Guilin University of Electronic Technology, Guilin 541004, China

Correspondence should be addressed to Tao Wang; wangtao@seu.edu.cn

Received 15 July 2021; Accepted 26 October 2021; Published 25 November 2021

Academic Editor: Darong Huang

In order to improve the safety design of roadside areas and reduce the loss of roadside accidents, this paper uses PC-Crash software to perform an accident simulation analysis. By recording the track of the vehicle after entering the roadside, the recommended widths of the roadside clear zone for different operating speeds and horizontal curve radii in straight and curved sections are given. According to our previous research data, the conditions for setting the roadside clear zone are proposed. Finally, based on a cost-benefit ratio analysis, a comprehensive risk index method is adopted to evaluate the social stability risk of the project and conduct research on the design method of the roadside clear zone. The results show that the width of the roadside clear zone has an exponential relation with the departure speed and a power relation with the horizontal curve radius. The research results realize the accurate calculation of the roadside clear zone width and fill in the gaps of the relevant specifications and guidelines in the setting conditions of the roadside clear zone.

1. Introduction

Despite improvements in road safety, the "Global Status Report on Road Safety" issued by the World Health Organization shows that approximately 1.25 million people still die in traffic accidents yearly. In the twentieth century, approximately 25.85 million people died from traffic injuries worldwide, which is greater than the number of deaths in World War I. RISER, a European research project, has shown that roadside accidents account for 19%, 22%, and 19% of the total accidents in Germany, Austria, and Greece, respectively, but account for 33%, 36%, and 34% of all traffic fatalities, respectively [1]. According to the statistical report from the Roadside Safety Research Program of the Federal Highway Administration (FHWA), roadside accidents accounted for more than 50% of the total deaths in 2018 [2]. Additionally, the Road Traffic Accident Statistical Annual Report of China also revealed that roadside accidents accounted for approximately 8% of the total number of traffic accidents each year but had a 13% fatality rate [3]. Given the high death rate of roadside accidents, it is urgent to conduct research on roadside safety design.

Since the late 1960s, roadside safety design has been a hotly debated topic in highway design. The concept of the roadside clear zone (RCZ) first appeared in a meeting document of the Highway Research Board (HRB) in 1963, which was formally written into the Highway Safety Design Manual in May 1965. After two revisions in 1973 and 1978, the manual was incorporated into the practical application of road project construction. In 1989, the first edition of the Roadside Design Guide (RDG) was published by AASHTO, followed by the second edition in 2002, and the third and fourth editions were published in 2006 and 2011, respectively.

In the fourth edition of the RDG, the RCZ is defined as "*Clear Zone-The unobstructed, traversable area provided beyond the edge of the through traveled way for the recovery of errant vehicles. The clear zone includes shoulders, bike lanes, and auxiliary lanes, except those auxiliary lanes that function such as through lanes*" [4]. According to the definition of the RCZ, this area should provide a fault-tolerance space with sufficient width, gentle slope, and no hazards for runaway vehicles. When implementing the RCZ design, how to mitigate existing hazards in the RCZ in order to reduce loss

of out of control vehicles and how to determine the slope and width of the RCZ according to different road characteristics, regional characteristics, and traffic conditions have always been a topic of concern in academia.

The fourth edition of the RDG gives the recommended width of the RCZ in straight sections and the correction coefficient of the curved section. The recommended width of the RCZ in the straight section is determined based on the annual average daily traffic (AADT), design speed, slope gradient, and slope form (i.e., fill or cut) [4]. Meanwhile, this edition also gives the setting standard of the RCZ slope [4], as shown in Figure 1. A slope gradient equal to or less than 1:4 is considered to be recoverable, and a slope gradient between 1:3 and 1:4 is considered to be unrecoverable. Additionally, when roadside conditions do not allow for setting a recoverable slope, an additional buffer zone (clear run-out zone) can be placed at the foot of the nonrecoverable slope to reduce accident losses, and it is stipulated that the slope gradient in this field should be equal to or less than 1:6. Note that these research results are widely used internationally. Given that the recommended widths from the fourth edition of the RDG are deduced based on limited empirical data, even though these empirical data provide a reference for roadside safety design in most countries, the road network characteristics, vehicle ownership, driver characteristics, roadside hazard distribution, and economic level of each country vary greatly. Therefore, scholars have taken various measures to analyze roadside safety design according to their national conditions.

China's Specifications for the Highway Safety Audit (JTG B05-2015) provides a graphic method for the RCZ width on fill and cut subgrades. The determination of the RCZ width in straight sections is based on the one-way AADT and running speed, and the RCZ width in the curved sections is selected by the correction coefficient according to operating speeds and the curve radii; furthermore, when the slope gradient is greater than 1:3.5, the filling slope cannot be considered to be an effective safe area; when the fill slope gradient is between 1:3.5 and 1:5.5, the 1/2 slope width can be regarded as a safe zone; when the slope gradient is less than 1:6, the entire slope width can be used as a safe clear zone [5]. Australian scholars have recommended that a clear area width of at least 2 m from the curb could significantly reduce injuries in vehicles leaving the road [6]. Sax et al. proposed that setting the RCZ at a width of 4~5 feet could reduce collisions with roadside objects by 90% [7]. By analyzing the main factors influencing the design of the RCZ width, such as the braking reaction time, the distance traveled by vehicles, and the state of vehicles leaving the road, Fan and Xing constructed a calculation model of the RCZ width [8], where the basic assumption is that the slope is flat, which is not consistent with the actual state. The United Kingdom studied the relationship between the increase in the RCZ width and the number of accidents and found that increasing the RCZ width is correlated with a gradual decrease in the number of roadside accidents [9]. However, a study from Jurewicz and Pyta proposed that even an RCZ with a width of 29.5 feet still did not prevent numerous roadside accidents [10]. For this reason, in a study on driver behavior, by investigating the impact of two variables (i.e., the RCZ width and the density of roadside vegetation) on driver behavior from the perspective of vehicle speed and lateral position, Fitzpatrick et al. determined that the wider the RCZ, the faster the vehicle speed. As the RCZ width increases, drivers tend to drive closer to the edge of the road [11, 12].

In the setting method of the RCZ, a safety benefit-cost analysis system called ROADSIDE was developed by the first version of the RDG for roadside designers on specific sections [13]. Road design decision-makers need to weigh the risk of injury and death to traffic participants against the cost of installing and maintaining safety facilities. Based on the National Cooperative Highway Research Program (NCHRP) 22-9 and 22-9 (2) projects, the first Roadside Safety Analysis Program (RSAP) was developed to assess the effectiveness of roadside safety improvement after the publication of the RDG (2002 version). The RSAP is designed based on the method of "encroachment probability" and consists of two complete procedures: the user interface procedure, which provides a user-friendly environment for data input and results review, and the main analysis procedure, which includes benefit-cost procedures and algorithms. Compared with the ROADSIDE system, the RSAP has shown significant improvement in how encroachments and final crashes are allocated by using a random solution rather than a deterministic method [14].

Furthermore, Ayati and Shahidian proposed an optimal RCZ width by weighting the engineering costs and safety improvement benefits resulting from increasing the roadside area. The results show that for curves with a radius less than 195 m, the wider roadside recovery zone is a key factor that affects roadside safety [15]. Considering the lack of land resources, Ayati and Shahidian implemented a reasonable allocation of limited resources among different roadside safety schemes in the form of charts through resource surveys [16]. According to the established chart, road design engineers could choose to install a guardrail or set a gentle embankment slope from the two aspects of roadside safety and economic safety. By proposing a series of traffic protection measures and considering various factors, including average collision cost, AADT, and discount rate, Roque and Cardoso discussed the comprehensive safety benefits brought by implementing a variety of measures and the generated investment cost, operation cost, and maintenance cost. They then selected an alternative scheme according to the proposed incremental benefit-cost ratio [17].

In roadside safety design, there is a lack of discussion on a comprehensive consideration of benefits involving the roadside slope, which is not consistent with the actual state. In light of the lack of land resources and social stability problems that result from land requisition, the concept of a tolerance roadside design is impractical to effectively implement in all regions. Therefore, the RCZ should be set according to the land use indexes of different grade roads, the probability of roadside accidents, social stability, engineering costs, safety benefit, and so on. However, existing technical specifications, such as Technical Standards of Highway Engineering (JTG B01-2014) and Design

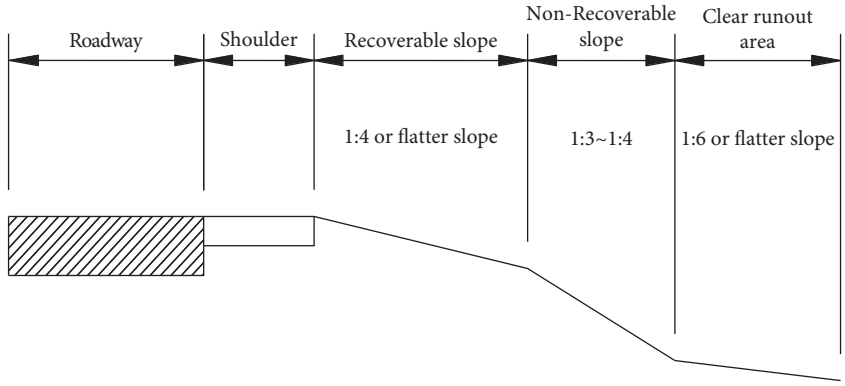


Figure 1: Setting standard of the RCZ slope.

Specification for Highway Alignment (JTG D20-2017) of China do not involve setting conditions for the RCZ. In addition, although China's Specifications for the Highway Safety Audit (JTG B05-2015) and the fourth edition of the RDG give a method for selecting the RCZ width in the curved section, they do not provide a theoretical basis. Therefore, in terms of the roadside safety design, there is still a lack of an accurate quantitative calculation of the RCZ width and the RCZ design method based on a benefit-cost analysis. How to develop a set of reasonable and effective methods for roadside safety design in order to reduce the frequency and severity of roadside accidents is an important research topic.

According to the research status and gaps of roadside safety, this paper uses PC-Crash software to conduct accident simulation analysis and proposes the recommended width of the RCZ in straight and curved sections for different operating speeds and curve radii by recording the vehicle's track after entering the roadside. Referring to our previous research results and data [18, 19], this paper utilizes a binary logistic regression analysis and multiple regression analyses to construct the probability prediction model of roadside accidents and an occupant injury risk assessment model. Based on the above research results and on the premise of considering the highway land use index, this paper puts forward the setting conditions for the RCZ, constructs the calculation model of the safety benefit brought by setting the RCZ, and gives the engineering cost calculation model from the two aspects of earthwork costs and land requisition costs. On the basis of a benefit-cost ratio analysis and the social stability risk assessment of a highway construction project, a set of RCZ design methods is therefore proposed in this paper.

2. Calculation of the RCZ Width

By analyzing the main factors influencing the design of the RCZ width, including the shoulder width, slope gradient, braking response time, departure speed, and departure angle, this paper uses the PC-Crash software to perform the simulation test. According to the track of vehicles after entering the roadside, the RCZ width is determined by calibrating the lateral distance between the right front of the vehicle and the edge of the carriageway.

2.1. Influence Factors. After a vehicle enters the roadside, most drivers will instinctively perform emergency braking, so the process of vehicles traveling can be equivalent to moving at a constant speed and moving at a uniform deceleration in kinematics. However, the distance required for a vehicle from the departure from the lane to a safe stop depends on a variety of factors, such as the shoulder width, slope gradient, braking response time, braking deceleration, departure speed, departure angle, and so on. Thus, before the simulation test, the values of the above factors should be discussed and analyzed.

2.1.1. Shoulder Width and Slope Gradient. The RCZ is generally composed of the shoulder and a gentle slope, so the width of the RCZ includes the shoulder width and the width of the slope clear zone (SCZ), as shown in Figure 2. The Specifications for Highway Safety Audit (JTG B05-2015) of China stipulates that when the slope gradient is less than 1:6, the entire slope width can be considered to be a safe area [5]. Therefore, this simulation test takes a slope gradient of 1:6 as the test condition when conducting research on the RCZ width. In addition, the shoulder width is determined in accordance with the relevant provisions in the Technical Standard of Highway Engineering (JTG B01-2014) [20].

2.1.2. Braking Response Time. The braking response time refers to the time required for the driver to identify the danger, the right foot to switch from the accelerator pedal to the brake pedal, and the vehicle to generate a braking force. Through an experimental detection of the drivers' decision time, Li et al. determined that most drivers' decision times are less than 0.5 s, and 95% of drivers' decision times are less than 1.0 s. When the driver identifies the danger, it takes approximately 0.2 s for the right foot to move from the accelerator pedal to the brake pedal and approximately 0.1 s from when the brake pedal is pressed to when the vehicle begins braking [21]. Therefore, the braking response time is approximately 0.8 s–1.3 s.

2.1.3. Braking Deceleration. The braking deceleration taken by drivers in the face of emergencies is greatly affected by the adhesion coefficient, which mainly depends on the road

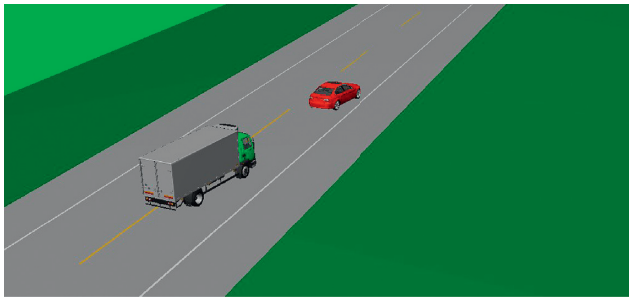


Figure 2: Test scenario.

material and road conditions. According to the stipulation of identification for the speed of vehicle involved in road traffic accident, a hard shoulder is generally asphalt pavement with an adhesion coefficient of 0.7; a dirt shoulder is generally sandy soil or soil; and its adhesion coefficient can be set at 0.55 [22]. Since most of the slope surface is covered by vegetation and the moisture content will affect the soil strength and make its adhesion coefficient change, so 0.55 can be used under dry soil conditions.

2.1.4. Departure Speed and Departure Angle. Many factors determine the angle of the vehicle leaving the lane. Most drivers will suddenly turn the steering wheel once they recognize danger. Therefore, the actual departure angle of the vehicle is of great randomness and uncertainty, which cannot be calculated quantitatively. Since there is currently no relevant research in China, according to the European RISER project on the relationship between vehicle departure angles and departure speeds in actual roadside accidents [1], the departure angles corresponding to different departure speeds can be approximately estimated, as shown in Table 1.

2.2. Simulation Test. PC-Crash is a computer simulation software developed by Austrian scholars. With powerful functions and a simple operation, this software is widely used in the world for simulations and reconstruction analyses of traffic accidents. PC-Crash software is based on the theory of energy conservation, momentum conservation, and Newton's three laws and combines the vehicle's rigid body, the road, and a multirigid body module to build the vehicle collision simulation model, which has the following advantages: (1) it allows vehicles to enter the roadside; (2) it contributes to achieving quantitative and accurate calculation of RCZ width on the basis of the assigned variables and distance calibration function; and (3) it is widely used in forensic identification and liability determination of traffic accidents, so it can recreate the accident process more realistically. For example, Cheng et al. carried out the test of a vehicle entering the roadside by PC-Crash software and predicted the probability of roadside accidents for curved sections on highways [18]. Zou et al. used PC-Crash to reconstruct freely available real-world vehicle-pedestrian video collisions and explored the cause mechanism of pedestrian injuries [23]. Wu et al. utilized a multibody system model in PC-Crash to reconstruct the elderly vulnerable road users' kinematics and head injuries [24]. These studies have proven that PC-Crash software has high accuracy in simulating traffic accidents.

Table 1: Departure angle.

Variable	Value								
Departure speed ($km \cdot h^{-1}$)	40	50	60	70	80	90	100	110	120
Departure angle (°)	12	10	9	8	7	6	5	5	4

2.2.1. Test Scenario Construction. To improve the reliability of the research results, before the simulation test, we selected a case of a roadside accident to calibrate the simulation scene of the PC-Crash software. The accident section is a two-way four-lane road with a lane width of 3.75 m, a shoulder width of 3 m, a horizontal curve radius of 600 m, a slope gradient of 1:6, a subgrade height of 0.7 m, dry asphalt pavement, and a slope covered with vegetation. According to traces at the accident scene, the departure angle of the vehicle was approximately 6°, and the departure distance was approximately 44.3 m. These findings, combined with the driver's confession, indicate that the vehicle speed was approximately 90 km/h. After the vehicle entered the roadside, the driver braked and took a slight left turn, as shown in Figure 3(a). PC-Crash software is first used to construct the road model, and various road parameters are set according to the characteristics of the accident section, with the adhesion coefficients of the road and slope set as 0.7 and 0.55, respectively. Subsequently, the test vehicle conforming to the accident vehicle is selected from the simulation software, and the speed, departure angle, braking, and steering of the test vehicle are set according to the accident information, with the braking response time set as 1.3 s. Finally, the simulation results show that the departure distance of the test vehicle is approximately 44.5 m, and its stop position and pose are basically consistent with those of the accident vehicle (as shown in Figure 3(b)). The effectiveness of the simulation scene is therefore verified.

According to the calibrated simulation scene, for straight sections, this paper uses PC-Crash to establish a two-way two-lane road model with a lane width of 3.75 m, a slope gradient of 1:6, and a shoulder width of 3 m. For the curved section, referring to the Design Specification for Highway Alignment (JTG D20-2017) [25], the horizontal curve radius is set as 200 m, 300 m, 400 m, 500 m, and 600 m based on the above road model. BMW-116d Autom and ASCHERSLEBEN KAROSS are selected as the representative models for the car and truck, respectively, and their parameters are set in accordance with the literature [18, 19]. The constructed test scenario is shown in Figure 2.

The vehicle fully brakes by setting the simulation motion sequence module. The braking response times are set at 0.8 s and 1.3 s, and the departure speeds are set as 50 km/h, 60 km/h, 70 km/h, 80 km/h, 90 km/h, 100 km/h, 110 km/h, and 120 km/h. The departure angle is set according to Table 1.

2.2.2. Test Process. According to the above variable values, $2 \times 8 = 16$ test conditions can be constructed. In order to ensure the accuracy of the simulation results, each vehicle

(a) (b)

Figure 3: Simulation calibration: (a) accident scene and (b) simulation scene.

type is simulated repeatedly 5 times under each test condition. The two vehicle types could be simulated 10 times, and the average value of the results is finally calculated. After each simulation, the ranging tool of PC-Crash software is used to measure the lateral distance between the right front of the vehicle and the road edge, that is, the guaranteed width of the SCZ. Combined with the shoulder width of a specific section, the recommended width of the RCZ can be obtained, as shown in Figure 4.

2.3. Recommended RCZ Width. According to the above test process, $2 \times 8 \times 10 = 160$ simulation tests are conducted in the straight section, and $2 \times 8 \times 10 \times 5 = 800$ simulation tests are conducted in the curved section. When determining the standard of the RCZ width, the vehicle departure speed should take the operating speed as the basis, rather than the design speed.

2.3.1. RCZ Width in the Straight Section. By calculating the average value of the test results under different test conditions, for the straight section, Table 2 gives the recommended width of the RCZ corresponding to eight operating speeds when the shoulder width is 3 m. The upper and lower limits of the RCZ width in Table 2 are the measured average values corresponding to brake response times of 0.8 s and 1.3 s, respectively. The upper limit should be taken into consideration when the road designer sets the RCZ, which can significantly improve the roadside safety level. Since the driver's basic braking response time is 0.8 s, the setting of the RCZ should at least meet the requirements of the lower limits.

2.3.2. RCZ Width in the Curved Section. For the curved section, according to the collected simulation data, Figures 5 and 6 show the relationships between the departure speed and the SCZ width and between the horizontal curve radius and the SCZ width, respectively. As can be seen from Figure 5, with an increase in the departure speed, the width of the SCZ gradually increases, and the increasing amplitude becomes larger and larger; the width tends to be dispersed from concentration. It can be seen from Figure 6 that with an increase in the horizontal curve, the width of the SCZ gradually decreases, and the decreasing amplitude becomes smaller and smaller; the width tends to be concentrated from dispersion.

Referring to the trends analysis of the departure speed, the horizontal curve radius, and the SCZ width, SPSS software is used to fit the calculation models of the lower and upper limits of the RCZ width m in the curve segment, as shown in Table 3. The correlation coefficient R^2 is usually used to test the fitting goodness of the model; its value is between 0 and 1. The higher the R^2 is, the better the fitting goodness of the model is, that is, the stronger the ability of the model is to explain the dependent variable. By comparing the R^2 of different models, models (b) and (f) are the optimal models for the lower and upper limit value of the RCZ width, respectively. As can be seen from these two optimal models, the width of the RCZ has an exponential relation with the departure speed and a power relation with the horizontal curve radius.

3. Setting Conditions for the RCZ

3.1. Overall Framework. This paper considers the limitations of the highway land use index (related to the land scope of the SCZ l). According to the subgrade height h and shoulder width w and combined with the recommended width m of the RCZ obtained from this paper, the setting conditions of the RCZ are studied. For sections equipped with the RCZ setting conditions, based on our previous research results and data, the probability model of roadside accident occurrence and the occupant injury risk assessment model are first fitted; then the calculation model of the safety improvement benefit brought by the RCZ is constructed. Finally, the engineering cost caused by the RCZ is calculated from the two aspects of earthwork costs and land requisition costs. If the benefit-cost ratio is great than 1, then the social stability risk generated by setting the RCZ is further evaluated. If the social stability risk is less than 0.36, it is recommended that the RCZ be set; otherwise, it is unadvisable to set the RCZ. The design idea of the RCZ is shown in Figure 7.

Given the shortage of land resources and the limitation of a land acquisition index in China, it is difficult to realize the design of the RCZ in all regions. Therefore, before setting the RCZ, it is necessary to consider whether the conditions for setting the RCZ are available on the roadside. According to the relevant provisions of the Specifications for Highway Safety Audit (JTG B05-2015), when the slope gradient is 1:6 or lower, the entire slope width can be used as a safe clear zone [5]. Therefore, h and l should satisfy the following relationship:

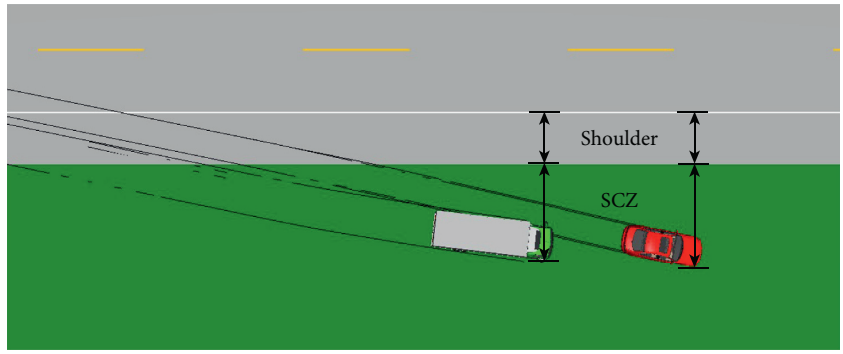


Figure 4: Definition of the RCZ width.

Table 2: Recommended width of the RCZ in the straight section where the shoulder width is 3 m.

Operating speed (km·h^{-1})	Recommended width of the RCZ (m)
50	3.085 (0.085)~3.098 (0.098)
60	4.055 (1.055)~4.168 (1.168)
70	5.2001 (2.001)~5.260 (2.260)
80	5.795 (2.795)~6.052 (3.052)
90	6.380 (3.380)~6.507 (3.507)
100	6.610 (3.610)~7.201 (4.201)
110	8.355 (5.355)~9.420 (6.420)
120	9.645 (6.645)~11.532 (8.532)

Note. The width of the SCZ is in brackets. When setting the RCZ, the width of the RCZ can be determined according to different shoulder widths.

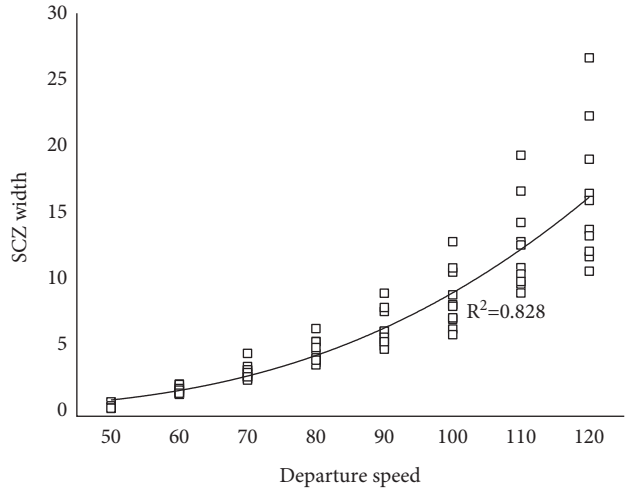


Figure 5: Relationship between the SCZ width and departure speed.

$$\frac{h}{l} \leq 1{:}\ 6 \approx 0.167, \tag{1}$$

where l depends on the highway land scope L, which is composed of the subgrade width W (including w), l, the berm width l_1, the width of side ditch l_2, and the width of the outer area of side ditch l_3, as shown in Figure 8. According to the provisions of China's Highway Engineering Project Construction Land Index (Construction Standard [2018] 124) [26] and Design Specification for Highway Alignment

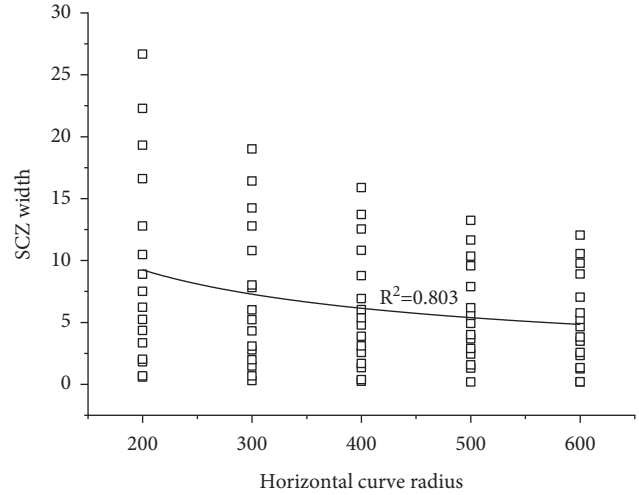


Figure 6: Relationship between the SCZ width and horizontal curve radius.

(JTG D20-2017) [25], for the subgrade width and land scope corresponding to highways of different grades, l can be calculated as follows:

$$l = \frac{L - W - l_1 - l_2 - l_3}{2}. \tag{2}$$

In practical engineering implementations, the land scope of the highway is often limited due to the impact of structures or residential areas along the road, and it is difficult to fully comply with the regulations. Therefore, the actual land scope should be determined according to the specific conditions on the roadside. In addition, in the case of conforming to formula (1), the RCZ should also meet the following relationships:

$$l + w \geq \text{lower limit value of } m. \tag{3}$$

That is, the RCZ width should not be less than the lower limit of the recommended RCZ width obtained in Section 2.

3.2. Safety Benefit Calculation Model. Although the RCZ can bring considerable traffic safety benefits, given that the setting of the RCZ will generate additional engineering costs and land requisition costs, setting the RCZ should be considered on the basis of measuring safety benefits and

Table 3: Model building.

No.		Models	R^2
a		$m = 1.05 \times 10^{-5} v^3 - 0.032R - 3.49 \times 10^{-14} R^2 + w$	0.674
b		$m = 387.49e^{-(395.448/(v-1.44))} + 5.027 \times 10^6 (1+R)^{-2.669} + w$	0.919
c	Lower limit value (braking reaction time is 0.8 s)	$m = 0.003v^2 - 0.208v - 70.717(1+R)^{0.044} + 96 + w$	0.790
d		$m = -2.786e^{-0.013v} - 0.029R + 2.339 \times 10^{-5} R^2 + w$	0.587
e		$m = -7.09 \times 10^5 e^{(2.3\times10^4)/(v-2.7\times10^3)} - 0.035R + 2.62 \times 10^{-5} R^2 + w$	0.864
f		$m = 284.711e^{-(314.869/(v-11.273))} + 2.888 \times 10^6 (1+R)^{-2.52} + w$	0.911
g		$m = 1.06 \times 10^{-5} v^3 - 83.34(1+R)^{0.038} + 104.18 + w$	0.857
h	Upper limit value (braking reaction time is 1.3 s)	$m = 0.0025v^2 - 0.204v - 0.032R + 2.644 \times 10^{-5} R^2 + 12.747 + w$	0.898
i		$m = 21.442e^{-(86.63/v)} - 0.033R + 2.654 \times 10^{-5} R^2 + w$	0.593
j		$m = -12.25e^{-22.95v} - 1.75 \times 10^{-8} R^3 + 8.121 + w$	0.479

Note. v is the departure speed, km·h^{-1}; R refers to the horizontal curve radius, m; and w presents the shoulder width, m.

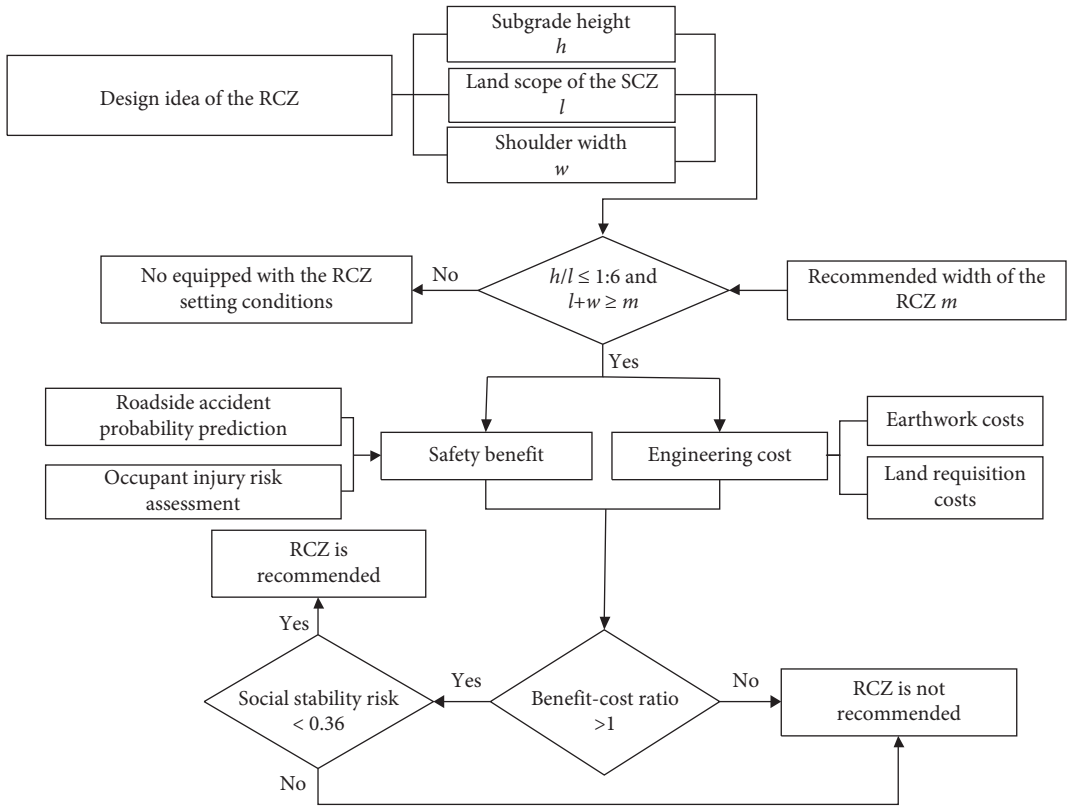


Figure 7: Design idea of the RCZ.

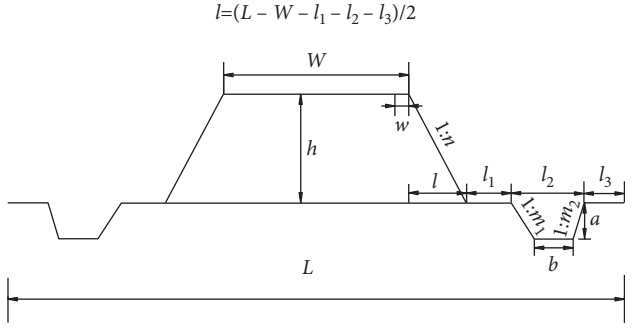


Figure 8: Highway land scope.

engineering costs for sections equipped with the RCZ setting conditions. If the benefit-cost ratio is great than 1, it can be further considered when setting the RCZ. In addition, when determining the probability prediction of roadside accidents and the occupant injury risk assessment, this paper uniformly collates and analyzes the accident data of different vehicle types from two studies [18, 19] in order to obtain the applicability of the research results to different vehicle types.

3.2.1. Probability Prediction of Roadside Accidents. Based on the obtained 12,800 roadside accident data and the identified significant risk factors (i.e., departure speed, horizontal

curve radius, vehicle type, adhesion coefficient, and shoulder width) by us in the early stage (see the literature [18]). This paper uses a binary logistic regression analysis to construct a probability prediction model of roadside accidents involving the horizontal curve radius, shoulder width, adhesion coefficient, and departure speed.

(1) Model Building. According to the obtained data of roadside accidents, this paper takes the occurrence of roadside accidents as the dependent variable, which is divided into two grades: $y = 1$ (roadside accidents occur) and $y = 2$ (no roadside accidents occur). It then takes the departure speed v, horizontal curve radius R, shoulder width w, and adhesion coefficient μ as independent variables and uses SPSS software to perform a binary logistic regression analysis, as shown in Table 4.

It can be seen from Table 4 that the above four variables are significantly correlated with the occurrence of roadside accidents. The following regression equation can be constructed according to the parameter estimates:

$$\ln\left(\frac{P_1}{P_2}\right) = -4.085 + 0.078v - 2.7 \times 10^{-4}R - 4.011\mu - 1.515w, \tag{4}$$

where P_1 and P_2 are the prediction probabilities of occurrence and nonoccurrence of roadside accidents, respectively. Formula (4) can be converted into the cumulative logistic probability prediction model, which represents the occurrence of roadside accidents:

$$P(y \le 1) = \frac{\exp\left(-4.085 + 0.078v - 2.7 \times 10^{-4}R - 4.011\mu - 1.515w\right)}{1 + \exp\left(-4.085 + 0.078v - 2.7 \times 10^{-4}R - 4.011\mu - 1.515w\right)}. \tag{5}$$

Model (5) is mainly used to predict the probability of roadside accidents in the curved section. It is stipulated in the Design Specification for Highway Alignment (JTG D20-2017) that the maximum radius of the curve should not exceed 10,000 m [25]. In order to realize the probability prediction of roadside accidents in the straight section, this paper defines that model (5) is applicable to a straight segment when R is set as 10,000 m.

(2) Model Verification. According to the model prediction results, the ROC curve is drawn using SPSS software, as shown in Figure 9. As can be seen from Figure 9, the ROC curve is very close to the upper left corner, and the area under the curve (AUC) is $0.969 > 0.9$, indicating that the fitting effect of the probability prediction model of roadside accidents is good, with a confidence level of 95%.

3.2.2. Risk Assessment of Occupant Injuries in Roadside Accidents. Based on our previous collected 1,500 data of roadside accidents (see the literature [19]), by considering the departure speed, horizontal curve radius, slope gradient, and subgrade height, this paper uses multiple regression analyses to fit the occupant injury risk assessment model of roadside accidents based on the acceleration severity index (ASI) and utilizes a Fisher optimal segmentation method to rationally classify the occupant injury risk levels.

(1) Model Building. According to the relationships between the ASI and various risk factors explored in the literature [19], by comparing the correlation coefficients R^2 of different models, the optimal model of the occupant injury risk assessment for the straight and curved sections are respectively fitted as follows:

Straight section:

$$\mathrm{ASI} = 0.008v + 0.123h + 4.76 \times 10^{-5}e^{0.033\alpha} + 0.876. \tag{6}$$

Curved section:

$$\mathrm{ASI} = 0.01v + 0.127h + 0.306e^{0.0419\alpha} + 7.832R^{-0.165} - 2.392, \tag{7}$$

where α is the slope gradient, °. Among them, the correlation coefficient R^2 of model (6) is 0.976, and that of model (7) is 0.963.

(2) Classification of Occupant Injury Risk. This paper uses a Fisher optimal segmentation algorithm to classify the ASI data in order to obtain the classification scheme of roadside accidents risk, including the optimal number of classifications and the corresponding threshold of each level. Fisher's principle of optimal segmentation is to ensure a minimum sum of squares of the intragroup deviation of the segmented ordered sample data, and the corresponding group has optimal segmentation.

Based on the 1,200 data groups obtained from the literature [19], the data for ASI ≤ 1 are screened as a group, and the risk is set to level I. The remaining 269 groups of data are numbered by the order of the ASI values from small to large to generate ASI-ordered samples, which are denoted as x_i ($i = 1, 2, \ldots, 269$). For the specific segmentation steps of ordered samples, refer to the literature [19]. The algorithm code is written by MATLAB software to calculate the category diameter and minimum error function of the ordered sample, and the curve of the minimum error function with different classification numbers k is drawn, as shown in Figure 10.

Table 4: Estimate of the parameters.

Variable	Parameter	Standard deviation	Wald	Degree of freedom	Significance
Departure speed	0.078	0.003	1,877	1	<0.01
Horizontal curve radius	$-2.7e-04$	0.001	1,749	1	<0.01
Adhesion coefficient	−4.011	0.186	466	1	<0.01
Shoulder width	−1.515	0.073	434	1	<0.01
Constant	−4.085	0.211	373	1	<0.01

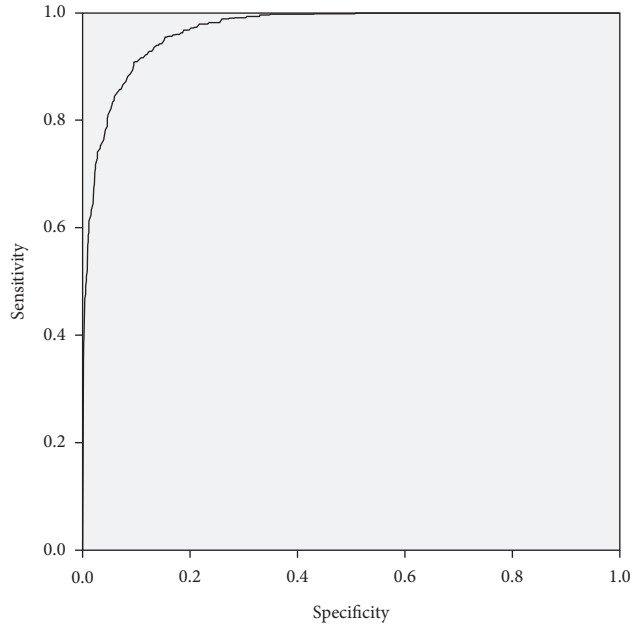


Figure 9: ROC curve.

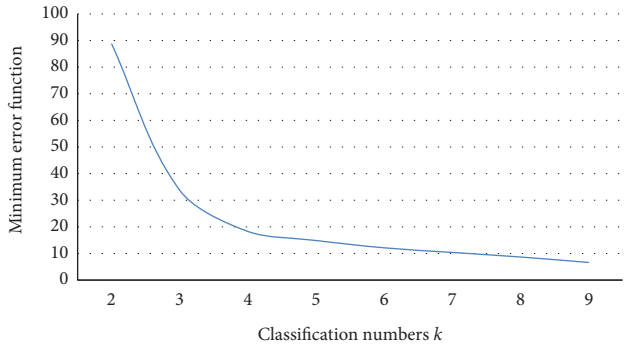


Figure 10: Relationship between the minimum error function and classification number.

As shown in Figure 10, when $k = 3$ and 4, the curvature of the ordered sample significantly varies. To further determine the optimal classification number, the ratios of the minimum error function $\beta(k)$ of the adjacent classification number are calculated in Table 5.

According to Table 5, β (3) is greater than β (4). Thus, $k = 3$ is the optimal classification number for the ordered samples. Additionally, the codes that correspond to the optimal classification can be obtained from Table 5, so the corresponding injury index threshold can be determined. Based on the above research results and considering the classification of occupant injuries in traffic accidents in China, the criteria for the occupant injury risk assessment are proposed, as shown in Table 6.

3.2.3. Safety Improvement Benefit Calculation

(1) Number of Roadside Accidents. According to the probability prediction model of roadside accidents constructed in Section 3.2.1, the following prediction model of the number of roadside accidents is proposed by introducing AADT:

$$A = \mathrm{AADT}\left(\frac{\exp\left(-4.085 + 0.078v - 2.7 \times 10^{-4}R - 4.011\mu - 1.515w\right)}{1 + \exp\left(-4.085 + 0.078v - 2.7 \times 10^{-4}R - 4.011\mu - 1.515w\right)}\right), \tag{8}$$

Table 5: Classification results.

k	Minimum error function	Classification	β
2	88.675	{1~156} {157~269}	—
3	32.521	{1~89} {90~156} {157~269}	1.74
4	18.671	{1~43} {44~89} {90~156} {157~269}	1.24
5	15.013	{1~43} {44~89} {90~156} {157~202} {203~269}	—

Table 6: Criteria for the occupant injury risk assessment.

Risk level	ASI threshold value	Occupant injury level
I	≤1	Noninjury
II	(1, 1.31]	Moderate injury
III	(1.31, 1.78]	Disabling injury
IV	>1.78	Fatal injury

where A is the number of roadside accidents.

(2) Reduction of Occupant Injuries per Accident Brought by the RCZ. The reduction of occupant injuries brought by setting the RCZ can be calculated quantitatively by the occupant injury risk level, that is: occupant injury risk level without setting the RCZ – the occupant injury risk level with setting the RCZ. Since setting the RCZ can effectively avoid the occurrence of roadside accidents, this paper assumes that the setting of the RCZ will not cause casualties of passengers in roadside accidents, that is, the risk is level I (as shown in Table 6). Therefore, the reduction of the occupant injuries per accident brought by the RCZ can be equivalent to the occupant injury risk level per accident without setting the RCZ, which can be obtained according to the occupant injury risk assessment model constructed in Section 3.2.2. Additionally, according to the Specifications for Highway Safety Audit (JTG B05-2015), when the slope gradient is steeper than 1:3.5, it cannot be regarded as an effective safety clear zone [5]. Therefore, this paper takes the slope gradient of 1:3.5 as the basic index for not setting the RCZ.

(3) Standard Economic Losses per Accident at Different Occupant Injury Risk Levels. There are currently few studies on the statistics of casualties, medical compensation, and property losses of traffic accidents in China. Therefore, this paper collects a total of 347 economic losses of casualties in traffic accidents of different severity via a questionnaire survey and telephone inquiry among lawyers, traffic polices, and insurance practitioners, as shown in Table 7. According to the statistical analysis, the average economic loss of moderate injury per accident is approximately 60,000 yuan; the average economic loss of disabling injury per accident is approximately 120,000 yuan; and the economic loss of fatal injuries per accident is approximately 370,000 yuan.

Based on the above analysis, the calculation model for a safety improvement benefit brought by setting the RCZ is given as follows: AADT × the occurrence probability of roadside accidents × reduction in occupant injuries per accident brought by the RCZ × standard economic losses per accident at different occupant injury risk levels.

3.3. Engineering Cost Calculation Model. By establishing the calculation models of earthwork costs and land requisition costs, this paper calculates the difference in the economic costs with and without the RCZ in order to obtain the additional engineering costs caused by setting the RCZ.

3.3.1. Earthwork Costs. It is assumed that the slope gradient is $1{:}n_1$ in case the roadside is not set in a clear zone; otherwise, the slope is $1{:}n_2$, as shown in Figure 8. Then, the earthwork volume of the highway per kilometer (set as V) is as follows:

$$V = \frac{1}{2} \times 1000h[(W + 2nh) + W]. \tag{9}$$

The increment in the subgrade earthwork per kilometer generated by setting the RCZ is as follows:

$$\Delta V = V_2 - V_1 = 1000h^2(n_2 - n_1). \tag{10}$$

The average cost of the subgrade earthwork per cubic meter is set as T_0 yuan, so the cost of the subgrade earthwork is as follows:

$$T = T_0 \Delta V = 1000T_0h^2(n_2 - n_1). \tag{11}$$

3.3.2. Land Requisition Costs. According to the highway land scope shown in Figure 8, the area per kilometer of the highway (set as S) is as follows:

$$S = 1000[W + 2nh + 2l_1 + 2(m_1 + m_2)a + 2b + 2l_3]. \tag{12}$$

The growth in road area per kilometer generated by setting the RCZ is as follows:

$$\Delta S = S_2 - S_1 = 2000h(n_2 - n_1). \tag{13}$$

The actual land requisition costs in terms of construction engineering consist of opportunity costs, new resource consumption costs, and transfer payments. In the national economic evaluation, the opportunity cost and new resource consumption cost are adjusted according to the shadow price, while the transfer payment is not included in the land requisition costs, as shown in the following formula:

$$E_0 = \mathrm{OC} + Y = \mathrm{NB}_0(1 + g)^{\tau+1}\left[\frac{1 - (1 + g)^n(1 + i)^{-n}}{i - g}\right] + Y, \tag{14}$$

where E_0 is the land requisition cost, 10,000 yuan/mu; OC is the opportunity cost, 10,000 yuan/mu; Y is the new resource

Table 7: Statistics of economic losses caused by traffic accidents.

Occupant injury level	Economic losses (10,000 yuan)	Number of accidents	Percentage (%)
Moderate injury	0~2	26	13.98
	2~4	18	9.68
	4~6	54	29.03
	6~8	60	32.26
	8~10	21	11.29
	More than 10	7	3.76
Total	1049	186	100
Disabling injury	0~4	0	0.00
	4~8	1	1.02
	8~12	37	37.76
	12~16	36	36.73
	16~20	12	12.24
	More than 20	12	12.24
Total	1372	98	100
Fatal injury	0~12	0	0.00
	12~24	1	1.59
	24~36	20	31.75
	36~48	28	44.44
	48~60	10	15.87
	More than 60	4	6.35
Total	2594	63	100

consumption cost, 10,000 yuan/mu; NB_0 is the annual net benefit; g is the average annual net benefit growth rate; τ is the number of years from the base year to the engineering commencement year; n is the economic evaluation period; and i is the social discount rate. The land requisition cost per kilometer generated by setting the RCZ is as follows:

$$E = 0.0015E_0\Delta S. \tag{15}$$

3.4. Risks to Social Stability. Since the problems of land expropriation and demolition, farmers' burden, environmental damage, and social security caused by engineering construction tend to arouse social contradictions, the risk assessment of social stability should be determined before the engineering implementation. According to the relevant provisions of the Compilation Outline and Explanation of Social Stability Risk Assessment Reports of Major Fixed Asset Investment Projects (Trial; Issue and Reform Investment No. [2013] 428) issued by the National Development and Reform Commission (NDRC) of China [27], this paper adopts a comprehensive risk index method (CRIM) to evaluate the social stability risk caused by setting the RCZ. The specific steps are as follows:

(1) Through the analysis of the feasibility study report and the design data of the project, the significant risk factors of setting the RCZ are identified from the four aspects of legality, rationality, feasibility, and controllability of the engineering construction

(2) According to data analysis and questionnaire survey, the probability-impact matrix (PIM; as shown in Figure 11) is used to evaluate and predict the risk level of each risk factor

(3) Referring to the rules of social stability risk assessment and the expert experience of similar highway construction engineering projects, the weight of each risk factor is determined

(4) According to the weight and risk level of each risk factor, a comprehensive risk index of the engineering project is obtained via weighted calculation

(5) According to the comprehensive risk level, whether or not the engineering project should be implemented will be determined

PIM is also called the risk assessment matrix, which takes the possibility of the occurrence of risk factors as the abscissa and the impact degree of risk factors after the occurrence as the ordinate. It is mainly used to analyze specific risk points and determine their risk levels.

4. Case Analysis

Huashan south road in the Coastal Advanced Equipment Industry Cluster area in southern Zhejiang is in urgent need of renovation because of the increasing traffic volume. According to the preliminary design data of the project, the reconstructed highway adopts the design standard of a first-grade highway with two-way six-lane traffic, a designed speed of 80 km/h, and a subgrade width of 44 m. According to the traffic volume prediction results, the annual one-way AADT is approximately 5,000. K2 + 960~K16 + 358 are the filling sections; considering that there are structures and residential areas on the roadside, the design of the RCZ should be studied. According to the speed calculation model of the Specifications for Highway Safety Audit (JTG B05-2015) [5], the operating

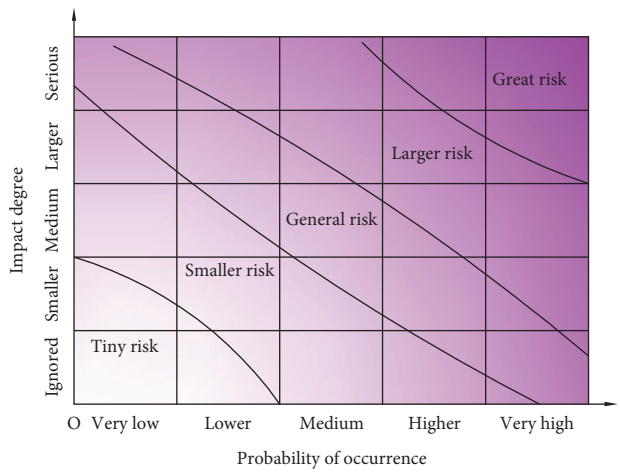


Figure 11: PIM.

Table 8: Results.

No.	Mileage peg	v (km·h^{-1})	R (m)	w (m)	h (m)	h/l	Lower limit value of m (m)	Allowable RCZ width (m)	Benefit-cost ratio (benefit/cost/ 10,000 yuan)	Social stability risk	Setting the RCZ
1	K2 + 960~K4 + 070	105	+∞	3	1.65	0.161	7.48	10.97	3.18 (379/119)	0.216	Yes
2	K4 + 070~K5 + 182	99	+∞	3	1.43	0.166	6.59	10.97	1.59 (196/123)	0.216	Yes
3	K5 + 182~K6 + 225	88	3,900	3	1.78	0.223	—	—	—	—	No
4	K6 + 225~K7 + 330	88	3,900	3	2.23	0.280	—	—	—	—	No
5	K7 + 330~K8 + 416	83	3,500	3	2.51	0.315	—	—	—	—	No
6	K8 + 416~K9 + 550	70	3,800	3	1.63	0.151	5.41	10.97	1.13 (124/110)	0.216	Yes
7	K9 + 550~K10 + 615	75	3,800	3	1.34	0.164	5.79	10.97	0.66 (81/122)	—	No
8	K10 + 615~K11 + 718	69	1,800	3	1.08	0.136	5.52	10.97	0.89 (87/97)	—	No
9	K11 + 718~K12 + 800	67	2,000	3	1.89	0.237	—	—	—	—	No
10	K12 + 800~K13 + 910	73	2,000	3	1.14	0.143	5.55	10.97	2.16 (224/103)	0.216	Yes
11	K13 + 910~K15 + 105	69	2,000	3	1.43	0.179	—	—	—	—	No
12	K16 + 105~K16 + 358	69	1,500	3	1.05	0.132	5.78	10.97	2.00 (188/94)	0.216	Yes

speed of each section is predicted. Table8 summarizes the design indexes and operating speeds of each section.

4.1. Determination of the Setting Condition for the RCZ. According to the cross-sectional design data of the project, the width of the berm, side ditch, and outer area of the side ditch are 1 m, 2 m, and 1 m, respectively. By referring to the provisions of the highway land scope in China's Highway Engineering Project Construction Land Index (Construction Standard [2018] 124) [26], formula (4) is used to calculate the land scope of the SCZ: $l = (63.935 - 44 - 1 - 2 - 1)/2 = 7.9675$ m. According to the subgrade height of each section (as shown in Table 8), the calculation results show that sections 1, 2, 6, 7, 8, 10, and 12 conform to the relationship shown in formula (1). In addition, if the RCZ is set, the RCZ width should be further verified to meet the recommended width given in Section 2. According to formula (3), the allowable RCZ widths of sections 1, 2, 6, 7, 8, 10, and 12 are all larger than the corresponding lower limit widths. Among them, the recommended RCZ width in the straight section is obtained via the line interpolation method according to Table 2.

Aimed at the above 7 sections that meet the conditions for setting the RCZ, the safety benefit calculation model and engineering cost calculation model are used to calculate the safety improvement benefit and additional engineering cost of each section brought by the RCZ and further consider whether to set the RCZ by analyzing their benefit-cost ratios.

4.2. Calculations of the Safety Benefit and Engineering Cost. In the process of the safety improvement benefit assessment, when calculating the probability of roadside accidents, the adhesion coefficient is set to 0.7 because the reconstructed road is dry asphalt pavement. In the process of engineering cost accounting, according to the relevant provisions of the Specifications for Highway Safety Audit (JTG B05-2015) [5], the slope gradient should be 1:3.5 when there is no RCZ and 1:6 when there is an RCZ.

According to the project feasibility report, the average earthwork cost is proposed to be 55 yuan/m^3, so the total earthwork costs can be calculated by formula (11). Since rice is mainly planted along the road, according to the relevant regulations of the local price bureau, the annual net benefit is

Table 9: Significant risk factors.

Category	Risk factors
Legality	Policy risk
	Legal risk
Rationality	Ecological environment risk
Feasibility	Land requisition and demolition risk
Controllability	Social security risk
	Public opinion risk

Table 10: Statistical results of the questionnaire survey.

Question			Frequency	Percentage (%)
Ecological environment risk	What do you think is the ecological impact of setting the RCZ?	Beneficial	61	53.04
		Nothing	49	42.61
		Harmful	1	0.87
		Unknown	4	3.48
Land requisition and demolition risk	Are you satisfied with the current way of land requisition and demolition?	Satisfied	90	78.26
		Dissatisfied	6	5.22
		General	19	16.52
	Are you satisfied with the current compensation standard?	Satisfied	88	76.52
		Dissatisfied	7	6.09
		General	20	17.39
	To set the RCZ, if the land requisition and demolition continue according to the current way and compensation standard, can you accept it?	Yes	89	77.39
		No	7	6.09
		Unconcerned	19	16.52

1,106 yuan/mu, and the average annual net benefit growth rate is 6%. In addition, the base year for engineering is the commencement year, so τ is set as 0. Referring to the relevant provisions of the Economic Evaluation Methods and Parameters of Construction Projects (third edition) [28], the economic evaluation period takes 25 years, and the social discount rate is 8%. In this project, new resource consumption costs mainly include elderly maintenance, endowment insurance, relocation compensation, resettlement fees, and compensation for attachments to land, whose prices are taken as 5,400 yuan/mu, 1,000 yuan/mu, 34,800 yuan/mu (adjusted by the shadow price conversion factor 1.1), 16,200 yuan/mu, and 6,000 yuan/mu, respectively, for a total of 63,400 yuan/mu. According to formulas (14) and (15), the land requisition cost can be calculated.

Table 8 shows the safety benefit and the engineering costs of the seven sections. Among them, the benefit-cost ratio of sections 1, 2, 6, 10, and 12 is >1, that is, the safety benefit is greater than the engineering cost, so they can be considered to set the RCZ. However, the risks to social stability caused by the setting of the RCZ in these sections should be further assessed.

4.3. Social Stability Risk Assessment. Combined with the construction background of the project and the construction characteristics of the RCZ, the significant risk factors of setting the RCZ are identified from the four aspects of legality, rationality, feasibility, and controllability, as shown in Table 9.

The risk level of each significant risk factor is evaluated and predicted by using the PIM as follows.

4.3.1. Policy Risk. According to the project planning, the RCZ is included in the road structure and conforms to the relevant provisions of the Specifications for Highway Safety Audit (JTG B05-2015). Therefore, the RCZ complies with the national industrial policy; the probability of policy risk is very low; and the impact degree is medium, so the risk is considered a smaller risk.

4.3.2. Legal Risk. The decision-maker of this project is the Wenzhou municipal government, and relevant external supporting documents, land preexamination, and environmental impact assessments have all passed the examination. From the perspective of legal procedure, the overall construction of this project (including the RCZ) meets the requirements of current laws and regulations, so the probability of legal risk is very low and the impact degree is medium, which is considered to be a smaller risk.

4.3.3. Risk of the Ecological Environment and Risk of Land Requisition and Demolition. For farmland, residential areas, and enterprises involved in land requisition and demolition of sections 1, 2, 6, 10, and 12, the questionnaire survey is conducted to collect public opinion on the environmental damage, land requisition, and demolition caused by the RCZ. The entire process of public participation follows the principle of representativeness and randomness. Considering that respondents under the age of 20 and over the age of 75 have biased understandings of the interview content, the above age groups are excluded. Finally, there were 68 males and 47 females in all the collected samples. The sample distribution basically reflects the overall intention of the

Table 11: Social stability risk assessment before project implementation.

Risk factor	Weight	Risk level					Risk index
		Tiny 0.04	Smaller 0.16	General 0.36	Larger 0.64	Great 1	
Policy risk	0.14		√				0.0224
Legal risk	0.13		√				0.0208
Ecological environment risk	0.15	√					0.006
Land requisition and demolition risk	0.37			√			0.1332
Social security risk	0.11		√				0.0176
Public opinion risk	0.10		√				0.016
Σ	1						0.216

survey group, and the location of the sample covers the area affected by the RCZ. The survey results are shown in Table 10.

As can be seen from the statistical results in Table 10, only 0.87% of all respondents believe that the setting of the RCZ is harmful to the ecological environment, so the probability of ecological environment risk is very low and the impact degree can be ignored; therefore, the risk is considered a tiny risk. Moreover, 5.22% and 6.09% of respondents are not satisfied with the methods of land requisition and compensation standards, respectively, which means that this part of the group cannot accept the following requisition. However, the remaining 94% of respondents are in favor of or do not oppose setting the RCZ, so the probability of land requisition and demolition risk is lower and the impact degree is larger, which is considered a general risk.

4.3.4. Social Security Risk. As the constructor has formulated a reasonable construction organization and supervision measures, strengthened close communication with the local government, and comprehensively created a safe, harmonious, and stable public security environment, the probability of a social security risk is very low, and the impact degree is larger, so this risk is considered to be a smaller risk.

4.3.5. Public Opinion Risk. Although the project has been strongly supported by the local residents, the construction of the RCZ will have a certain impact on the surrounding environment. Therefore, publicity and positive guidance are particularly important. The constructor performed a significant amount of publicity work through mainstream media and networks in the early stage, so the probability of public opinion risk is lower and the influence degree is medium. Therefore, this risk is considered a smaller risk.

Referring to the rules of some social stability risk assessments of highway construction projects and expert experience, the weight of each risk factor in the setting of the RCZ is determined. The weight and risk level of each factor are weighted to obtain the comprehensive risk index of the project, as shown in Table 11.

In general, when the comprehensive risk index is lower than 0.36, which indicates that the social stability risk level of the project is smaller (i.e., the majority of people understand and support the project), a small number of people have opinions on the project, but effective persuasion and guidance can prevent and resolve the conflicts. As can be seen from Table 11, the social stability risk caused by setting the RCZ is 0.216, which is a smaller risk. Therefore, sections 1, 2, 6, 10, and 12 are feasible for setting the RCZ.

5. Conclusion

This paper explores the main factors affecting the design of the RCZ width, including the shoulder width, slope gradient, braking response time, departure speed, and departure angle. PC-Crash software is used to establish the vehicle and road models to perform simulations of a vehicle entering the roadside by setting different departure speeds and curve radii. By recording the track of the vehicle after entering the roadside and measuring the lateral distance between the right front of the vehicle and the lane edge, the recommended widths of the RCZ in the straight and curved sections for different operating speeds and curve radii are given. The Fourth Edition of the RDG and China's Specifications for Highway Safety Audit (JTG B05-2015) provide the RCZ width determination method. For the straight section with fill and excavation subgrades, the Fourth Edition of the RDG provides a recommended RCZ width corresponding to different design speeds, AADT, and slope. For the curved section, it provides a correction coefficient for a curved section that is based on the horizontal curve radius, but this guide does not provide a specific calculation method. Referring to the Fourth Edition of the RDG, China's Specifications for Highway Safety Audit (JTG B05-2015) show how to determine the recommended RCZ width in the form of a graph. However, the calculation accuracy of the graphical method is inadequate, while the proposed method in this paper can accurately calculate the width of the RCZ via modeling.

Based on our previous research results and data, the probability prediction model of roadside accidents and the occupant injury risk assessment model are constructed by means of the binary logistic regression analysis, the multiple regression analysis, and the Fisher optimal segmentation algorithm, and the quantitative analysis of the probability and occupant injury risk of roadside accidents is realized.

Based on the above research results, this paper determines the setting conditions for the RCZ by considering the

land use index, the subgrade height, and the shoulder width. A safety benefit calculation model based on AADT, the occurrence probability, and the occupant injury risk of roadside accidents is constructed, and an engineering cost calculation model is given from the two aspects of earthwork costs and land requisition costs. Thus, an RCZ design method based on the safety benefit and the engineering cost is proposed. Based on the benefit-cost ratio analysis, the CIDM is adopted to evaluate the social stability risk of the project. Finally, the research results are applied by a case analysis. This result fills in the gaps of the Fourth Edition of the RDG and China's Specifications for Highway Safety Audit (JTG B05-2015) in the setting conditions of the RCZ by first involving social stability risk into the RCZ design.

The results of this paper are helpful for reducing the occurrence and loss of roadside accidents and can provide a reference and guidance for traffic control departments and highway designers to optimize the design of roadside safety in the future. However, because the recommended width of the RCZ obtained in this paper is based on the test condition of a safety slope of 1:6, in practical engineering implementations, the RCZ width should be adjusted when another slope gradient is adopted, but their relationship is not yet clear. Therefore, the precise quantitative calculation between the RCZ width and the safety slope gradient can be further determined. Additionally, since the research results of this paper are based on simulation data, certain variables are assigned values in advance according to professional standards and experience. Moreover, the human factor cannot be considered in the simulation software, which is an important factor affecting the occurrence of roadside accidents. Therefore, validation of the results and research on roadside design involving human factors should be implemented by collecting real accident cases in the future.

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 Fund for Less Developed Regions of the National Natural Science Foundation of China (No. 71861006), Guangxi Natural Science Foundation (No. 2020GXNSFAA159153), and Guangxi Science and Technology Base and Talent Special Project (No. AD20159035).

References

[1] RISER Consortium, *Guidelines for Roadside Infrastructure on New and Existing Roads. Roadside Infrastructure for Safer European Roads, Deliverable 6. 5th Research Framework Programme "Growth"*, European Commission, Brussels, Belgium, 2005.

[2] Federal Highway Administration (FHWA), *Roadway Departure Safety*, Federal Highway Administration (FHWA), Washington, DC, USA, 2018.

[3] Traffic Management Bureau of the Ministry of Public Security (TMBMPS), *Road Traffic Accident Statistical Annual Report*, TMBMPS, Beijing, China, 2018.

[4] American Association of State Highway and Transportation Officials (AASHTO), *Roadside Design Guide*, American Association of State Highway and Transportation Officials (AASHTO), Washington, DC, USA, 4th edition, 2011.

[5] Ministry of Transport of the People's Republic of China (MTPRC), *Specifications for Highway Safety Audit (JTG B05-2015)*, Ministry of Transport of the People's Republic of China (MTPRC), Beijing, China, 2015.

[6] Michigan Department of Transportation (MDOT), *Road Design Manual*, Michigan Department of Transportation (MDOT), Cadillac, MI, USA, 2011.

[7] C. R. Sax, T. H. Maze, R. R. Souleyrette, N. Hawkins, and A. L. Carriquiry, "Optimum urban clear zone distance," *Transportation Research Record*, vol. 2195, no. 1, pp. 27–35, 2010.

[8] X. Fan and Y. Xing, "Study on the width design of highway roadside clear zone considering tolerance concept," *Journal of Transportation Engineering and Information*, vol. 11, no. 4, pp. 49–54, 2013.

[9] K. W. Ogden, *Safer Roads: A Guide to Road Safety Engineering*, Ashgate Publishing Company, Burlington, Canada, 1996.

[10] C. Jurewicz and V. Pyta, "Effect of clear zone widths on run-off-road crash outcomes," in *Proceedings of the 2010 Australasian Road Safety Research, Policing and Education Conference*, pp. 1–12, Canberra, Australia, 2010.

[11] C. D. Fitzpatrick, S. Samuel, and M. A. Knodler Jr., "Evaluating the effect of vegetation and clear zone width on driver behavior using a driving simulator," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 42, pp. 80–89, 2016.

[12] C. D. Fitzpatrick, C. P. Harrington, M. A. Knodler Jr., and M. R. Romoser, "The influence of clear zone size and roadside vegetation on driver behavior," *Journal of Safety Research*, vol. 49, pp. 97–104, 2014.

[13] American Association of State Highway and Transportation Officials (AASHTO), *Roadside Design Guide*, American Association of State Highway and Transportation Officials (AASHTO), Washington, DC, USA, 1998.

[14] F. D. B. de Albuquerque, D. L. Sicking, R. K. Faller, and K. A. Lechtenberg, "Evaluating the cost-effectiveness of roadside culvert treatments," *Journal of Transportation Engineering*, vol. 137, no. 12, pp. 918–925, 2011.

[15] E. Ayati and M. Shahidian, "Safety and cost-effectiveness of clear zones in Iran," *Civil Engineering Innovation*, vol. 2, no. 1, pp. 37–46, 2007.

[16] E. Ayati and M. Shahidian, "Decision aid for allocation of transportation funds to roadside safety enhancement," *Iranian Journal of Science & Technology Transaction B Engineering*, vol. 31, no. 2, pp. 143–154, 2007.

[17] 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.

[18] G. Cheng, R. Cheng, Y. Pei, and L. Xu, "Probability of roadside accidents for curved sections on highways," *Mathematical Problems in Engineering*, vol. 2020, Article ID 9656434, 18 pages, 2020.

[19] G. Cheng, R. Cheng, L. Xu, and W. Zhang, "Risk assessment of roadside accidents based on occupant injuries analysis,"

Journal of Jilin University (Engineering and Technology Edition), vol. 51, 2020.

[20] Ministry of Transport of the People's Republic of China (MTPRC), *Technical Standard of Highway Engineering (JTG B01-2014)*, MTPRC, Beijing, China, 2014.

[21] L. Li, X. Zhu, and Z. Ma, "Driver brake reaction time under real traffic risk scenarios," *Automotive Engineering*, vol. 36, no. 10, pp. 1225–1229, 2014.

[22] Standardization Administration of China, *Identification for the Speed of Vehicle Involved in Road Traffic Accident (GB/T 33195—2016)*, Standardization Administration of China, Beijing, China, 2016.

[23] T. F. Zou, Q. Liu, A. M. Zha, C. Simms, and T. Chen, "New observations from real-world vehicle-pedestrian collisions in reducing ground related injury by controlling vehicle braking," *International Journal of Crashworthiness*, vol. 2020, Article ID 1827848, 2020.

[24] H. Wu, Y. Han, D. Pan et al., "The head AIS 4+ injury thresholds for the elderly vulnerable road user based on detailed accident reconstructions," *Frontiers in Bioengineering and Biotechnology*, vol. 9, no. 6, Article ID 682015, 2021.

[25] Ministry of Transport of the People's Republic of China (MTPRC), *Design Specification for Highway Alignment (JTG D20-2017)*, MTPRC, Beijing, China, 2017.

[26] Ministry of Land and Resources of the People's Republic of China (MLRPRC), *China's Highway Engineering Project Construction Land Index (Construction Standard [2018] 124)*, MLRPRC, Beijing, China, 2018.

[27] National Development and Reform Commission (NDRC), *Compilation Outline and Explanation of Social Stability Risk Assessment Reports of Major Fixed Asset Investment Projects (Trial) (Issue and Reform Investment No. [2013] 428)*, NDRC, Beijing, China, 2013.

[28] National Development and Reform Commission (NDRC), *Economic Evaluation Methods and Parameters of Construction Projects*, NDRC, Beijing, China, 3rd edition, 2006.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 2983605, 10 pages
https://doi.org/10.1155/2021/2983605

Research Article

Malicious Mining Behavior Detection System of Encrypted Digital Currency Based on Machine Learning

Mu Bie[1] **and Haoyu Ma**[2]

[1]*Information Technology Center, Chongqing Jianzhu College, Chongqing 400072, China*
[2]*School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China*

Correspondence should be addressed to Haoyu Ma; mlsxeby@163.com

Received 26 July 2021; Revised 16 August 2021; Accepted 16 September 2021; Published 18 November 2021

Academic Editor: Yong Chen

With the gradual increase of malicious mining, a large amount of computing resources are wasted, and precious power resources are consumed maliciously. Many detection methods to detect malicious mining behavior have been proposed by scholars, but most of which have pure defects and need to collect sensitive data (such as memory and register data) from the detected host. In order to solve these problems, a malicious mining detection system based on network timing signals is proposed. When capturing network traffic, the system does not need to know the contents of data packets but only collects network flow timing signals, which greatly protects the privacy of users. Besides, we use the campus network to carry out experiments, collect a large amount of network traffic data generated by mining behavior, and carry out feature extraction and data cleaning. We also collect traffic data of normal network behavior and combine them after labeling. Then, we use four machine learning algorithms for classification. The final results show that our detection system can effectively distinguish the normal network traffic and the network traffic generated by mining behavior.

1. Introduction

With the rapid popularity of the Internet and the continuous emergence of information technology, the attack means for emerging technologies are also constantly upgraded and evolving. With the help of various means, the media and carriers for the implementation of network security threats are unpredictable, and the network security situation is always not optimistic. In recent years, with the development of encryption currency trading market, and encryption of monetary value, malicious attack has become the most widely used means in mining and a kind of network security threats, affecting the enterprises, organizations, and individuals at the end of 2008, in the hearing on the network publishing a paper on the currency, called the currency: peer-to-peer electronic cash system [1]. In Satoshi Nakamoto's white paper, the authors describe a new kind of "currency" trading system that can work in a nontrust environment. That "currency" is cryptocurrency. Subsequent researchers then extended this concept and introduced the mining system into the mining of Bitcoin, which led to the rapid development of Bitcoin. In addition to Bitcoin, other types of cryptocurrencies have been released, bringing broader market space [2]. Mining itself is not a malicious activity. For Bitcoin mining, it is a computational activity, using hardware resources to perform mathematical computations in the Bitcoin network, which are paid in Bitcoins. As a result, countless people began to join the Bitcoin mining team [1]. In addition, with the more and more types of electronic currency, mining currency is also expanding, and mining forms and algorithms are also constantly evolving. At first, the mining behavior mainly uses CPU, GPU, and other resources, and the cost of these hardware resources is relatively low. Later, as the cryptocurrency market gradually expanded, people were no longer satisfied with using these hardware resources for mining but began to design various professional mining tools, such as programmable array for mining. Such professional mining tools greatly improved the mining speed [3].

Malicious mining or mining hijacking is a malicious practice that uses infected devices to mine cryptocurrency. The attacker exploits the computing power and bandwidth of the victim (the computer) to mine (in the vast majority of cases, this is done without the victim's knowledge or consent). Typically, malicious mining software responsible for such activities is designed to use sufficient system resources to perform mining operations without being noticed or detected for as long as possible. Since cryptocurrency mining requires a lot of computing resources, hackers try to break into multiple devices on the same LAN, so they can make more money by increasing computing power. Malicious cryptocurrency mining usually takes place in two ways: first, mining with browser-based cryptography, typically by embedding a script containing the mining code into a website. Take Coinhive scripting as an example, a cryptocurrency mining service that relies on a small piece of code embedded in a website that uses some or all of the computing power of a browser visiting a particular website to mine Monero cryptocurrency [4]. Due to the special ring signature scheme [5] used in the Monero coin protocol, it is difficult for law enforcement officers to determine the identity of the attacker through the collection address. Many attackers use Coinhive's services to turn multiple compromised sites and routers into mining machines for malicious mining operations.

The second way is to use binary-based malicious mining software for mining. When a user inadvertently clicks an e-mail containing mining software or a related malicious advertising link, the malicious mining software will be downloaded to the user's host computer, and malicious mining will be carried out without the victim's knowledge [6]. The earliest mining Trojan appeared in 2012. With the price of cryptocurrency skyrocketing from 2017, mining Trojan has become a major security threat in the Internet since 2018. Malicious mining attack not only brings performance loss to users, but also may greatly waste power resources and increase carbon emissions. Therefore, it is of great significance to conduct an in-depth study on mining software and find a more effective detection method for malicious cryptocurrency mining software by analyzing its characteristics.

In order to deal with the security threats brought by malicious cryptocurrency mining software, researchers have proposed a variety of protection schemes against cryptocurrency mining attacks; it is found that these existing solutions have certain limitations through comprehensive analysis [7].

In 2018, Hong [8] from Fudan University proposed a method to detect web mining behavior based on hash function. The author uses the Chrome Remote Interface (based on the Chrome debugging protocol, which supports debugging the Chrome browser) to carry out remote debugging of the web page and obtains the function call information of the web mining operation. Then, select the commonly used hash operation function for monitoring, according to the total time spent on hash operation to judge whether it is mining software.

In the same year, Konoth et al. from VU Amsterdam proposed Minesweeper [9], which is a defense strategy for web mining based on the characteristics of mining algorithm itself. Through inspection, the author finds that the current malicious web page mining operations mostly adopt WebAssembly (WASM) technology to improve the efficiency of web page encryption mining and use WASM code to realize the mining algorithm Cryptonight. Therefore, by analyzing the WASM code, the author matches the mining algorithm based on the detected encryption and hash operations, so as to determine whether a web page operation belongs to mining behavior.

Conti et al. [10] proposed the method of detecting mining software by using Hardware Performance Counter (HPC). The change of HPC value is used as the characteristic of the mining algorithm, and the curve conforming to the mining algorithm is fitted with the machine learning algorithm, so as to judge whether the program conforms to the rules of the mining algorithm and determine whether it is mining software [11]. However, this approach is highly controversial.

Based on the existing research work, this paper proposes a set of mining flow detection systems based on machine learning, aiming at the deficiency of the existing malicious mining behavior detection schemes. The method presented in this paper has three obvious advantages: first, it purely utilizes network stream timing signals for detection with strong privacy, without the need to install any local software and the need to read user data. Second, for network traffic transmitted with SSL encryption, there is no need to decrypt, as well as to detect; it is a warning, and classification can be realized through packet flow. Third, the scheme is highly adaptable and can be adapted to different mining algorithms and pool strategies at any time through new data training.

2. Network Traffic Classification Detection Algorithm

Various machine learning algorithms are used in this paper, including support vector machines (SVM), K-Nearest Neighbor (KNN), AdaBoost, and Convolutional Neural Networks (CNN). The performance differences of various algorithms were compared through experiments, and the detection efficiency was evaluated.

2.1. Linearly Separable Support Vector Machines. First, our task is to distinguish between normal network traffic and malicious mining traffic. And it is a typical binary classification task; the basic idea of support vector mechanism model is to create an optimal decision hyperplane and make the plane on both sides of the plane maximize the distance between the two classes of samples recently, and support vector machine was used first, in order to solve the problem of sample classification binary classification problems, so the algorithm has a good applicability for such tasks.

We assume that the sample network traffic data set to be trained is as follows:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}, \quad y_i \in \{-1, +1\}, \tag{1}$$

wherein x_j, $i = 1, \ldots, k$ represents the training sample data, and y_i represents the class symbol corresponding to each training sample.

The method of network traffic classification is to find a partition hyperplane in the sample space based on the collected training set M. The ultimate purpose of support vector machine is to construct a decision function, which can correctly classify every information. In two-dimensional space, intuitiveness is said to exist in the two-dimensional space of different points corresponding to the coordinates of separation, and the line can be a straight line and can also be a curve; for this task, characteristics exist in multidimensional space, thread classification obviously cannot meet the requirements, and this time, structure can only be used for classification of plane to classify a point in space, And such lines and planes are classification lines and classification hyperplanes, which can be expressed as

$$(\omega \cdot x) + b = 0. \tag{2}$$

In Equation (2), ω is the direction of the normal of the hyperplane, and b is the distance between the hyperplane and the origin. The location of the hyperplane is determined by these two unknowns. It is assumed that the distance between any point x in the network traffic training set and the hyperplane (ω, b) can be expressed as

$$r = \frac{|\omega \cdot x + b|}{\|\omega\|}. \tag{3}$$

As shown in Figure 1, in the SVM schematic diagram, the training sample points falling on the dotted line in the figure are the "support vectors" of the SVM algorithm, namely, the points with correct classification. We can calculate the sum of the distances from any two support vectors belonging to different classes to the hyperplane as

$$\gamma = \frac{2}{\|\omega\|}. \tag{4}$$

Through analysis, it can be found that there are many hyperplanes that can separate the two types of data, but there is only one optimal hyperplane that meets the two conditions at the same time, and only the points closest to the hyperplane are closely related to the hyperplane, and the other points will not directly affect the classification results. We give the basic lemma for support vector machines [12]:

$$\begin{aligned} &\min_{\omega,b} \frac{1}{2}\|\omega\|^2 \\ &\text{s.t. } y_i(\omega \cdot x_i + b) \geq 1, \\ &i = 1, 2, \ldots, k. \end{aligned} \tag{5}$$

Through observation, it can be found that Equation (5) is a convex quadratic programming problem with a unique minimum point. Formula (5) is transformed into a "duality problem" by Lagrange multiplier method. By introducing

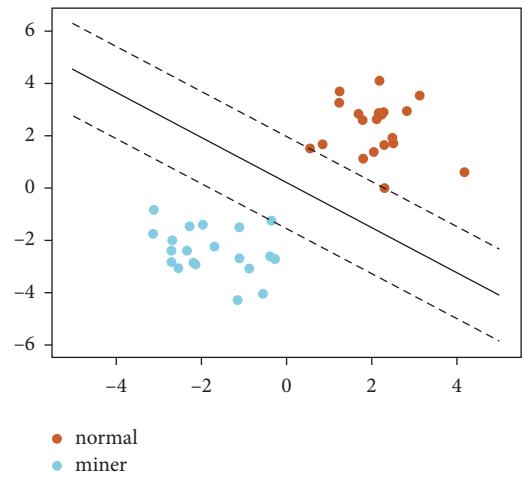


Figure 1: Use SVM to classify network traffic.

Lagrange multiplier, the formula can be processed more easily. Although the effect is not clear in linear cases, such change can be generalized to solve nonlinear problems and can be converted into kernel function for calculation. In the algorithm in this paper, we use the linear kernel function, which is defined as follows:

$$\kappa(x_1, x_2) = \langle x_1, x_2 \rangle. \tag{6}$$

2.2. Suitable AdaBoost Algorithm. Adaptive Boosting is the full name of AdaBoost algorithm, and Boosting, as a typical algorithm, was proposed in 1995 [13]. Boosting algorithm's main intention is to enhance the "weak classifier" into "strong classifier," and each "weak classifier" has its own functions and deficiencies. The advantages of each "weak classifier" will be brought into play, and the deficiencies of one "weak classifier" will be compensated by another or more "weak classifiers" to enhance the whole process into a "strong classifier." The AdaBoost algorithm enhances several "weak classifiers" into "strong classifiers" by an adaptive way. AdaBoost algorithm is reflected in practical application, mostly used in dichotomy. Recently, AdaBoost algorithm is mostly used in image recognition technology with remarkable effect and excellent performance in feature selection technology [14].

The algorithm ideas applicable to this task are as follows:

Assign a set of labeled network traffic data sets $(x_1, y_1), \ldots, (x_n, y_n)$ and $y_i \in \{+1, -1\}$, y as a lable, and x as a feature. Then, initialize weight:

$$D_i(i) = \frac{1}{n} (i = 1, \ldots, n). \tag{7}$$

Select a weak classifier h_t to reduce the weight of the error:

$$h_t = \text{argmin}_{h_{j \in H}} e_j = \sum_{i=1}^{n} D_t(i) I\left[y_i \neq h_j(x_i)\right]. \tag{8}$$

And calculate the proportion of $h_t(x)$ in the strong classifier:

$$a_t = \frac{1}{2}\ln\frac{1-e_i}{e_i}. \tag{9}$$

Until $e_t = \min_j e_j > 1/2$, set $T = t-1$ and then stop the loop. Update a new weight:

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \tag{10}$$

where Z_t is the normalized factor to ensure that D_{t+1} could be allocated.

Finally, we got a strong classifier:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} a_t h_t(x)\right). \tag{11}$$

2.3. Unsupervised Learning K-Nearest Neighbor Algorithm. K-Nearest neighbor algorithm is a famous statistical method of pattern recognition, which plays an important role in machine learning classification algorithms. In the KNN algorithm, the distance between the given test object and each object in the training set is firstly calculated; then, the K training objects nearest to the test object are selected as the nearest neighbors of the test object; and finally, the test objects are classified according to the main categories belonging to the K nearest neighbors. In general, KNN uses the "voting method" in the classification task; that is, it chooses the marker category with the most occurrence times in k instances as the prediction result. Euclidean distance is used to calculate the similarity when measuring the sample distance:

Assume that the category of network traffic is determined by n attribute; that is, the sample to be tested is an n-dimensional vector. Assume that the representation of sample i is $X^i = (x_1^i, x_2^i, \ldots, x_n^i)$. x_j^i represents the $j-$th attribute of sample i, so the Euclidean distance between any two samples can be obtained:

$$d\left(x^i, x^j\right) = \sqrt{\sum_{p=1}^{n}\left(x_p^i - x_p^j\right)^2}. \tag{12}$$

In a feature space as shown in Figure 2, a blue circle represents normal traffic, an orange rectangle represents abnormal traffic, and a red star represents an unknown anomaly detection point to be detected. According to the anomaly detection principle of KNN algorithm, the comparison of Euclidean distances will be used to judge whether a new unknown point is abnormal. Thus, the intrusion points A and B can be marked as abnormal traffic by the KNN algorithm, because the Euclidean distance from these two points to the abnormal class is closer than that to the normal class.

2.4. End-to-End Convolutional Neural Network Algorithm. In our previous research [15], deep learning has been applied to the detection of dark web traffic, and because network traffic can extract hundreds of relevant features, the feature extraction and analysis are also a question worth considering; for the first three algorithms, we use the principal causes analysis (PCA) [14] for dimension reduction; for deep learning, we can not undertake feature extraction or reduce it; this is called the end-to-end learning solutions. In Figure 3, we show the steps of the end-to-end learning algorithm.

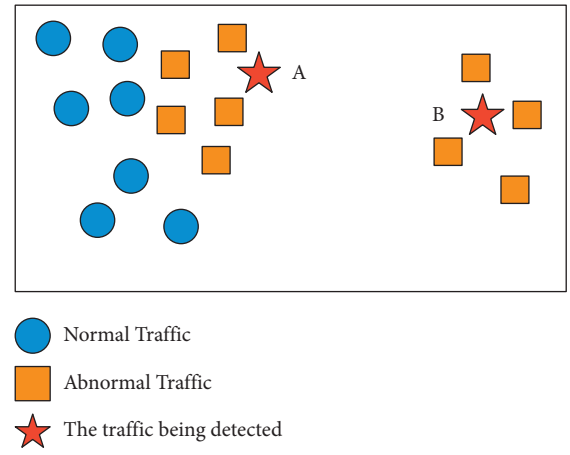


Figure 2: KNN.

Deep learning can automatically extract high-level features from original data and learn the internal rules and levels of samples. It has a high adaptability to massive high-dimensional data and well solves the problem of feature engineering in traditional machine learning.

In the overall structure of the model, the convolutional layer operation is the core part of CNN, responsible for automatically extracting a variety of abstract features from the original flow generated images, and different convolutional kernel learning features in the layer are different. The size parameters of the convolution kernel are specified manually, and the internal weight parameters need to be adjusted continuously in round training. The convolution kernel completes the convolution operation with each feature channel in the form of sliding window on the input matrix. Let us now define the $i-$th traffic data in the data set that is represented by k-dimensional vector, and a traffic with a characteristic length of n is represented as follows:

$$x_{i:n} = x_1 \oplus x_2, \ldots, \oplus x_n. \tag{13}$$

Convolutional operation: we define a convolution kernel $w \in \mathbb{R}^{hk}$, and the filter operates on a set of network traffic bytes of window width k and outputs a new feature. Refer to formula (14); a set of network traffic bytes generates characteristics by operating the following:

$$c_i = f\left(w \cdot x_{i:\,i+h-1} + b\right). \tag{14}$$

In formula (14), b is the bias term, and f is a nonlinear function.

In the dichotomy task, Sigmoid function is often used as the activation function [14]. In the actual training comparison, we find that the effect of this activation function is better than other activation functions.

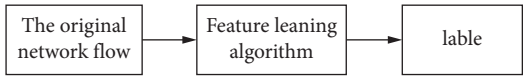


Figure 3: The end-to-end learning scheme.

$$f(x)_{\text{Sigmoid}} = \frac{1}{1+e^{-x}}, \tag{15}$$

$$c = [c_1, c_2, \ldots, c_{n-h+1}]. \tag{16}$$

Feature mapping: in formula (16), filters operate in every possible traffic byte window and generate a corresponding feature map, where $c \in \mathbb{R}^{n-h+1}$.

Pooling operation: we use maximum pooling for sampling. After the pooling layer completes the convolution operation between the image and the feature map in the convolution layer, the output data will enter the pooling layer for subsampling. By reducing the data dimension, overfitting is reduced, and the original feature information abstracted from the traffic data is extracted.

$$\hat{c} = \max\{c\}. \tag{17}$$

After the above operation, the resulting data will be entered into the full connection layer for classification. The operation of the full connection layer is described as follows: in formula (18), we define a linear function.

$$f(x)_{\text{probability}} = \sum_{i=1}^{d} w_i x_i + b. \tag{18}$$

We assume that there is a network traffic with d attributes described; we need to continuously update w and b for the purpose of training. And we choose Root Mean Square Prop (RMSProp) algorithm [13] as the optimization algorithm to update parameters:

$$s_{dw} = \beta v_{dw} + (1-\beta) dw^2, \tag{19}$$

$$s_{db} = \beta v_{db} + (1-\beta) db^2, \tag{20}$$

$$W = W - \alpha \frac{dW}{\sqrt{s_{dw}} + \varepsilon}, \tag{21}$$

$$b = b - \alpha \frac{db}{\sqrt{s_{db}} + \varepsilon}. \tag{22}$$

In formulas (19) and (20), s_{dw} and s_{db} are, respectively, the gradient momentum accumulated by the loss function in the first $t-1$ round of iteration. And β is an index of gradient accumulation, which we can define freely. In formulas (21) and (22), α is the learning rate in the training process, and to prevent the denominator from becoming zero, a small value ε is used for smoothing.

And we continue to use the sigmoid function as the activation function of the neuron. Then, in order to overcome the Sigmoid function brought by the characteristics of slow parameter update, we use the cross-entropy loss function to speed up the training process:

$$L(\hat{y}, y)_{\text{cross-entropy-loss}} = -y \log(\hat{y}) - (1-y)\log(1-\hat{y}). \tag{23}$$

Figure 4 shows the structure of the neural network. After the above processing, the trained neural network model can give the classification results of the input network traffic data.

3. Design of Malicious Mining Flow Detection System

In the early 1990s, the concept of abnormal traffic detection technology was first proposed in the literature [16]. Relevant scholars classified large-scale network traffic and divided it into normal traffic and abnormal traffic. Abnormal traffic detection technology can model network traffic characteristics to find abnormal patterns to predict the occurrence of abnormal, and at the same time, it can dig out new attack types for prevention. Abnormal traffic detection technology mainly includes five aspects, namely, network data traffic collection, traffic feature extraction, behavior modeling, abnormal detection, and result presentation and feedback. The specific technical process of abnormal traffic detection is shown in Figure 5.

In this paper, the detection model we proposed is roughly the same as that in Figure 5, and the specific implementation process will be introduced below.

3.1. Attacker Model. In this paper, it is assumed that the attacker will use the existing vulnerability exploitation program to remotely attack and exploit the hosts and services with vulnerabilities on the public network to achieve the purpose of planting malicious mining programs. Attackers may also be targeted at the target Server and host open Web services and applications for brute force cracking access, such as brute force cracking of Nginx Server or SQL Server, violent guess of SSH, VNC login credentials, and so on. There may also be unauthorized access vulnerabilities due to incorrectly configured application services and components deployed on the server. The hacker group carries out batch scanning of relevant service ports, and when the host and server with unauthorized access vulnerability are detected, further download and implant malicious mining programs by injecting execution scripts and commands [17]. Malicious mining attacks usually use remote code execution vulnerabilities or unauthorized vulnerabilities to execute commands and download and release subsequent malicious mining scripts or Trojans. Secondly, according to the our understanding and research of mining software, all mining software has network behavior, and this is an important basis for the study of mining software in this paper, but also an important support for the detection of mining software in this paper. Therefore, this paper assumes that mining software can access the network unconstrained and carry out network communication. Figure 6 describes in detail the communication flow between the miner and the mining pool.

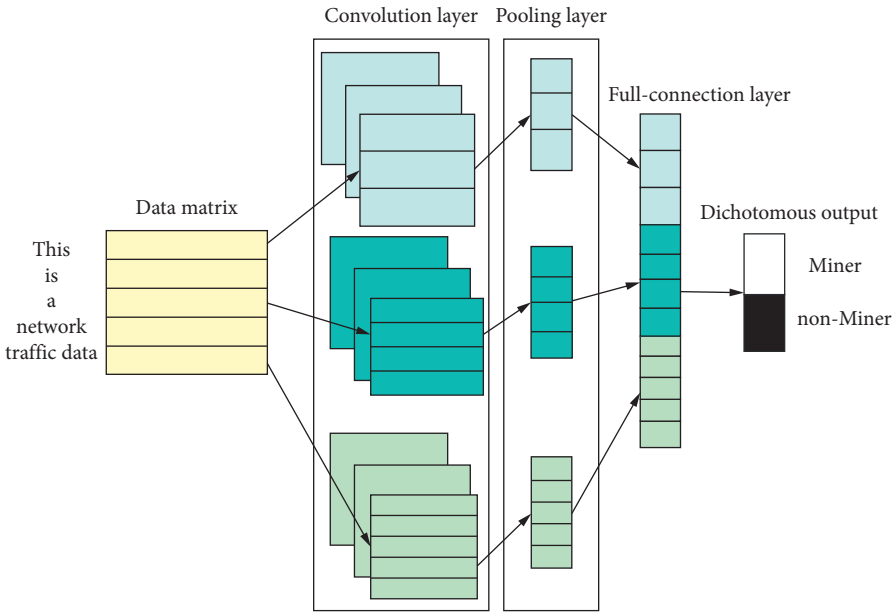


Figure 4: Neural network diagram.

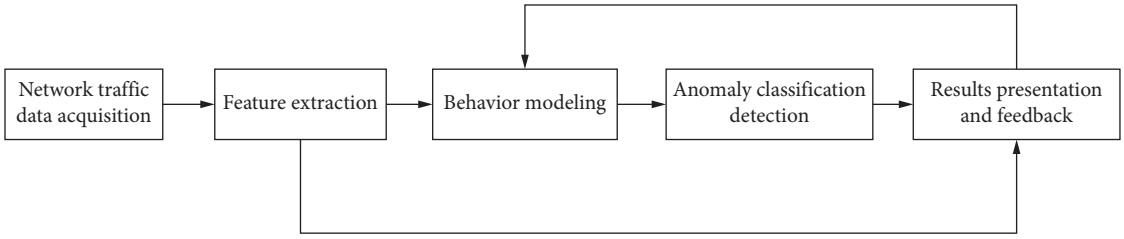


Figure 5: Flow for detecting abnormal traffic.

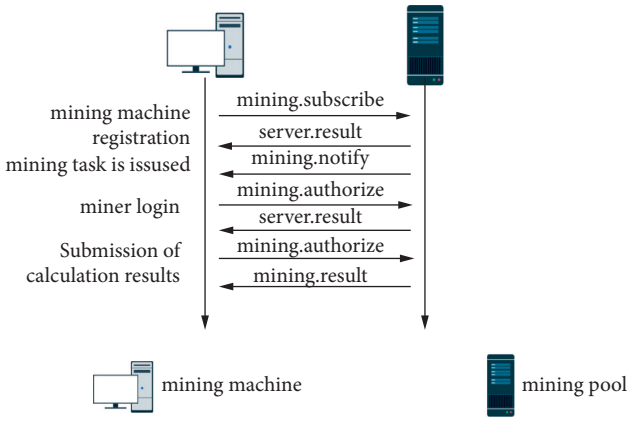


Figure 6: The communication flow between the controlled host and the mine pool.

We assume that all communication sessions in the figure above are transmitted using the Transport Layer Security protocol and cannot be decrypted. And the invaded host cannot be scanned for virus or Trojan because of objective reasons. We set that the attacker does not use a proxy server to forward network traffic, so that we can quickly lock the original IP address for blocking.

3.2. Detection and Defense Models. Anomaly traffic detection technology [18] is done by comparing the differences between normal mode and abnormal mode to judge whether it, apparently many kinds of attacks on attack, along with the rapid change of the large-scale traffic data, using the characteristics of these changes, can detect abnormal flow, so as to do a series of measures, including closing the relevant network port or shielding attack IP; this is undoubtedly a guarantee to the network environment. Anomaly traffic detection technology collected various hosts and servers of a large number of discrete traffic; the traffic data contains the network connection characteristics, building the model according to the data characteristics, analysis, and calculation of the secondary data, getting the multidimensional data flow characteristics, and achieving the normal behavior pattern, with the similar degree of normal behavior patterns to determine whether the behavior is unusual [19]. Finally, the feedback analysis of the test results is carried out. In this paper, we define abnormal traffic as the network traffic that mining software (Trojan) and mining pool communicate through.

Figure 7 shows the overall architecture diagram of the malicious mining flow detection system we proposed. In an environment where detection and warning are desired, network traffic can be collected from the interfaces of various

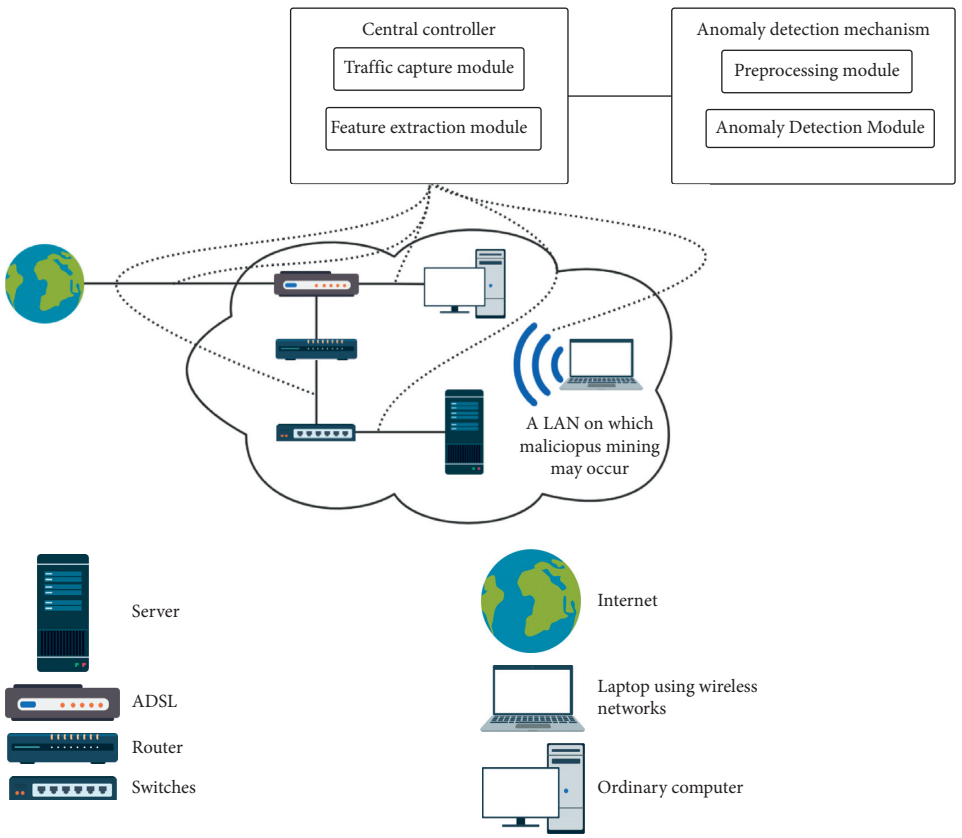


Figure 7: System architecture diagram.

devices for detection and analysis, or the underlying network can be analyzed directly through the device's operating system. The central controller will undertake this function. When the feature extraction module completes the feature extraction and feature selection of the data stream, it will send the selected optimal feature subset to the anomaly detection mechanism for processing and detection, so as to improve the efficiency and accuracy of the abnormal traffic detection mechanism and reduce the burden of the central controller.

Anomaly detection mechanism is a centralized data flow information analysis and anomaly detection processing mechanism. It obtains the stream feature vectors after feature selection through the central controller, preprocesses these vectors, and classifies them through the anomaly detection algorithm in Chapter 3 to determine whether there is an anomaly. Anomaly detection mechanism is divided into preprocessing module and anomaly detection module. The preprocessing module is mainly responsible for standardizing and normalizing the stream feature vectors sent by the controller. Suppose that there are n stream feature vectors, each of which contains t features, denoted as $X_{ij}\,(1 \le i \le n, 1 \le j \le t)$. The preprocessing of X_{ij} includes the following processes:

3.2.1. Standardized.

$$X_{ij}^{*} = \frac{X_{ij} - \text{Mean}_j}{\text{AvgDev}_j} \tag{24}$$

Among them,

$$\begin{aligned} \text{Mean}_j &= \frac{X_{1j} + X_{2j} + \cdots + X_{nj}}{n}, \\ \text{AvgDev}_j &= \frac{\left|X_{1j} - \text{Mean}_j\right| + \left|X_{2j} - \text{Mean}_j\right| + \cdots + \left|X_{nj} - \text{Mean}_j\right|}{n} \end{aligned} \tag{25}$$

3.3. Normalized. Normalization refers to the normalization of standardized data to the interval $[0, 1]$. Let X_{ij}' be the normalized value of X_{ij}^{*}:

$$X_{ij}' = \frac{X_{ij}^{*} - X_{\min}}{X_{\max} - X_{\min}}. \tag{26}$$

Among them,

$$X_{\min} = \min\left\{X_{ij}^{*}\right\},$$
$$X_{\max} = \max\left\{X_{ij}^{*}\right\}. \tag{27}$$

After the normalization and normalization of the preprocessing module of the anomaly detection module, the preprocessing stream feature vector will be handed over to the anomaly detection module for anomaly detection. The anomaly detection module uses the detection algorithm proposed in Chapter 3 to detect the traffic. Figure 8 shows the overall flow chart of the detection system we proposed.

4. Experiment

In order to simulate the real attack scenario, we used two hosts in the same campus network for experiments. One server was the target machine, and the other was Wireshark to capture the network traffic of the target machine. In order to make the collected data as fair and reliable as possible, we used the TLS protocol for the mine pool address, and all the features related to privacy in the data set were deleted (such as IP address).

4.1. Experiment Conditions. We use Wireshark [20] for network traffic capture, and then we use CICFlowMeter [21] for flow feature extraction. The specific experimental environment parameters are listed in Table 1.

4.2. Model Parameters. For the machine learning algorithm in Chapter 3, we have tried many parameters for training and evaluated under different parameter conditions. Finally, the parameters we selected are listed in Table 2–5.

The captured features of the train and test file are shown in Table 6. These data have gone through the preprocessing stage and removed the unobvious features.

4.3. Experimental Result. In Table 7, we use accuracy, precision, recall, and F1-score to compare several machine learning algorithms.

Accuracy is our most common evaluation index, and it is easy to understand, that is, the number of samples divided by all the samples. Generally speaking, the higher the accuracy, the better the classifier.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{28}$$

Precision is defined as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{29}$$

This represents the proportion of instances that are divided into positive instances that are actually positive instances.

Recall rate is a measure of coverage, which measures that multiple positive cases are divided into positive cases.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{30}$$

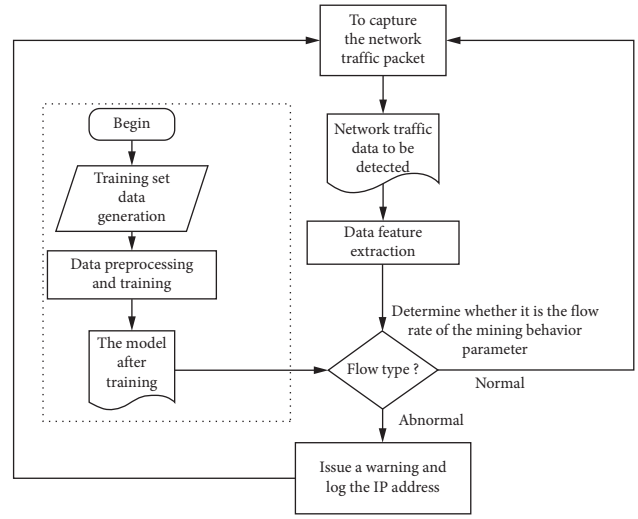


Figure 8: Flow chart of detection system.

Table 1: The experimental environment parameters.

Category	Parameters
CPU	Intel xeno 4210R
RAM	32 GB
GPU	Nvidia RTX3090
Operation system	Ubuntu 18.04 LTS
CUDA version	10.2
Machine learning platform	Pytorch 1.7.0 + scikit-learn 0.23

Table 2: Parameters used in support vector machine.

Parameter name	Parameter
Penalty coefficient of error term	0.3
Kernel	Linear
Probability	False
Max_iter	10000000
Random state	None

Table 3: Parameters used in AdaBoost.

Parameter name	Parameter
Base estimator	Decision tree
Number of base classifier cycles	60
Probability	False
Learning rate	1.2
Random state	None
Algorithm	SAMME.R

Sometimes, there are contradictions between precision and recall indicators, so we need to consider them comprehensively. The most common method is F-measure (also known as F-score):

$$F1 - \text{score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{31}$$

It can be seen from Table 7 that the effects of different machine learning methods on such tasks vary greatly, but the accuracy of these methods is high, but the precision

Table 4: Parameters used in KNN.

Parameter name	Parameter
N_neighbors	3
Weights	Uniform
Algorithm	Auto

Table 5: Training parameters used in CNN.

Parameter name	Parameter
Epoch	5
Iteration	30000
Batch size	128
Learning rate	0.01
Weight decay	1e-8
Momentum	0.9

Table 6: Description of traffic characteristics.

Name	Description
Protocol	Type of the protocol used
Flow duration	Length of connection
Flow Bytes (s)	Number of data bytes
Flow Packets (s)	Number of data packets
Flow IAT	Packets flow interarrival time
Fwd IAT	Forward interarrival time
Bwd IAT	Backward interarrival time
Active time	The amount of time a flow was active before becoming idle
Idle	The amount of time a flow was idle before becoming active

Table 7: Classification performance statistics.

ML algorithm	Accuracy	Precision	Recall	F1-Score
SVM	0.98	0.66	0.66	0.66
AdaBoost	0.99	0.78	0.81	0.79
KNN	0.91	0.64	0.63	0.64
CNN	1.00	0.96	0.86	0.91

performance is quite different. If we evaluate the classification effect according to F1-Score, deep learning is the most potential method to provide relatively correct prediction results.

5. Conclusion

The experimental results show that our detection system can effectively alert malicious mining behavior and protect user privacy. Our detection system can run on any node or host of the network and can quickly learn the characteristics of network traffic generated by different mining behaviors. For the malicious use of server computing resources for cryptocurrency mining, the existing detection methods all have some deficiencies; we used machine learning and data mining technology to analyze network traffic and greatly improve the efficiency of the detection system and effectively protect user privacy.

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

At the same time, the authors also need to state that this study was funded by the key Project of the Ministry of Education of China under grant ZD2020070201 and Science and Technology Research Program of Chongqing Municipal Education Commission under grant KJZD-K201904301.

References

[1] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," *Decentralized Business Review*, Article ID 21260, 2008.

[2] J. D. P. Rodriguez and J. Posegga, "Rapid: Resource and Api-Based Detectionagainst In-Browser miners," in *Proceedings of the 34th Annual Computer Security ApplicationsConference*, San Juan, PR, USA, December 2018.

[3] R. Tahir, M. Huzaifa, A. Das et al., "mining on someone else's dime: mitigating covert mining operations in clouds and enterprises," *Research in Attacks, Intrusions, and Defenses*, Springer International Publishing, in *Proceedings of the International Symposium on Research in Attacks,Intrusions, and Defenses (RAID)*, pp. 287–310, September 2017.

[4] S. Eskandari, A. Leoutsarakos, T. Mursch, and J. Clark, "A first look at browser-based cryptojacking," in *Proceedings of the 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 58–66, IEEE, London, UK, April 2018.

[5] S. Noether, "Ring signature confidential transactions for Monero," *IACR Cryptology*, vol. 1098, 2015.

[6] R. Konoth, "Malicious cryptocurrency miners: status and outlook," 2019, https://arxiv.org/abs/1901.10794.

[7] B. Ainapure, D. Shah, and A. Ananda Rao, "Performance analysis of virtual machine introspection tools in cloud environment," in *Proceedings of the International Conference on Informatics and Analytics*, Pondicherry, India, August 2016.

[8] G. Hong, "How you get shot in the back: a systematical study about cryptojacking in the real world," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, October 2018.

[9] R. K. Konoth, "Minesweeper: an in-depth look into drive-by cryptocurrency mining and its defense," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, Toronto, Canada, October 2018.

[10] A. Gangwal, "Detecting covert cryptomining using hpc," in *Proceedings of the International Conference on Cryptology and Network Security*, December 2020.

[11] R. Tahir, "Mining on someone else's dime: mitigating covert mining operations in clouds and enterprises," in *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses*, Springer, Cham, Switzerland, October 2017.

[12] J. A. K. Suykens, J. Vandewalle, and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[13] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and Computation*, vol. 121, no. 2, pp. 256–285, 1995.

[14] S. Wold, E. Kim, and G. Paul, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[15] H. Ma, "Dark web traffic detection method based on deep learning," in *Proceedings of the 2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*, May 2021.

[16] K. Hara, D. Saito, and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, July 2015.

[17] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and Its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[18] R. A. Maxion, "Anomaly detection for diagnosis," in *Proceedings of the Digest of Papers. Fault-Tolerant Computing: 20th International Symposium*, IEEE Computer Society, Tyne, UK, June 1990.

[19] Lashkari and A. Habibi, "Characterization of tor traffic using time based features," *ICISSp*, vol. 3, 2017.

[20] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[21] A. Orebaugh, G. Ramirez, and J. Beale, *Wireshark & Ethereal Network Protocol Analyzer Toolkit*, Elsevier, Amsterdam, Netherlands, 2006.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5503107, 11 pages
https://doi.org/10.1155/2021/5503107

Research Article

Weighted Reconstruction and Improved Eigenclass Combination Method for the Detection of Bearing Faults

Zhengyu Du,[1] Jie Ma,[1] Chao Ma,[2] Min Huang,[1] and Weiwei Sun[1]

[1]*School of School of Mechatronics Engineering, Beijing Information Science and Technology University, Beijing, China*
[2]*Key Laboratory of Modern Measurement and Control, Ministry of Education, Beijing, China*

Correspondence should be addressed to Jie Ma; mjbeijing@163.com

Received 4 August 2021; Accepted 26 October 2021; Published 10 November 2021

Academic Editor: Yong Chen

Aiming at the difficulty of extracting and classifying early bearing faults, a fault diagnosis method based on weighted average time-varying filtering empirical mode decomposition and improved eigenclass is proposed in this paper. Firstly, the bearing fault signal is decomposed into a series of intrinsic mode functions by the signal decomposition method, and the amplitude of the component is modulated by the weighted average method to enhance the fault impulse component. Then, the fractional Fourier transform is used to filter the reconstructed signal. Regarding classification issues, the eigenclass classifier is optimized by the IDE method that can be used for feature dimensionality reduction. Finally, the optimal features are selected and input into the IDE-EigenClass model. The experimental results show that the bearing fault diagnosis method proposed in this paper has higher accuracy and stability than the traditional PNN, SVM, BP, and other methods.

1. Introduction

In the process of bearing fault diagnosis, one difficult point is to extract the fault impulse component. Because the background noise is present in the process of equipment operation, the fault impulse component is likely to be drowned by the noise, resulting in the failure of diagnosis [1–3]. To solve this problem, Huang et al. proposed the empirical mode decomposition (EMD) for the first time, which can effectively reduce the noise in the signal. However, EMD has some problems such as end effect and mode aliasing [4]. Based on EMD, Huang et al. proposed the ensemble empirical mode decomposition (EEMD). By adding noise to the EMD decomposition process, the effect of mode aliasing has been significantly reduced [5]. In 2014, the variational mode decomposition (VMD) was proposed. VMD takes into full consideration the narrow-band property of the components, so the filtering frequency band is more centralized and the signal-to-noise ratio of the signal components is higher [6]. However, there remains the problem of mode aliasing. The time-varying filtering empirical mode decomposition (TVFEMD) is a new signal decomposition method proposed by Li et al. in 2017 [7]. The method provides an effective solution to the above problems by using B-spline approximation as a time-varying filter.

In addition, the noise can be removed by filtering the signal, as by the method of combining fast spectral kurtosis with VMD to successfully extract the features of early rolling bearing faults [8]. On the basis of the fast spectral kurtosis method, autocorrelation spectral kurtosis has been improved to effectively eliminate the interference of non-periodic noise and improve the recognition accuracy [9]. In recent years, the fractional Fourier transform (FRFT) has been applied intensively in fault diagnosis. The fractional Fourier transform has good time-frequency focusing property for LFM signals [10]. In the early stage of the fault, since the fault impulse frequency changes steadily, the fractional Fourier transform proves effective in extracting the fault impulse components. Jiao and Ma used FRFT to extract the features of the bearing signal, and the effect was significant [11]. In addition, the advantages of FRFT can be exploited to the full under variable conditions. Jia and Huang [12] successfully diagnosed the fault of variable-speed bearing using FRFT and LMS noise reduction methods.

Machine learning algorithms are used mainly for classification issue. For example, Zhang et al. used the EMD-SVM combination method to classify faults [13]. In addition, Jung and Koh used the DWT-KNN to classify faults and achieved certain results [14]. However, there remain many problems in classification. For example, the accuracy is not high enough, and the operation is too time-consuming. Standardized variable distance (SVD) is an improved algorithm based on MDE and proposed in 2020 to solve the problem of long operation time [15]. However, its classification accuracy is heavily dependent on features. In the neural network algorithm, Ma et al. used the CSBP method to improve the fault recognition rate by optimizing the BP neural network [16]. In addition, the convolutional neural network (CNN) [17] has also been used to classify the bearing faults. Eigenclass is a new machine learning classification method proposed by Uğur in 2020. In this procedure, a feature matrix is constructed before the proximity based on the eigenvalues and is used to classify the samples [18]. This method ensures not only a high accuracy but also a high speed and efficiency when the sample size is not very large.

The above studies mainly focus on the vibration signal. Nevertheless, for some devices which are not easily embedded with the acceleration sensor, it would be a good choice to use the acoustic pressure sensor to collect the acoustic signal for analysis and processing. Its advantage of noncontact measurement makes it possible to diagnose the fault of some devices in which it would otherwise be difficult or even infeasible to embed the acceleration sensor [19]. However, there are few studies on fault diagnosis methods based on acoustic signals probably because of the relatively low signal-to-noise ratio of the acoustic signal. In addition, the generated noise still contains certain impulse noise, which is not necessarily periodic but can have a great effect on the kurtosis of the signal. Therefore, the impulse noise may lead to the wrong selection of sensitive components.

To solve the above problems, a fault diagnosis method based on weighted average time-varying filtering empirical mode decomposition and improved eigenclass is proposed in this paper. Firstly, the acoustic array sensor is used to collect data. Then, the signal is preprocessed to reduce noise interference. Finally, the features are input into IDE-Eigenclass model for fault classification. Compared with the existing bearing fault diagnosis methods, the innovative point of this proposal is that the acoustic array sensor is used to collect the bearing fault signal. This method aims to realize the multichannel acquisition, which can not only improve the accuracy of bearing sound signal but also realize the localization of the sound source of fault bearing. In the process of actual equipment operation, there are usually more than one bearing, usually multiple or even multiple groups of bearings. In such a scenario, failure to accurately locate whichever of the bearings will make it inconvenient to maintain the equipment. The acoustic array technology has made it possible to accurately locate the pitch angle and azimuth of the bearing as well as the faulty bearing. This paper is organized as follows. Section 2 mainly introduces the principle of signal preprocessing. Section 3 introduces the principle of classification for optimization of IDE-Eigenclass. Section 4 deals with the diagnosis process. Section 5 is concerned on the experimental verification. Finally, a summary with relevant discussion is made in Section 6.

2. Signal Preprocessing

2.1. Time-Varying Filtering Empirical Mode Decomposition. Time-varying filtering empirical mode decomposition (TVFEMD) is an improved method based on EMD. Because the cut-off frequency in those methods based on filtering is constant over time, they are not suitable for nonstationary signals. This method can solve the problem of mode mixing while maintaining the time-varying characteristics of different modes. The detailed introduction on TVFEMD has been covered in [20] and will not be repeated here.

2.2. Signal Reconstruction. A new signal reconstruction method is proposed in this paper.

Firstly, the kurtosis of each component is calculated. The kurtosis can reflect the degree of a signal's transient state, thereby identifying the fault impulse characteristics. The expression is given as follows:

$$k = \frac{(N-1)\sum_{i=1}^{N}(x_i - \overline{x})^4}{\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)^2}. \tag{1}$$

Next, the coefficient of correlation between each component and the original signal is calculated, with the coefficient of correlation between two signals defined as

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}, \tag{2}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y, respectively, and N is the size of the dataset.

The product of kurtosis and coefficient of correlation is termed as correlation kurtosis:

$$s(i) = k(i) \times \rho(i). \tag{3}$$

The mean correlation kurtosis is calculated as

$$S(i) = \frac{s(i)}{\sum_{i=1}^{N} s(i)}, \tag{4}$$

where $s(i)$ represents the correlation-kurtosis value of each IMF component.

The signal is weighted and reconstructed by

$$\mathrm{IMF}_K^i = \sum_{i=1}^{N} S(i)\mathrm{IMF}^i. \tag{5}$$

The above method is the so-called weighted average time-varying filtering empirical mode decomposition.

This method is tantamount to modulating the amplitude at different frequency components of the signal. The kurtosis correlation coefficient index is proposed to modulate the amplitude of each component. By multiplying the index

factor by each component, the effective component can be enhanced, while the noise component can be reduced in amplitude, thereby pinpointing the fault component. At the same time, this method retains all IMF components to ensure the intactness of useful information, thereby improving the accuracy of diagnosis.

2.3. Fractional Fourier Transform. The fractional Fourier transform is based on the Fourier transform. Its physical significance is shown in Figure 1.

As shown in Figure 1, a signal is transformed from the time domain into the frequency domain through the Fourier transform, a process equivalent to rotating the signal by 90°. In other words, the physical significance of the fractional Fourier transform is equivalent to the whole transformation process of the signal at 0° to 90°. The Fourier transform is applied to every internal angle to find the optimal rotary angle through two-dimensional peak search. The principle of selection for the optimal rotary angle is described as follows. The rotary angle is optimal when the projection of the desired fault impulse component on the optimal plane of rotation does not interfere with the projection of the noise. Since the projection of noise at any rotary angle is invariable, the only consideration is the projection of the fault impulse component.

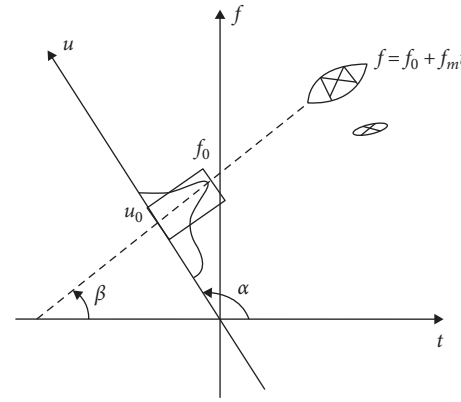


Figure 1: The physical significance of FRFT.

3. Feature Optimization and Fault Classification

3.1. IDE Feature Optimization. In recent years, machine learning has been applied more and more frequently in the research on classification problems. Two key factors that affect the effectiveness of classification are the quality of feature selection and the classification method used. The foremost step of classification is to select features, whereby to use the corresponding features to represent a section of data. The quality and the number of features can have a great impact on the effectiveness of classification. In addition, the number of features should be moderate. Too few features are not enough to make a distinction between the types of samples, while too many features may lead to the disaster of dimensionality, resulting in a sharp decline in classification accuracy. Overemphasizing the distinction between features may lead to overtraining, which means training the wrong features in the original data as correct ones at certain probability. Anyway, either an extreme number of features or overemphasis on the distinction between features may lead to a sharp decline in classification accuracy.

The improved distance evaluation (IDE) technique is a feature optimization method. The ratio of the average distance between the same type of samples to the average distance between different types of samples is calculated to judge whether the feature can make a distinction between the types [21]. All features are solved and normalized according to the above method. The numerical results are sorted, and the optimal features are selected by setting a certain threshold. For example, a feature set contains class B samples, and its expression is given as follows:

$$\left\{p_{n,b,j}, n = 1, 2, \ldots N_b;\ b = 1, 2, \ldots, B;\ j = 1, 2, \ldots, J\right\}, \tag{6}$$

where $p_{n,b,j}$ represents the j^{th} eigenvalue of the n^{th} sample in the class B sample, N_b represents the sample size of class B, J represents the eigenvalue, and B stands for the class. Therefore, the distinguishing degree of a feature to all classes can be calculated by the mathematical formula.

First, the average distance between the eigenvalues of the same samples is calculated as

$$d_{b,j} = \frac{1}{N_b \times (N_b - 1)} \sum_{l,n=1}^{N_b} \left|p_{n,b,j} - p_{l,b,j}\right|, \quad l, b = 1, 2, \ldots, N_b,\ l \neq m. \tag{7}$$

Then, we calculate the average value $d_j^{(w)}$:

$$d_j^{(w)} = \frac{1}{B} \sum_{b=1}^{B} d_{b,j}. \tag{8}$$

Define the variance factor of $v_j^{(w)}$ as

$$v_j^{(w)} = \frac{\max\left(d_{b,j}\right)}{\min\left(d_{b,j}\right)}. \tag{9}$$

Calculate the mean value of all samples under the same condition:

$$u_{b,j} = \frac{1}{N_b} \sum_{n=1}^{N_b} q_{n,c,j}. \tag{10}$$

Calculate the average distance between different samples:

$$d_j^{(b)} = \frac{1}{B \times (B - 1)} \sum_{b,e=1}^{B} \left|u_{e,j} - u_{b,j}\right|, \quad b, e = 1, 2, \ldots, B,\ b \neq e. \tag{11}$$

Define the variance factor $v_j^{(b)}$ as

$$v_j^{(b)} = \frac{\max\left(\left|u_{e,j} - u_{b,j}\right|\right)}{\min\left(\left|u_{e,j} - u_{b,j}\right|\right)}, \quad b, e = 1, 2, \ldots, B,\ b \neq e. \tag{12}$$

A compensation coefficient λ is introduced as

$$\lambda_j = \frac{1}{\left(v_j^{(w)}/\left(\max\left(v_j^{(w)}\right)\right)\right) + \left(v_j^{(b)}/\left(\max\left(v_j^{(b)}\right)\right)\right)}. \quad (13)$$

The ratio between the two average distances is calculated and the compensation coefficient is allocated:

$$\alpha_j = \lambda_j \frac{d_j^{(b)}}{d_j^{(w)}}. \quad (14)$$

Finally, $\overline{\alpha_j}$ is normalized as

$$\overline{\alpha_j} = \frac{\alpha_j}{\max\left(\alpha_j\right)}. \quad (15)$$

After normalization, the features are sorted in the size order, with the optimal features selected according to the threshold to form a feature matrix, which is then input into the machine learning model for training.

3.2. Eigenclass. According to the method, the feature matrix is first made up out of each set of features; next, the eigenvalues of the feature matrix in the training set and the test set are determined; finally, the distance between them is calculated.

Step 1: divide the dataset A into the matrix of the training set A_{train} and the test set A_{test}.

Step 2: if 0 is an element of the matrix, replace it with a very small positive number, such as 0.0001.

Step 3: for each class r, divide the training set into the training matrix A_{train}^r according to the number of classes.

Step 4: each line element of the test set is composed of the feature matrix, so is each line element of each class in the training set. Then, the eigenvalues of the test matrix and all the training matrices are obtained by

$$b_{rt}^i = q\left(\text{diag}\left(A_{t-\text{train}}^r\right), \text{diag}\left(A_{i-\text{test}}\right)\right). \quad (16)$$

Step 5: calculate the quasi-distance between the training matrix and the test matrix by the following definition:

$$q(A, B) = \sum \left| \begin{bmatrix} 1 \\ \cdot \\ \cdot \\ 1 \end{bmatrix} - \text{eig}(A, B) \right|. \quad (17)$$

Step 6: derive the quasi-distance between the test set and all the training sets of the same type, and then, construct a distance matrix and reorganize it in the ascending order.

Step 7: calculate the r-mean value of each line element. The test set of the line with the minimum mean value belongs to the class represented by this line.

The steps of the eigenclass algorithm are shown in Table 1.

4. Procedure

According to the method proposed in this paper, a flowchart can be drawn as shown in Figure 2. The diagnostic steps in this paper are shown in Table 2.

5. Experimental Verification

5.1. Feature Extraction. To further verify the effectiveness of the method proposed in this paper, the NU1004 cylindrical roller bearing is used to test the outer ring fault, with bearing parameters shown in Table 3. The rolling bearing fault testbed is shown in Figure 3. The test platform consists of a rolling bearing, a DC motor, a support bench, a loading bolt, and other components. The speed of the bearing is adjusted by the motor, and the loading bolt is used to load the bearing. The test speed has reached 780 rpm. ReSpeaker Mic Array v2.0 is used as an acoustic sensor. The sampling frequency is 44.1 kHz. The device is shown in Figure 4. Placed at a horizontal distance of 20 cm from the bearing, the sensor can implement 4-channel acquisition. Two arrays of such sensors are placed in the axial and radial directions of the bearing, respectively, to implement a total of eight channels of acquisition.

According to the main parameters of the bearing, the theoretical fault characteristic frequency is 64.48 Hz at a test speed of 780 rpm. The collected sound signal is decomposed by TVFEMD to obtain 15 intrinsic mode functions. The time-domain waveform of the original signal and some IMF components are shown in Figure 5. The purpose of selecting components 6, 7, 12, and 13 is that the kurtosis correlation coefficient index of the four components exceeds the threshold value of 0.7 after normalization, an indication that these components are most sensitive. The kurtosis and the coefficient of correlation are calculated as shown in Figure 6.

Figure 5 indicates that IMF component 1 has a relatively large kurtosis value. By the kurtosis selection principle, it should contain the richest fault information. However, the envelope spectrum shows that it has a peak value near 100 Hz, which is quite discrepant from the theoretical fault characteristic frequency of 64.48 Hz. The discrepancy may be ascribed to the impulse noise contained in this component rather than in the fault impulse component. By calculating the coefficient of correlation, it is found that the correlation with the original signal is very low, and therefore, it can be concluded that IMF 1 is not an effective component.

By multiplying the kurtosis correlation coefficient index by the amplitude of each IMF, the amplitude of each component is modulated to increase the ratio of sensitive component to noise component, thus fulfilling the purpose of feature enhancement.

Then, the reconstructed signal is filtered by undergoing the fractional Fourier transform, and the result is obtained at 90° to 180°, where the optimal rotary angle is found, as shown in Figure 7, and the best order is 1.01.

At the order of 1.01, narrow-band filtering is carried out, with the results shown in Figure 8.

Then, the inverse fractional Fourier transform is performed, with the results shown in Figure 9.

Table 1: Steps for eigenclass classification.

Eigenclass algorithm
Step 1: divide the data A into the matrix of the training set A_{train} and the test set A_{test}
Step 2: if $a_{ij-\text{train}} = 0$, $a_{ij-\text{train}} \longleftarrow 0.0001$ $a_{ij-\text{test}} = 0$, $a_{ij-\text{test}} \longleftarrow 0.0001$
Step 3: for each class r, extract the r-class training matrix A^r_{train} from A_{train}
Step 4: for each i, obtain the eigenvalues of the test matrix and all the training matrices: $b^i_{rt} = q(\text{diag}(A^r_{t-\text{train}}), \text{diag}(A_{i-\text{test}}))$
Step 5: calculate the quasi-distance between the training matrix and the test matrix by the following definition: $q(A, B) = \sum \left\| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \text{eig}(A, B) \right\|$
Step 6: construct a distance matrix C and reorganize it in ascending order
Step 7: calculate the r-mean of each row to form a discriminant matrix D as defined below: $D = \begin{bmatrix} d_1 \\ \vdots \\ d_r \end{bmatrix}$, $A_{\text{test}} \longrightarrow \min(d_1 \cdots d_r)$

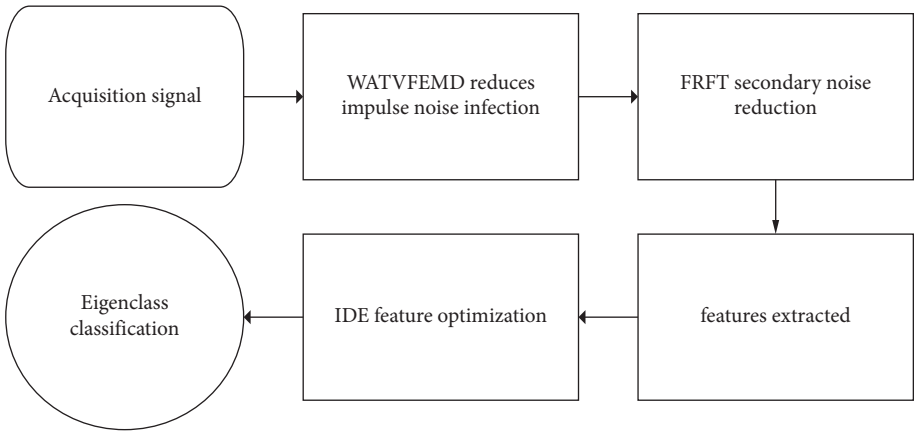


Figure 2: A flowchart of fault diagnosis.

Table 2: Fault diagnosis procedure.

Diagnostic procedure
Step 1: decompose the original signal by TVFEMD to extract the IMFs component
Step 2: obtain the correlation-kurtosis index of each modal function and weight and reconstruct the signal
Step 3: use an FRFT filter for secondary noise reduction to improve the signal-to-noise ratio
Step 4: extract 18 feature indexes including time domain and entropy, and then, use the IDE for feature selection to form a feature set
Step 5: divide the data into the training set and the test set and combine them into the eigenclass for classification prediction

Table 3: Main parameters of the NU1004 bearing.

Internal diameter (mm)	External diameter (mm)	Pitch diameter (mm)	Diameter of element (mm)	Number of balls
20	42	31	5.5	12

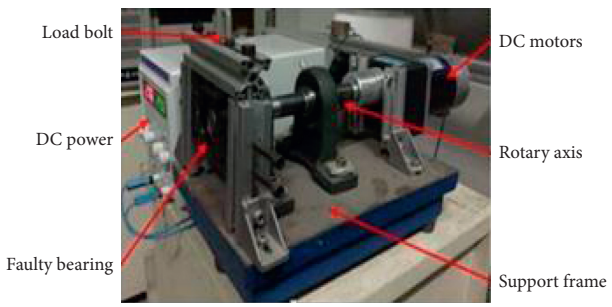


Figure 3: Bearing test platform.

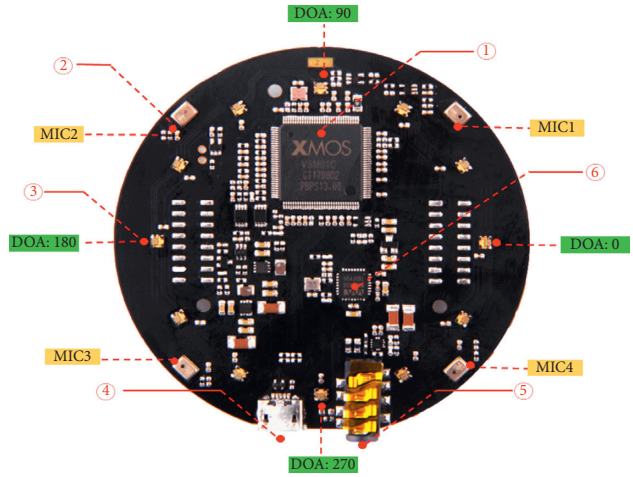


Figure 4: Schematic view of the sensor.

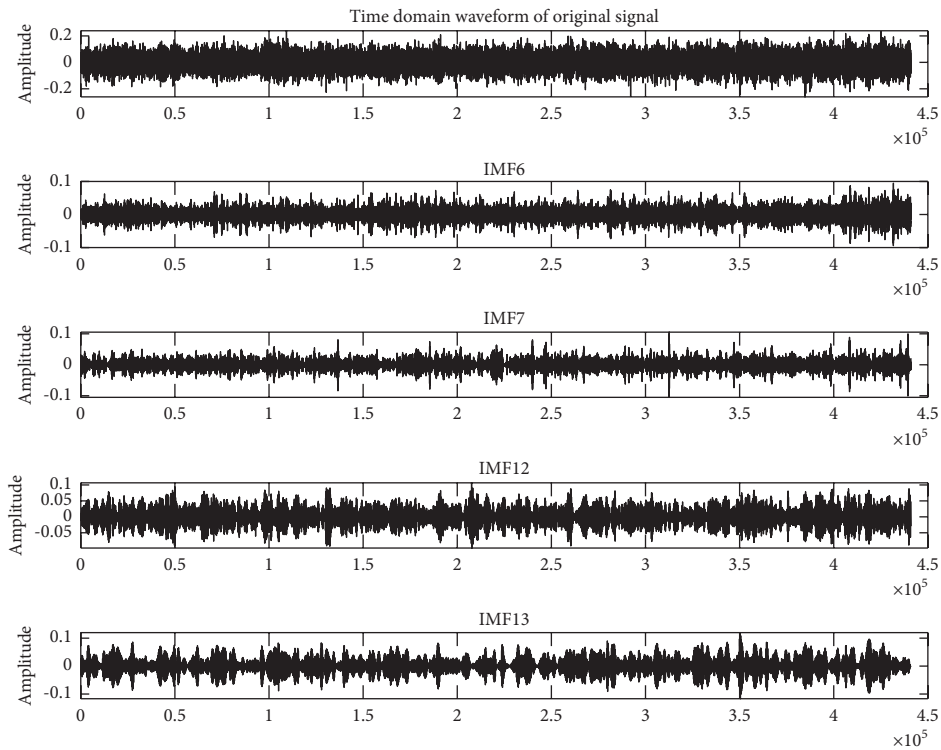


Figure 5: Time-domain waveforms of the original signal and some IMF components.

Figure 9 indicates that the fault characteristic frequency is 64.66 Hz. Given its bandwidth of 1 Hz, its harmonics and other noise components are filtered out. Therefore, the envelope spectrum contains only the fault characteristic frequency. It follows that the proposed method proves effective.

5.2. Fault Classification. The fault data of normal bearing, inner ring bearing, outer ring bearing, and rolling element bearing have been collected. More details of the dataset is presented in Table 4.

After all data are denoised, 15 time-domain indexes and 3 entropy indexes are selected to represent a segment of data. The time-domain characteristics are shown in Table 5. Three entropy indexes, namely, sample entropy, permutation entropy, and fuzzy entropy, have also been selected.

Then, the IDE is used for feature optimization, and the result is shown in Figure 10.

The threshold is set to 0.6. As shown in Figure 10, the threshold values of features 12, 13, 15, and 17 are greater than 0.6, an indication that these four features can make a good distinction between the experimental data.

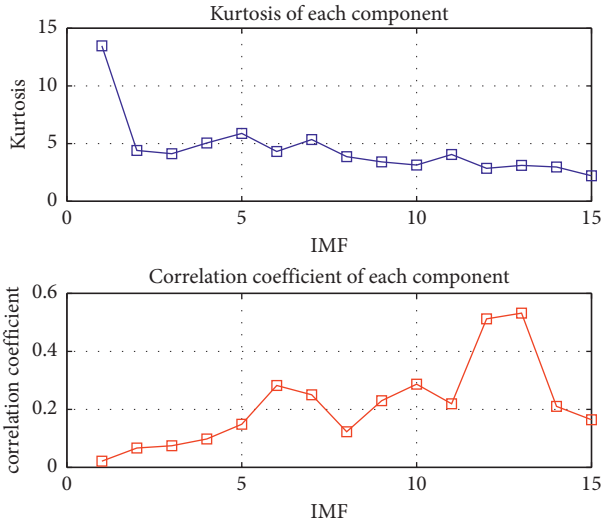


Figure 6: Kurtosis and correlation coefficient of each component.

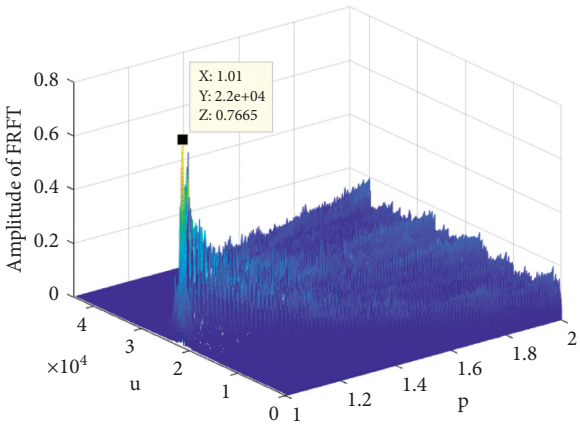


Figure 7: FRFT diagram.

Then, 400 datasets are combined according to the ratio of 1 : 4 to form the training set and the test set. Finally, the eigenclass is used to classify the data. The results are shown in Figure 11.

Figure 11 indicates that the eigenclass has a satisfactory effect of classification on the test set, with only 3 label classification errors, including only one classification error for normal bearing, inner circle, and outer ring, respectively. The classification accuracy of 96.25% demonstrates that the eigenclass has a good effect on bearing fault classification. The effectiveness of this method is preliminarily verified.

In addition, this paper makes the comparison and verification in two aspects. Firstly, to verify whether the eigenclass has sufficiently accurate classification effect, the processed data are input into PS0-SVM, BP, PNN, and SVD. 10 training sessions are conducted, respectively, with the results of one testing shown in Figure 12. The comparison of testing accuracy for 10 times is shown in Figure 13.

As can be seen from Figure 12, the eigenclass achieves the highest testing accuracy within the least operation time. Therefore, it can be verified that eigenclass has a better classification effect than the traditional classification methods.

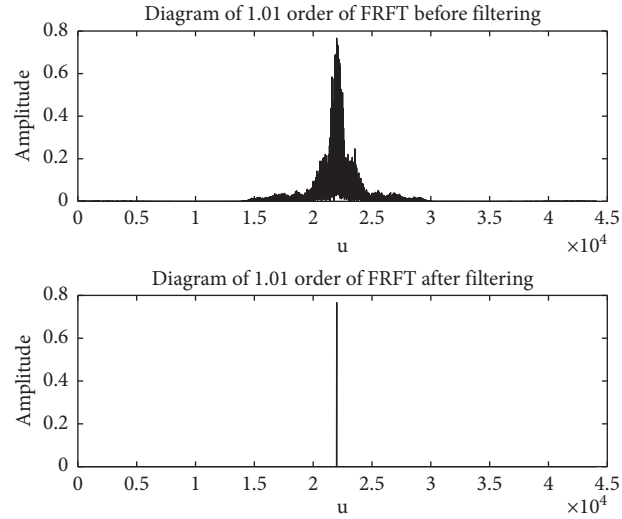


Figure 8: Fractional Fourier transform filter.

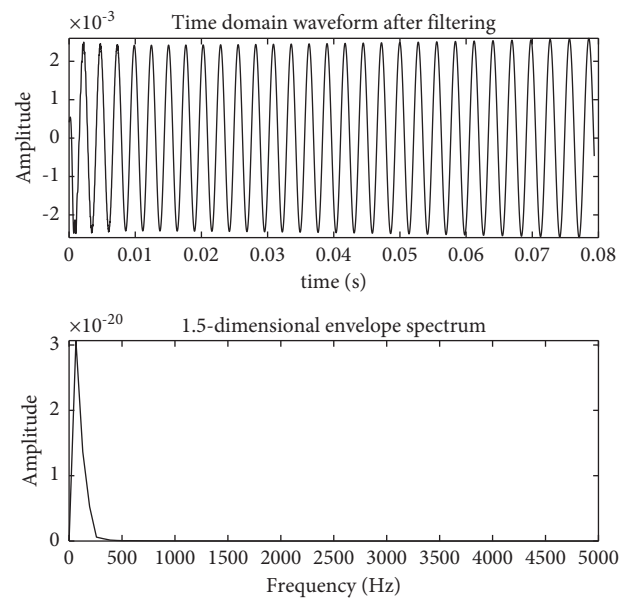


Figure 9: Partial time-domain waveform after filtering.

Table 4: Details of the bearing dataset.

Training samples	Testing samples	Operating condition	Label
80	20	Normal	1
80	20	Inner race	2
80	20	Out race	3
80	20	Ball	4

Secondly, to verify the effectiveness of the IDE method and the robustness of the eigenclass method, four features are randomly selected and input into the above five classification models, with the results shown in Table 6.

Table 4 indicates that the IDE method can improve the success rate of classification. In addition, no matter which four features are used, the eigenclass achieves the highest classification accuracy. Thus, the effectiveness of the proposed method is verified.

Table 5: Feature parameters.

Time-domain feature parameters
$T_1 = \max\|x_n\|$
$T_2 = \min\|x_n\|$
$T_3 = ((\sum_{n=1}^{N} x(n))/N)$
$T_4 = T_1 - T_2$
$T_5 = ((\sum_{n=1}^{N} \|x(n)\|)/N)$
$T_6 = (\sum (x_n - \overline{x})^2)/N$
$T_7 = \sqrt{((\sum (x_n - \overline{x})^2)/N)}$
$T_8 = ((1/n-1)\sum_{i=1}^{n} (x_i - \overline{x})^2)^{1/2}$
$T_9 = E[((X-\mu)/\sigma)^3]$
$T_{10} = \sqrt{(\sum_{i=1}^{N} X_i^2/N)}$
$T_{11} = (T_{10}/T_5)$
$T_{12} = (T_4/T_{10})$
$T_{13} = (T_{10}/T_5)$
$T_{14} = ((1/N)\sum_{i=1}^{N} \sqrt{\|x_i\|})^2$
$T_{15} = (T_4/T_{14})$

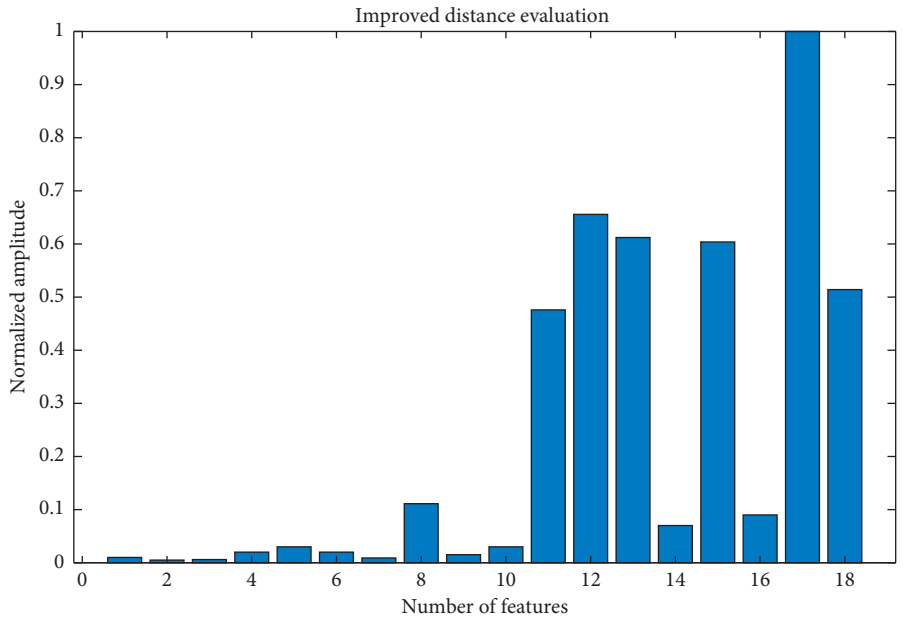


Figure 10: IDE feature optimization.

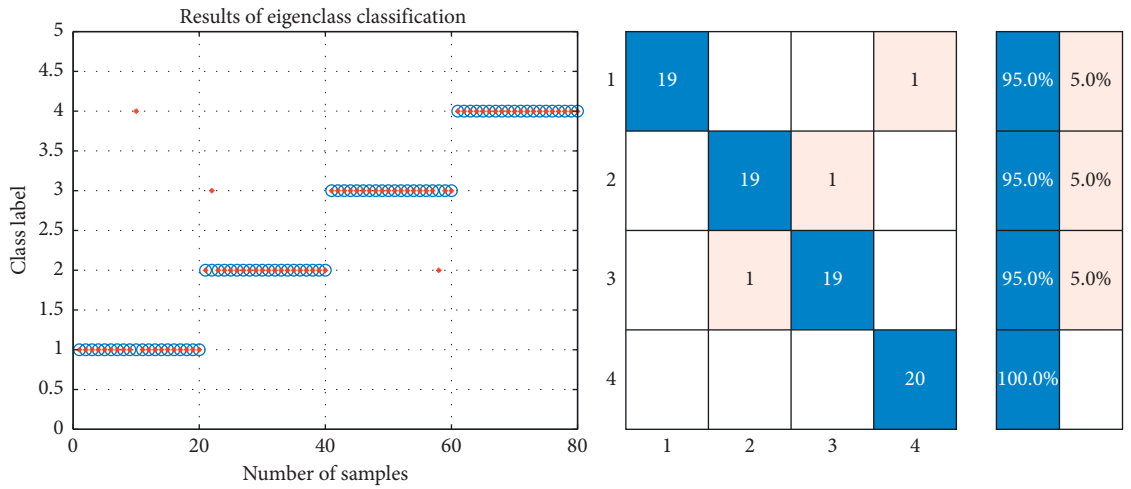


Figure 11: Classification results.

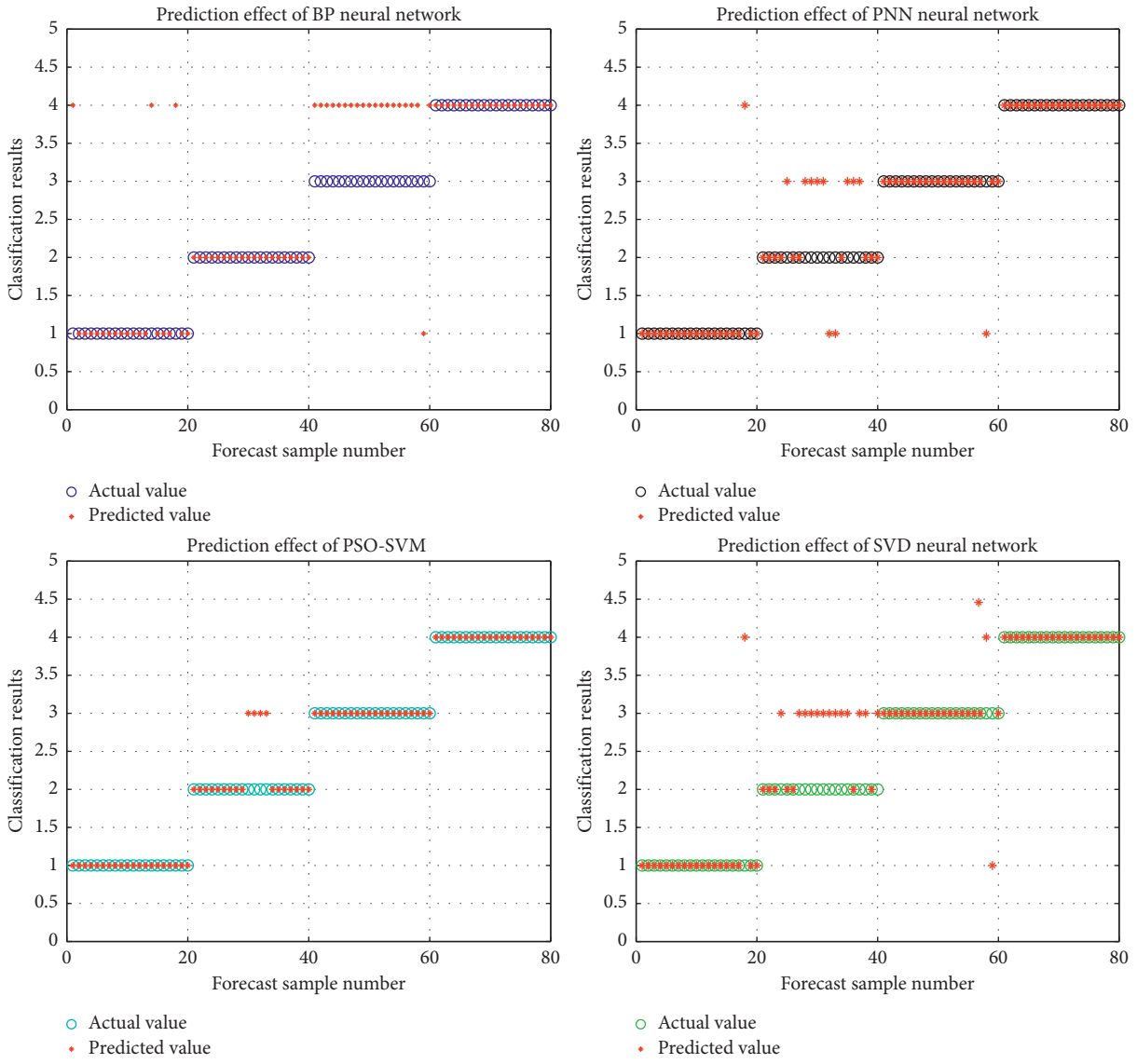


Figure 12: Comparison of classification results of different methods.

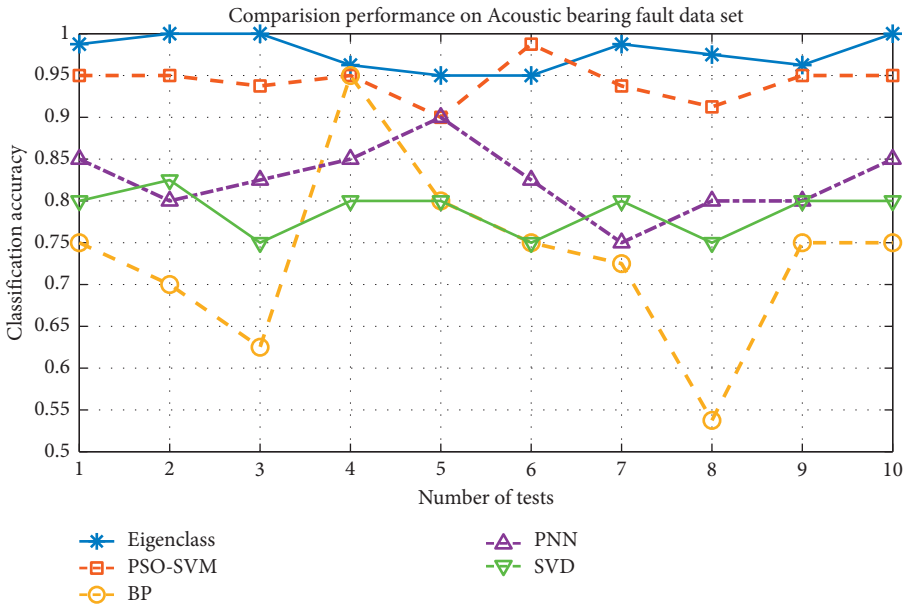


Figure 13: Classification accuracy of different methods for 10 times.

Table 6: Performance comparison for different features.

Accuracy	Superior features	Four features selected randomly		
		Min	Mean	Max
Eigenclass (%)	100	75.00	87.50	100
PSO-SVM (%)	97.50	62.50	78.75	95.00
BP (%)	95.00	50.00	72.50	95.00
PNN (%)	90.00	65.00	77.50	90.00
SVD (%)	82.50	60.00	71.25	82.50

6. Conclusion

In this paper, a bearing fault diagnosis method has been proposed, based on weighted average time-varying filter modal decomposition and improved eigenclass. Through experimental verification, the following conclusions can be drawn:

(1) The method of weighted reconstruction of decomposed signal proposed in this paper can effectively modulate the amplitude of signal, increase the ratio of the fault impulse component to the noise component, and improve the accuracy of classification.

(2) By optimizing the input of the classifier, the classification accuracy of the eigenclass has been improved with effectiveness. Compared with traditional methods, this novel method ensures both the highest accuracy and stability, as well as the verified effectiveness.

(3) In this paper, the eigenvalue attributes between indexes are used to construct the eigenvalue matrix. Concerning its originality, this method is innovative and open to further exploration and improvement in the follow-up research.

Data Availability

The data of this paper can only be obtained with the consent of the corresponding author.

Conflicts of Interest

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

Acknowledgments

This work was supported by the National Key Research and Development Program of China (no. 2019YFB1705402) and National Natural Science Foundation of The People's Republic of China (Grant 61973041).

References

[1] Y. Sun and J. Yu, "Fault detection of rolling bearing using sparse representation-based adjacent signal difference," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–16, 2021.

[2] Z. Liao, X. Song, B. Jia, and P. Chen, "Bearing fault feature enhancement and diagnosis based on statistical filtering and 1.5-dimensional symmetric difference analytic energy spectrum," *IEEE Sensors Journal*, vol. 21, no. 8, pp. 9959–9968, 2021.

[3] J. Ma and J. Xu, "Fault prediction algorithm for multiple mode process based on reconstruction technique," *Mathematical Problems in Engineering*, vol. 2015, Article ID 348729, 8 pages, 2015.

[4] N. E. Huang, Z. Shen, S. R. Long, and M. C. Wu, "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, pp. 903–995, 1998.

[5] N. E. Huang, M.-L. C. Wu, and S. R. Long, "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 459, no. 2037, pp. 2317–2345, 2003.

[6] R. Jegadeeshwaran, V. Sugumaran, and K. P. Soman, "Vibration based fault diagnosis of a hydraulic brake system using variational mode decomposition (VMD)," *Structural Durability & Health Monitoring*, vol. 10, pp. 81–97, 2014.

[7] H. Li, Z. Li, and W. Mo, "A time varying filter approach for empirical mode decomposition," *Journal of Signal Processing*, vol. 138, pp. 146–158, 2017.

[8] X. P. Ren, P. Li, and C. G. Wang, "Early fault diagnosis of rolling bearing based on VMD and fast spectral kurtosis," *Bearings*, vol. 12, pp. 39–43, 2017.

[9] A. Moshrefzadeh, A. Fasana, and J. Antoni, "The autogram:an effective approach for selecting the optimal demodulation band in rolling element bearing diagnosis," *Mechanical Systems and Signal Processing*, vol. 105, pp. 294–318, 2018.

[10] H. M. Ozaktas and D. Mendlovic, "Fractional Fourier transforms and their optical implementation II," *Journal of the Optical Society of America a-option Image Science and Vision*, vol. 10, no. 12, pp. 2522–2531, 1993.

[11] L. Jiao and J. Ma, "Feature extraction of rolling bearing faults based on VMD and FRFT," in *Proceedings of the 2020 IEEE 9th Data Driven Control and Learning Systems Conference, DDCLS*, pp. 167–172, Liuzhou, China, June 2020.

[12] J. J. Jia and Y. J. Huang, "Fault diagnosis of variable speed rolling bearing based on FRFT filter and LMS noise reduction," *Journal of Beijing University of Chemical Technology*, vol. 48, pp. 84–91, 2021.

[13] 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.

[14] U. Jung and B. H. Koh, "Wavelet energy-based visualization and classification of high-dimensional signal for bearing fault detection," *Knowledge and Information Systems*, vol. 44, no. 1, pp. 197–215, 2015.

[15] A. Elen and E. Avuçlu, "Standardized variable distances: a distance-based machine learning method," *Applied Soft Computing Journal*.vol. 98, Article ID 106855, 2021.

[16] C. Ma, S. H. Wang, and X. L Xu, "Intelligent fault diagnosis method of rolling bearing based on CSBP," *Journal of Electronic Measurement and Instrument*, vol. 33, pp. 58–63, 2019.

[17] W. Zhang, G. L. Peng, C. H. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, pp. 387–425, 2017.

[18] E. Uğur, "A precise and stable machine learning algorithm: eigenvalue classification (EigenClass)," *Neural Computing and Applications*, vol. 33, pp. 5381–5392, 2021.
[19] C. Ma, S. H. Wang, and X. L. Xu, "Fault diagnosis of acoustic array rolling bearing based on EEMD," *Journal of Electronic Measurement and Instrument*, vol. 31, pp. 1379–1384, 2017.
[20] K. D. Wang and K. K. Han, "Bearing fault diagnosis based on time varying filtering empirical mode decomposition," *Manufacturing Technology & Machine Tool*, vol. 12, pp. 42–46, 2018.
[21] Y. G. Lei, Z. J. He, and Y. Y. Zi, "A new approach to intelligent fault diagnosis of rotating machinery," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1593–1600, 2008.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1422006, 11 pages
https://doi.org/10.1155/2021/1422006

Research Article

Load Frequency Control for Power Systems with Actuator Faults within a Finite-Time Interval

Haifeng Qiu, Liguo Weng, Bin Yu, and Yanghui Zhang

State Grid Hangzhou Xiaoshan Power Supply Company, Hangzhou 311200, China

Correspondence should be addressed to Haifeng Qiu; 13857177475@sina.cn

Received 14 August 2021; Accepted 21 October 2021; Published 8 November 2021

Academic Editor: Meng Li

This paper is concerned with the issue of finite-time H_∞ load frequency control for power systems with actuator faults. Concerning various disturbances, the actuator fault is modeled by a homogeneous Markov chain. The aperiodic sampling data controller is designed to alleviate the conservatism of attained results. Based on a new piecewise Lyapunov functional, some novel sufficient criteria are established, and the resulting power system is stochastic finite-time bounded. Finally, a single-area power system is adjusted to verify the effectiveness of the attained results.

1. Introduction

Load frequency control (LFC), as an integral part of automatic generation control in power systems, has been adopted to regulate the frequency deviation and tie-line power exchanges [1–3]. Added by the LFC strategy, the high-quality electric energy can be maintained over a certain range [4]. In general, constant frequency deviation may lead to unreliable frequency devices, transmission lines overload, etc. Meanwhile, owing to the large size of the power grid, it raises the difficulty in frequency control. Therefore, it is a tough task to design suitable frequency control law. In practical applications, the loads are unexpected and unmeasurable, which indirectly regulate the system frequency. Accordingly, through the LFC strategy, the system performance can be guaranteed without affecting the generation capacity or frequency deviation. Up to now, the research on the LFC for power system gradually becomes a hot topic [5–7].

In networked control systems, various faults can be encountered due to the long-term utilization of components [8–10]. Note that the actuator faults are the source of instability and performance deterioration. To overcome the above shortage and improve the dependability, a great deal of attention has been shifted to actuator faults, and plenty of results have emerged [11, 12]. However, the actuator faults are assumed to be time-unchanged, which limits the potential applications. As stated in [13], the so-called failure probability is common in the reliability industry, where failure rates can be governed by the Markov switching chain [14–16]. Despite the significant achievement has attained, no suitable attention has been devoted to the power systems.

On the other hand, Lyapunov asymptotic stability is most common in the literature, where asymptotic behavior can be expected over the infinite-time domain. Nevertheless, in reality, the desirable transient performance is very important in many physical systems, which causes the inapplicability of the Lyapunov stability. Following this trend, finite-time stability (FTS) has been studied [17–19], which concerns the dynamic behavior within bound over a fixed time interval instead of an asymptotical case. As is well known that FTS is different from the Lyapunov case, it gives more solutions of transient performance control. Owing to the merits of the FTS, many valuable achievements have been made over the past years [20]. However, to our knowledge, most of the previous results are assumed that the data communication keeps continuous between sensors and controllers. In the fields of sampled-data control law, this assumption is not accurate. In general, with respect to the demand of actual systems, the sampler may encounter component aging, data losses, etc. [21–23]. These shortages may lead to unreliable periodic sampling. Fortunately, the aperiodic sampled-data control strategy is presented

[24, 25], which can efficiently deal with the aforementioned issues. However, the finite-time aperiodic sampled-data control for power systems remains unsettled, not mentioned to the LFC, which motivates us for this study.

Inspired by the above observations, we focus on the finite-time H_∞ load frequency control for power systems with actuator faults over the finite-time interval in this study. The main contributions can be summarized as follows: (1) different from the previous studies, to fully describe the randomly occurring actuator fault, the actuator fault is characterized by a homogeneous Markov chain. (2) To better characterize the actual demands of practical dynamics, a generalized framework of the actuator constraint is considered. (3) Apart from the traditional Lyapunov asymptotic stability, this study exploits the FTS for power systems and focuses on the finite-time control issue. By resorting to the piecewise Lyapunov theory, some novel results over the finite-time interval are reached. Finally, a numerical example is manifested to reveal the validity of the gained results.

The remainder of this study is listed as follows. Section 2 provides a description of the problem. Section 3 presents the main results, and the simulation validation is exhibited in Section 4. Section 5 concludes the study.

1.1. Notations. The notations of this paper are standard. $\|\cdot\|$ means the Euclidean norm. $\mathbb{R}$ indicates a set of n-dimensional matrix. E refers to the mathematical expectation. $(\lambda_{\max}(\mathrm{A})/\lambda_{\min}(\mathrm{A}))$ means the largest/smallest eigenvalue of matrix A. $\Pr\{\cdot\}$ means the occurrence probability. $\mathrm{diag}\{\cdot\}$ represents a block-diagonal matrix.

2. Problem Formulations

Block diagram of single-area LFC power model is exhibited in Figure 1 [6]. Accordingly, the dynamic equation of power model can be listed as follows:

$$\begin{cases} \dot{\overline{\delta}}(t) = \overline{A}\overline{\delta}(t) + \overline{B}u^F(t) + \overline{F}\omega(t), \\ \overline{y}(t) = \overline{C}\overline{\delta}(t), \end{cases} \tag{1}$$

where

$$\begin{aligned} \overline{\delta}(t) &= [\,\Delta f \;\; \Delta P_m \;\; \Delta P_v\,]^T, \\ \omega(t) &= \Delta P_d, \\ \overline{C} &= [\,\beta \;\; 0 \;\; 0\,], \\ \overline{A} &= \begin{bmatrix} -\dfrac{D}{M} & \dfrac{1}{M} & 0 \\ 0 & -\dfrac{1}{T_{ch}} & \dfrac{1}{T_{ch}} \\ -\dfrac{1}{RT_g} & 0 & -\dfrac{1}{T_g} \end{bmatrix}, \\ \overline{B} &= \begin{bmatrix} 0 & 0 & \dfrac{1}{T_g} \end{bmatrix}^T, \\ \overline{F} &= \begin{bmatrix} -\dfrac{1}{M} & 0 & 0 \end{bmatrix}^T, \end{aligned} \tag{2}$$

and the system parameters are expressed in Table 1.

In single-area, the area control error (ACE) is interpreted as $y(t) = \beta\Delta f$ due to the unaccessiblity of the tie-line power exchange. In reality, the actuator faults cannot be neglected for long-term utilization of components, which can be expressed as

$$u^F(t) = \alpha(r_t)u(t), \tag{3}$$

where $\alpha(r_t) = \mathrm{diag}\{\alpha_1(r_t), \alpha_2(r_t), \ldots, \alpha_f(r_t)\}$, and each element $\alpha_m(r_t) \in [0,1]\,(m = 1, 2, \ldots, f)$. More specifically, $\{r_t, t \geq 0\}$ is identified as a right-continuous Markov chain taking values over a set $\mathcal{S} = \{1, 2, \ldots, S\}$ with generator $\Pi = [\pi_{pq}]_{S\times S}$, and its transition probabilities are inferred as

$$\Pr\{r_{t+\Delta t} = q | r_t = p\} = \begin{cases} \pi_{pq}\Delta t + o(\Delta t), & \text{if } q \neq p, \\ 1 + \pi_{pq}\Delta t + o(\Delta t), & \text{if } q = p, \end{cases} \tag{4}$$

where $\Delta t > 0$ and $(\lim_{\Delta t \longrightarrow 0} o(\Delta t)/\Delta t) = 0$, for $q \neq p$ and $\pi_{pp} = -\sum_{q\neq p}\pi_{pq}$ for each $p \in \mathcal{S}$.

Taking the ACE as the desired controller input of LFC, the output of the proportional-integral (PI) controller is asserted as

$$u(t) = -K_P\overline{y}(t) - K_I\int \overline{y}(t), \tag{5}$$

where K_P and K_I signify the proportional and integral gains of the area, respectively.

Let $\delta(t) = [\overline{\delta}^T(t) \int_0^t \overline{y}^T(s)\mathrm{d}s]^T$, $y(t) = [\overline{y}^T(t) \int_0^t \overline{y}^T(s)\mathrm{d}s]^T$, the power model (1) is reformulated as

$$\begin{cases} \dot{\delta}(t) = A\delta(t) + B\alpha(r_t)u(t) + F\omega(t), \\ y(t) = C\delta(t), \end{cases} \tag{6}$$

where

$$A = \begin{bmatrix} -\dfrac{D}{M} & \dfrac{1}{M} & 0 & 0 \\ 0 & -\dfrac{1}{T_{ch}} & \dfrac{1}{T_{ch}} & 0 \\ -\dfrac{1}{RT_g} & 0 & -\dfrac{1}{T_g} & 0 \\ \beta & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \dfrac{1}{T_g} \\ 0 \end{bmatrix},$$

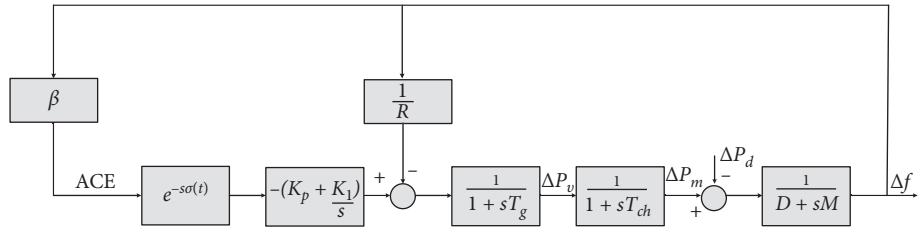


Figure 1: The structure of single-area power system.

$$F = \begin{bmatrix} -\frac{1}{M} \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

The purpose of this study is to solve the output feedback control for power system (6) with data sampling. Therefore, the sampling sequence attained at a set of time instants. Added by the data sampling technique, only the measured signal $y(t_k)$ can be released to the controller. Specifically, the sampling instants are represented as

$$0 = t_0 < t_1 < \cdots < t_k < t_{k+1} < \cdots < \lim_{k \longrightarrow \infty} t_k = \infty. \tag{8}$$

In light of periodic sampling instants, in this study, we consider the aperiodic sampling case. Following this trend, the sampling interval $[t_k, t_{k+1})$ is time-varying with the upper sampling period. Thus, one defines $0 < t_{k+1} - t_k \triangleq \tau_k \leq \tau$. Based on the input delay technique, we have that $t_k = t - (t - t_k) \triangleq t - \tau(t)$ with $\dot{\tau}(t) = 1$ for $t \neq t_k$. Summarizing the above discussion, we have $0 \leq \tau(t) < \tau_k \leq \tau$.

Letting $K = \begin{bmatrix} K_p & K_I \end{bmatrix}$, the PI-based sampled data LFC can be designed as

$$u(t) = Ky(t_k), \quad t \in [t_k, t_{k+1}). \tag{9}$$

Substituting (3) and (9) into (6), the closed loop power system can be governed by

$$\dot{\delta}(t) = A\delta(t) + B\alpha(r_t)KC\delta(t_k) + F\omega(t), \quad t \in [t_k, t_{k+1}). \tag{10}$$

Before further derivation, some important contents are stated as follows.

Assumption 1 (see [18]). The external disturbance $\omega(t)$ belongs to $L_2[t_0, T)$, and it is assumed that there exists a parameter $\overline{\omega} > 0$ such that $\int_{t_0}^{T} \omega^T(t)\omega(t)\mathrm{d}t \leq \overline{\omega}$.

Table 1: Physical meaning.

Parameters	Physical meaning
ΔP_v	Governer valve position
Δf	Deviations of frequency
ΔP_m	Mechanical output of generator
β	Frequency bias factor
ΔP_d	Load disturbance
M	Moments of inertia of the generator
R	Speed drop
T_d	Time constant of the governor
D	Damping coefficient of generator
T_{ch}	Time constant of turbine

Assumption 2. From the viewpoint of the physical limitation of actuators in power systems, the control torque is assumed to meet

$$|u_s(t)| \leq u_{\max}, \quad s = 1, 2, \ldots, n_u. \tag{11}$$

Definition 1 (see [26]). Given parameters $c_1 > 0$, $c_2 > 0$, time interval $T > 0$, and matrix $R > 0$, the closed loop power system (10) is called stochastic finite-time stability (SFTS) with respect to (c_1, c_2, T, R) and $\omega(t) = 0$, if inequality $\mathrm{E}\{\delta^T(t_0)R\delta(t_0)\} \leq c_1 \Rightarrow \mathrm{E}\{\delta^T(t)R\delta(t)\} \leq c_2$ holds for $\forall t \in [t_0, T]$.

Definition 2 (see [26]). Given parameters $c_1 > 0$, $c_2 > 0$, $\overline{\omega} > 0$, time interval $T > 0$, and matrix $R > 0$, the closed loop power system (10) is called stochastic finite-time boundedness (SFTB) with respect to $(c_1, c_2, T, R, \overline{\omega})$, if inequality $\mathrm{E}\{\delta^T(t_0)R\delta(t_0)\} \leq c_1 \Rightarrow \mathrm{E}\{\delta^T(t)R\delta(t)\} \leq c_2$ holds for $\forall t \in [t_0, T]$.

The object of this study is to design sampled-data-based controller (9) such that

(1) The closed-loop power system (10) is SFTS with $\omega(t) = 0$.

(2) When disturbances $\omega(t) \neq 0$, the power system (10) is called SFTB with H_∞ performance index γ such that

$$\mathrm{E}\{\|y(t)\|_2^2\} < \gamma^2 \|\omega(t)\|_2^2. \tag{12}$$

Lemma 1 (see [27]). *For any vectors $\hbar_1(t)$ and $\hbar_2(t)$, scalar $\iota \in (0, 1)$, a matrix* A*, and symmetric matrices* B *and* C*, the following inequality holds:*

$$-\frac{1}{\iota}\hbar_1^T(t)\mathrm{B}\hbar_1(t)-\frac{1}{1-\iota}\hbar_2^T(t)\mathrm{C}\hbar_2(t)\le-\begin{bmatrix}\hbar_1(t)\\ \hbar_2(t)\end{bmatrix}^T\begin{bmatrix}\mathrm{B} & \mathrm{A}\\ * & \mathrm{C}\end{bmatrix}\begin{bmatrix}\hbar_1(t)\\ \hbar_2(t)\end{bmatrix}. \tag{13}$$

3. Main Results

Theorem 1. *For given parameters $\rho>0$, $c_1>0$, $c_2>0$, $T>0$, $\overline{\omega}>0$, $u_{\max}$, and matrix R, the closed-loop power system (10) is called SFTB with respect to $(c_1,c_2,\overline{\omega},T,R)$, if there exists matrix $P_p>0$, $N_1>0$, $N_3>0$, $Q_1>0$, and $Q_2>0$, such that $\forall p\in\mathcal{S}$*

$$\mathcal{Q}_f\ge 0,\quad f=1,2, \tag{14}$$

$$\Theta^1<0, \tag{15}$$

$$\begin{bmatrix}\Theta^2 & \tau_p^T\mathcal{H}_p^T\\ * & -\tau\mathcal{Q}_1\end{bmatrix}<0, \tag{16}$$

$$\psi c_1+\gamma^2\overline{\omega}<e^{-\rho T}\min_{p\in\mathcal{S}}\left\{\lambda_{\min}\left(\widetilde{P}_p\right)\right\}c_2, \tag{17}$$

$$\begin{bmatrix}-u_{s\max}^2 & C^TK_s^T\\ * & -\dfrac{R}{c_2}\end{bmatrix}<0,\quad s=1,2,\ldots,n_u. \tag{18}$$

where

$$\begin{aligned}
\Theta^i&=\begin{bmatrix}\Theta_{11}^i & \Theta_{12}^i & \Theta_{13} & \Theta_{14}^i & \Theta_{15}\\ * & \Theta_{22}^i & H_8^\top & \Theta_{24}^i & H_{10}\\ * & * & \Theta_{33} & 0 & 0\\ * & * & * & \Theta_{44}^i & T_2^\top F\\ * & * & * & * & -\gamma^2 I\end{bmatrix},\\
\mathcal{Q}_1&=\begin{bmatrix}N_1 & U\\ * & N_1\end{bmatrix},\\
\mathcal{Q}_2&=\begin{bmatrix}Q_2 & V\\ * & Q_2\end{bmatrix},\\
\Theta_{11}^1&=\sum_{q=1}^{S}\pi_{pq}P_q-M_1+Q_1-Q_2+T_1^TA+A^TT_1-\rho P_p-\tau\rho M_1-H_1-H_1^T,\\
\Theta_{11}^2&=\sum_{q=1}^{S}\pi_{pq}P_q-M_1+Q_1-Q_2+T_1^TA+A^TT_1-\rho P_p-H_1-H_1^T,\\
\Theta_{12}^1&=P_p+M_1-M_2-N_2+T_1^TB\alpha_pKC-\tau\rho\left(-M_1+M_2\right)+H_6^T-H_2,\\
\Theta_{12}^2&=P_p+M_1-M_2-N_2+T_1^TB\alpha_pKC+H_6^T-H_2,\\
\Theta_{13}&=-H_3+V,\\
\Theta_{14}^1&=\tau M_1-T_1^T+A^TT_2-H_4,\\
\Theta_{14}^2&=-T_1^T+A^TT_2-H_4,\\
\Theta_{15}&=T_1^TF-H_5,\\
\Theta_{22}^1&=-M_1+M_2+M_2^T+N_2+N_2^T-\tau\rho\left(M_1-M_2-M_2^T\right)+H_7+H_7^T,\\
\Theta_{22}^2&=-M_1+M_2+M_2^T+N_2+N_2^T-\tau N_3+H_7+H_7^T,\\
\Theta_{24}^1&=-M_1^T+\tau\left(M_2^T+N_2^T\right)+\left(T_2^TB\alpha KC\right)^T+H_9,\\
\Theta_{24}^2&=-M_1^T+\left(T_2^TB\alpha KC\right)^T+H_9,\\
\Theta_{33}&=-Q_1-Q_2,\\
\Theta_{44}^1&=\tau N_1+\tau^2Q_2-T_2^T-T_2,\\
\Theta_{44}^2&=\tau^2Q_2-T_2^T-T_2,\\
\Theta_{44}^1&=\tau N_1+\tau^2Q_2-T_2^T-T_2,\\
\Theta_{44}^2&=\tau^2Q_2-T_2^T-T_2,\\
\psi&\triangleq\max_{p\in\mathcal{S}}\left\{\lambda_{\max}\left(\widetilde{P}_p\right)\right\}+\lambda_{\max}\left(\widetilde{Q}_1\right)+\lambda_{\max}\left(\widetilde{Q}_2\right),\\
\widetilde{P}_p&=R^{-(1/2)}Q_pR^{-(1/2)},\\
\widetilde{Q}_s&=R^{-(1/2)}Q_sR^{-(1/2)},\quad s=1,2.
\end{aligned} \tag{19}$$

Proof. Establishing a Lyapunov functional as $V(t,r_t)=\sum_{i=1}^{4}V_i(t,r_t)$, where

$$\begin{aligned}
V_1(t,r_t)&=\delta^T(t)P(r_t)\delta(t),\\
V_2(t,r_t)&=(\tau-(t-t_k))\vartheta_1^T(t)\mathscr{M}\vartheta_1(t),\\
V_3(t,r_t)&=(\tau-(t-t_k))\int_{t_k}^{t}e^{\rho(t-s)}\vartheta_2^T(s,t_k)\mathscr{N}\vartheta_2(s,t_k)\mathrm{d}s,\\
V_4(t,r_t)&=\int_{t-\tau}^{t}e^{\rho(t-s)}\delta^T(s)Q_1\delta(s)\mathrm{d}s+\tau\int_{-\tau}^{0}\int_{t+\theta}^{t}e^{\rho(t-s)}\dot{\delta}^T(s)Q_2\dot{\delta}(s)\mathrm{d}s\mathrm{d}\theta,
\end{aligned}\tag{20}$$

where

$$\begin{aligned}
\vartheta_1(t)&=\begin{bmatrix}\delta^T(t) & \delta^T(t_k)\end{bmatrix}^T,\\
\vartheta_2(s,t_k)&=\begin{bmatrix}\dot{\delta}^T(s) & \delta^T(t_k)\end{bmatrix}^T,\\
\mathscr{M}&=\begin{bmatrix}M_1 & -M_1+M_2\\ * & M_1-M_2-M_2^T\end{bmatrix},\\
\mathscr{N}&=\begin{bmatrix}N_1 & N_2\\ * & N_3\end{bmatrix}.
\end{aligned}\tag{21}$$

The weak infinitesimal operator $\mathscr{L}\{\cdot\}$ can be inferred as

$$\mathscr{L}V(t,r_t)=\frac{\partial V(t,r_t)}{\partial t}+\dot{\delta}^T(t)\frac{\partial V(t,r_t)}{\partial t}+\sum_{q=\epsilon\mathcal{S}}\pi_{pq}P_qV(t,r_t).\tag{22}$$

Applying the operator $\mathscr{L}\{\cdot\}$ along the power system (10), which yields

$$\mathscr{L}V_1(t,r_t)=\delta^T(t)P_p\dot{\delta}(t)+\dot{\delta}^T(t)P\delta(t)+\delta^T(t)\sum_{q=1}^{S}\pi_{pq}P_q\delta(t),\tag{23}$$

$$\mathscr{L}V_2(t,r_t)=(\tau-(t-t_k))\vartheta_1^T(t)\mathscr{M}\begin{bmatrix}\dot{\delta}^T(t) & 0\end{bmatrix}^T+(\tau-(t-t_k))\begin{bmatrix}\dot{\delta}(t) & 0\end{bmatrix}^T\mathscr{M}\vartheta_1(t)-\vartheta_1^T(t)\mathscr{M}\vartheta_1(t),\tag{24}$$

$$\begin{aligned}
\mathscr{L}V_3(t,r_t)\le&-\int_{t_k}^{t}\dot{\delta}^T(s)N_1\dot{\delta}(s)\mathrm{d}s-2\delta^T(t)U_2\delta(t_k)+\delta^T(t_k)\left(N_2+N_2^T-(t-t_k)N_3\right)\delta(t_k)\\
&+(\tau-(t-t_k))\dot{\delta}^T(t)N_1\dot{\delta}(t)+2(\tau-(t-t_k))\dot{\delta}^T(t)N_2\delta(t_k)+\rho V_3(t,r_t),
\end{aligned}\tag{25}$$

$$\mathscr{L}V_4(t,r_t)\le\delta^T(t)Q_1\delta(t)-\delta^T(t-\tau)Q_1\delta(t-\tau)+\tau^2\dot{\delta}^T(t)Q_2\dot{\delta}(t)-\tau\int_{t-\tau}^{t}\dot{\delta}^T(s)Q_2\dot{\delta}(s)\mathrm{d}s+\rho V_4(t,r_t).\tag{26}$$

Based on Lemma 1, the following inequality can be devised:

$$-\int_{t_k}^{t}\dot{\delta}^T(s)N_1\dot{\delta}(s)\mathrm{d}s\le-\frac{1}{t-t_k}\xi^T(t)\mathscr{W}_1^{\mathscr{T}}\mathscr{Q}_1\mathscr{W}_1\xi(t),\tag{27}$$

$$-\tau\int_{t-\tau}^{t}\dot{\delta}^T(s)Q_2\dot{\delta}(s)\mathrm{d}s\le-\xi^T(t)\mathscr{W}_2^{\mathscr{T}}\mathscr{Q}_2\mathscr{W}_2\xi(t),\tag{28}$$

where

$$\begin{aligned}
\xi^T(t)&=\begin{bmatrix}\delta^T(t) & \delta^T(t_k) & \delta^T(t-\tau) & \dot{\delta}(t) & \omega^T(t)\end{bmatrix},\\
\mathscr{W}_1&=\begin{bmatrix}I & 0 & 0 & 0 & 0\\ 0 & -I & 0 & 0 & 0\end{bmatrix},\\
\mathscr{W}_2&=\begin{bmatrix}I & 0 & 0 & 0 & 0\\ 0 & 0 & -I & 0 & 0\end{bmatrix}.
\end{aligned}\tag{29}$$

It is well known that for any matrices $\mathscr{H}$, one gets

$$\frac{1}{t-t_k}(\mathscr{Q}_1\mathscr{W}_1-(t-t_k)\mathscr{H})^T\mathscr{Q}_1^{-1}(\mathscr{Q}_1\mathscr{W}_1-(t-t_k)\mathscr{H})\ge0.\tag{30}$$

The aforementioned condition can be rewritten as

$$-\frac{1}{t-t_k}\mathscr{W}_1^T\mathscr{Q}_1\mathscr{W}_1\le-\mathscr{H}^T\mathscr{W}_1-\mathscr{W}_1^T\mathscr{H}+(t-t_k)\mathscr{H}^T\mathscr{Q}_1^{-1}\mathscr{H}.\tag{31}$$

On the other hand, for any matrices T_1 and T_2, it is clear that

$$\begin{aligned}
0=2&\left(\delta^T(t)T_1^T+\dot{\delta}^T(t)T_2^T\right)\left[-\dot{\delta}(t)+A\delta(t)\right.\\
&\left.+B\alpha_pKC\delta(t_k)+F\omega(t)\right].
\end{aligned}\tag{32}$$

Substituting (23)–(32) into (20), it can be deduced that

$$\mathrm{E}\{\mathscr{L}V(t,r_t)\}-\rho V(t,r_t)-\gamma^2\omega^T(t)\omega(t)\le\xi^T(t)\left[\Theta(t)+(t-t_k)\mathscr{H}^T\mathscr{Q}_1^{-1}\mathscr{H}\right]\xi(t),\quad t\in[t_k,t_{k+1}),\tag{33}$$

where

$$\begin{aligned}
\Theta(t)&=\begin{bmatrix}\Theta_{11}&\Theta_{12}&\Theta_{13}&\Theta_{14}&\Theta_{15}\\ *&\Theta_{22}&H_8^T&\Theta_{24}&H_{10}\\ *&*&\Theta_{33}&0&0\\ *&*&*&\Theta_{44}&T_2^TF\\ *&*&*&*&-\gamma^2I\end{bmatrix},\\
\Theta_{11}&=\sum_{q=1}^{S}\pi_{pq}P_q-M_1+Q_1-Q_2+T_1^TA+A^TT_1-\rho P_p-(\tau-(t-t_k))\rho M_1-H_1-H_1^T,\\
\Theta_{12}&=P_p+M_1-M_2-N_2+T_1^TB\alpha_pKC-(\tau-(t-t_k))\rho(-M_1+M_2)+H_6^T-H_2,\\
\Theta_{13}&=-H_3+V,\\
\Theta_{14}&=(\tau-(t-t_k))M_1-T_1^T+A^TT_2-H_4,\\
\Theta_{15}&=T_1^TF-H_5,\\
\Theta_{22}&=-M_1+M_2+M_2^T+N_2+N_2^T-(t-t_k)N_3-(\tau-(t-t_k))\rho\left(M_1-M_2-M_2^T\right)+H_7+H_7^T,\\
\Theta_{24}&=-M_1^T+(\tau-(t-t_k))\left(M_2^T+N_2^T\right)+\left(T_2^TB\alpha KC\right)^T+H_9,\\
\Theta_{33}&=-Q_1-Q_1,\\
\Theta_{44}&=(\tau-(t-t_k))N_1+\tau^2Q_2-T_2^T-T_2.
\end{aligned}\tag{34}$$

Note that (33) is a convex combination of $t-t_k$ and $\tau-(t-t_k)$, in accordance with Schur complement; one can deduce that $\Theta(t)+(t-t_k)\mathscr{H}^T\mathscr{Q}_1^{-1}\mathscr{H}<0$ if and only if (15) and (16) hold. Therefore, one can see that

$$\mathrm{E}\{\mathscr{L}V(t,r_t)\}<\rho V(t,r_t)+\gamma^2\omega^T(t)\omega(t),\quad t\in[t_k,t_{k+1}).\tag{35}$$

By integrating the both sides of (35) from t_k to t and simple derivation, it yields

$$\begin{aligned}
\mathrm{E}\{V(t,p)\}&<e^{\rho(t-t_k)}\mathrm{E}\{V(t_k,p)\}+\int_{t_k}^{t}e^{\rho(t-s)}\gamma^2\omega^T(s)\omega(s)\mathrm{d}s\\
&\le e^{\rho(t-t_{k-1})}\mathrm{E}\{V(t_{k-1},p)\}+e^{\rho(t-t_{k-1})}\int_{t_{k-1}}^{t_k}e^{\rho(t_k-s)}\gamma^2\omega^T(s)\omega(s)\mathrm{d}s\\
&\quad+\int_{t_k}^{t}e^{\rho(t-s)}\gamma^2\omega^T(s)\omega(s)\mathrm{d}s\le\cdots\le e^{t-t_0}\mathrm{E}\{V(t_0,p)\}+e^{\rho(t-t_1)}\int_{t_0}^{t_1}e^{\rho(t_1-s)}\gamma^2\omega^T(s)\omega(s)\mathrm{d}s\\
&\quad+\cdots+\int_{t_k}^{t}e^{\rho(t-s)}\gamma^2\omega^T(s)\omega(s)\mathrm{d}s\le e^{\rho(t-t_0)}\left(\mathrm{E}\{V(t_0,p)\}+\int_{t_0}^{t}\gamma^2\omega^T(s)\omega(s)\mathrm{d}s\right)\\
&\le e^{\rho T}\left(\mathrm{E}\{V(t_0,p)\}+\gamma^2\overline{\omega}\right).
\end{aligned}\tag{36}$$

Recalling the Lyapunov functional (20), we can get

$$\mathrm{E}\{V(t,r_t)\} \geq \lambda_{\min}\left(\widetilde{P}_p\right)\delta^T(t)R\delta(t), \tag{37}$$

$$\mathrm{E}\{V(t_0,r_t)\} \leq \left(\max_{p\in\mathcal{S}}\left\{\lambda_{\max}\left(\widetilde{P}_p\right)\right\} + \lambda_{\max}(Q_1) + \lambda_{\max}(Q_2)\right)\delta^T(t_0)R\delta(t_0). \tag{38}$$

Substituting (37) and (38) into (36), we can obtain

$$\inf_{p\in\mathcal{S}}\left\{\lambda_{\min}\left(\widetilde{P}_p\right)\right\}\mathrm{E}\left\{\delta^T(t)R\delta(t)\right\} < e^{\rho T}\left(\psi c_1 + \gamma^2\overline{\omega}\right). \tag{39}$$

In light of (17), it can be concluded from (39) that $\mathscr{E}\left\{\delta^T(t)R\delta(t)\right\} < c_2$. Thus, from Definition 2, we have to derive that power system (10) is SFTB over the time interval $[t_0, T]$.

In the following, the actuator constraints (18) will be discussed. In light of (9), one has

$$\begin{aligned}\left|u_s^F(t)\right| &\leq \left|u_s(t)\right| = \left|K_sC\delta(t_k)\right|^2 \\ &\leq \left|K_sCR^{-(1/2)}\right|^2 \times \left|R^{1/2}\delta(t_k)\right|^2 \\ &= K_sCR^{-1}C^TK_s^T\delta^T(t_k)R\delta(t_k) \\ &\leq c_2K_sCR^{-1}C^TK_s^T.\end{aligned} \tag{40}$$

Recalling Assumption 2, it yields

$$c_2K_sCR^{-1}C^TK_s^T < u_{s\max}^2, \quad s = 1,2,\ldots,n_u. \tag{41}$$

According to Schur complement, (18) can be guaranteed by (41), which completes the proof of Theorem 1. □

Theorem 2. *For given parameters $\rho > 0$, $c_1 > 0$, $c_2 > 0$, $T > 0$, $\overline{\omega} > 0$, $u_{\max}$, and matrix R, the closed-loop power system (10) is called SFTB with respect to $(c_1, c_2, \overline{\omega}, T, R)$ and meet an H_∞ performance index $\overline{\gamma} = \gamma\sqrt{e^{\rho T}}$, if there exists matrix $P_p > 0$, $N_1 > 0$, $N_3 > 0$, $Q_1 > 0$, $Q_2 > 0$, and matrices X, Y with suitable dimensions, such that $\forall p \in \mathcal{S}$*

$$\begin{bmatrix} \overline{\Theta}^1 & \mathscr{Z}_1 & \mathscr{Z}_2 \\ * & -\varepsilon_1\left(X + X^T\right) & 0 \\ * & * & -\varepsilon_2\left(X + X^T\right) \end{bmatrix} < 0, \tag{42}$$

$$\begin{bmatrix} \overline{\Theta}^2 & \tau\mathscr{H}^T & \mathscr{Z}_1 & \mathscr{Z}_2 \\ * & -\tau\mathscr{Q}_1^{-1} & 0 & 0 \\ * & * & -\varepsilon_1\left(X + X^T\right) & 0 \\ * & * & * & -\varepsilon_2\left(X + X^T\right) \end{bmatrix} < 0, \tag{43}$$

$$e^{\rho T}\overline{\psi}c_1 + e^{\rho T}\gamma^2\overline{\omega} < \lambda_1 c_2, \tag{44}$$

$$\lambda_1 R < P_p < \lambda_2 R, \quad Q_1 < \lambda_3 R, \quad Q_2 < \lambda_4 R, \tag{45}$$

$$\begin{bmatrix} -u_{s\max}^2 & YC & I - X \\ * & -\dfrac{R}{c_2} & \varepsilon_3C^TY^T \\ * & * & -\varepsilon_3\left(X + X^T\right) \end{bmatrix} < 0, \quad s = 1,2,\ldots,n_u, \tag{46}$$

where

$$\Theta^i = \begin{bmatrix} \overline{\Theta}_{11}^i & \Theta_{12}^i & \Theta_{13} & \Theta_{14}^i & \Theta_{15} \\ * & \Theta_{22}^i & H_8^T & \Theta_{24}^i & H_{10} \\ * & * & \Theta_{33} & 0 & 0 \\ * & * & * & \Theta_{44}^i & T_2^T F \\ * & * & * & * & -\gamma^2 I \end{bmatrix},$$

$$\overline{\psi} \triangleq \lambda_2 + \tau\lambda_3 + \tau^2\lambda_4,$$

$$\overline{\Theta}_{11}^1 = \sum_{q=1}^{S} \pi_{pq} P_q - M_1 + Q_1 - Q_2 + T_1^T A + A^T T_1 - \rho P_p - \tau\rho M_1 - H_1 - H_1^T - C^T C,$$

$$\overline{\Theta}_{11}^2 = \sum_{q=1}^{S} \pi_{pq} P_q - M_1 + Q_1 - Q_2 + T_1^T A + A^T T_1 - \rho P_p - H_1 - H_1^T - C^T C,$$

$$\mathscr{T}_1 = \begin{bmatrix} \alpha\left(T_1^T B - BX\right)^T & \varepsilon YC & 0 & 0 & 0 \end{bmatrix}^T,$$

$$\mathscr{T}_2 = \begin{bmatrix} YC & 0 & 0 & \alpha\left(T_2^T B - BX\right)^T & 0 \end{bmatrix}^T. \tag{47}$$

The controller gain is deduced as

$$K = X^{-1} Y. \tag{48}$$

Proof. By resorting to the same Lyapunov functional as displayed in (20), and adopting the similar derivations in the proof of Theorem 1, for $t \in [t_k, t_{k+1})$, the following inequality can be realized:

$$\mathrm{E}\{\mathscr{L}V(t, r_t)\} < \rho V(t, r_t) - y^T(t) y(t) + \gamma^2 \omega^T(t)\omega(t), \quad t \in [t_k, t_{k+1}). \tag{49}$$

Integrating the both sides of (49) from t_k to t, it yields

$$\mathrm{E}\{V(t, p)\} < e^{\rho(t-t_k)} V(t_k, p) + \int_{t_k}^{t} e^{\rho(t-s)} \left(-y^T(s) y(s) + \gamma^2 \omega^T(s)\omega(s)\right) \mathrm{d}s, \quad t \in [t_k, t_{k+1}). \tag{50}$$

Clearly, (50) can be rewritten as

$$\mathrm{E}\{V(t, p)\} < e^{\rho(t-t_0)} V(t_0, p) + \int_{t_0}^{t} e^{\rho(t-s)} \left(-y^T(s) y(s) + \gamma^2 \omega^T(s)\omega(s)\right) \mathrm{d}s. \tag{51}$$

Under the zero-initial conditions, one gets $V(t_0, p) = 0$. It follows from (51) that

$$\int_{t_0}^{t} e^{\rho(t-s)} y^T(s) y(s) \mathrm{d}s < \int_{t_0}^{t} e^{\rho(t-s)} \gamma^2 \omega^T(s)\omega(s) \mathrm{d}s. \tag{52}$$

Since $1 \le e^{\rho(t-s)} \le e^{\rho t}$ and $t \in [t_0, T]$, (52) signifies

$$\mathrm{E}\left\{\int_{t_0}^{t} y^T(s) y(s) \mathrm{d}s\right\} < \overline{\gamma}^2 \int_{t_0}^{t} \omega^T(s)\omega(s) \mathrm{d}s, \tag{53}$$

with $\overline{\gamma} = \sqrt{e^{\rho T}}\gamma$. Therefore, the H_∞ performance index $\overline{\gamma}$ is ensured.

Meanwhile, in light of Lemma 1 of [28], conditions (42) and (43) can be easily attained. Similarly, (46) can be ensured. The proof is completed. □

4. Computational Experiments

In this example, a single-area power system is studied for simulating the derived results. Similar to [6], the parameters of system (1) are expressed in Table 2.

Taking the actuator faults into consideration, which is characterized by a Markov process, the possible fault coefficients are $\alpha_1 = 1$, $\alpha_2 = 0.8$, and $\alpha_3 = 0.5$. Furthermore, the transition probability matrix among faults is given by

Table 2: Physical meaning.

Parameters	R	D	$M(s)$	β	$T_g(s)$	$T_{ch}(s)$
	0.05	1.0	10	21.0	0.1	0.3

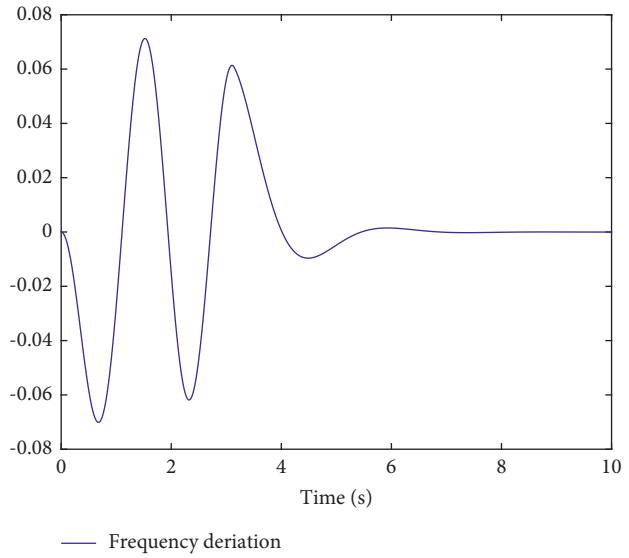


Figure 2: Evolution of frequency derivation.

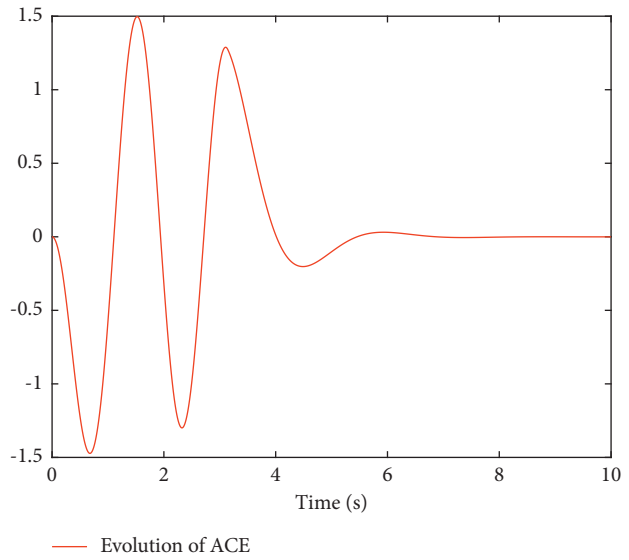


Figure 3: Evolution of ACE.

$$\Pi = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.8 & 0.1 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}. \tag{54}$$

On the other hand, other parameters are selected as $\tau = 0.2$, $\gamma = 0.6$, $\rho = 0.2$, $c_1 = 0.1$, $c_2 = 1$, $\overline{\omega} = 0.8$, $R = I_{4\times4}$, and $T = 8$. The control input $u(t)$ is supposed to be constrained by $|u(t)| \le u_{\max} = 2$. By solving the linear matrix inequalities of Theorem 2, the desired PI-type controller is derived as

$$K = \begin{bmatrix} -0.0002 & -0.0183 \end{bmatrix}. \tag{55}$$

For graphically verifying the achieved results, we select the initial state disturbance $\omega(t)$ as

$$\omega(t) = \begin{cases} 1.8 \sin(4t), & \text{if } t \le \pi, \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$

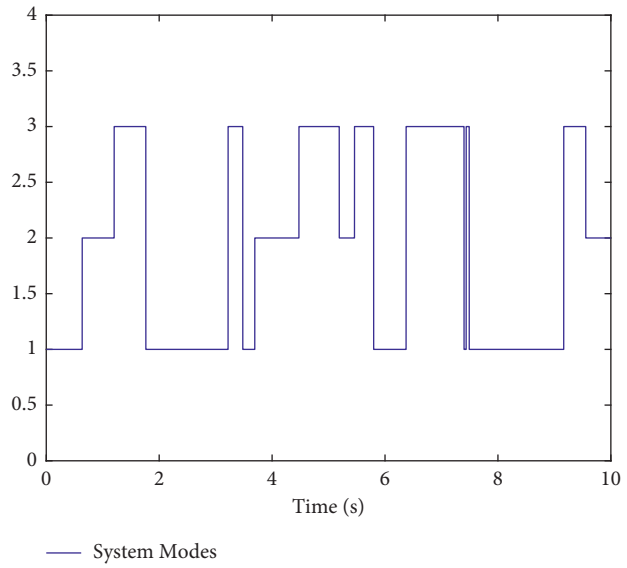


Figure 4: The mode switching of actuator faults.

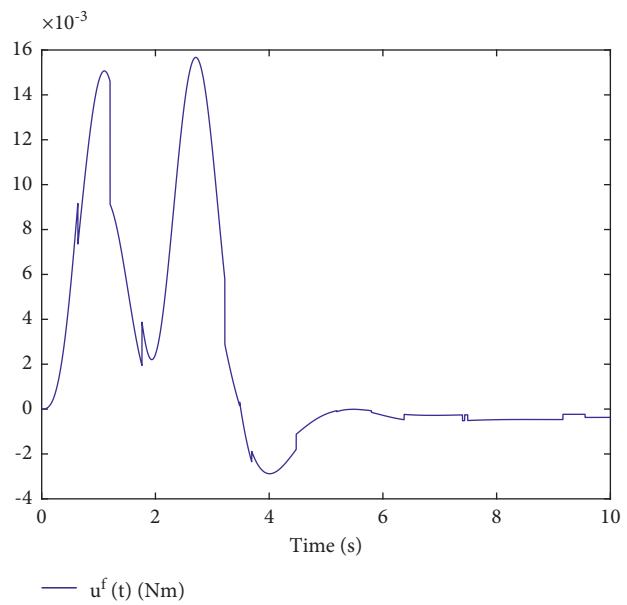


Figure 5: Control output.

Added by the aforementioned controller, the simulation results are plotted in Figures 2–7. Figure 2 plots the simulated frequency, and Figure 3 displays the evolution of ACE. Meanwhile, the mode switching of actuator faults is shown in Figure 4, and control output is presented in Figure 5. Furthermore, with the disturbance given in Figure 6, the evolution of $\delta^T(t)R\delta(t)$ is expressed in Figure 7. One can be observed from Figure 7 that the state of closed-loop system stays in the prefixed region, which implies the resulting system is SFTB. Meanwhile, the input constraint is also satisfied.

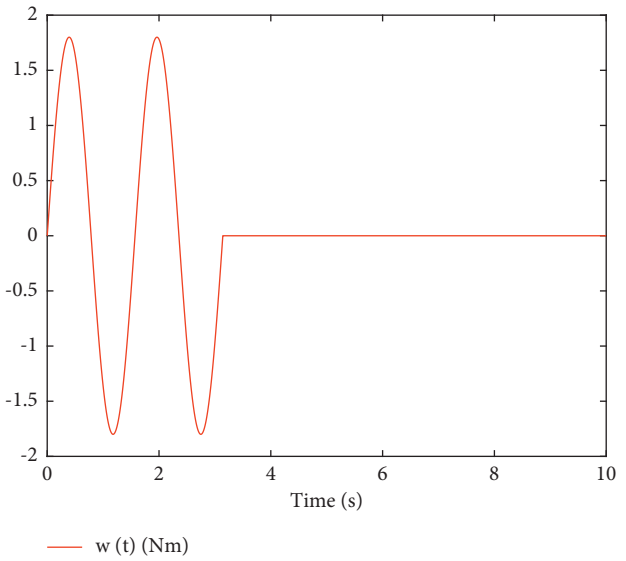


Figure 6: The disturbance $\omega(t)$.

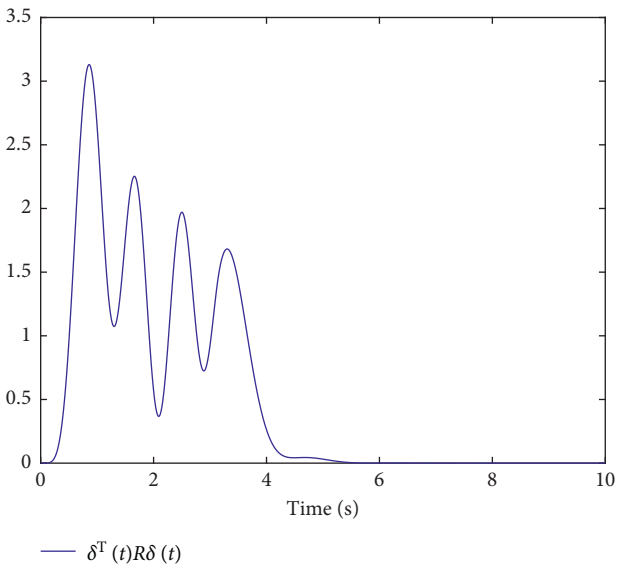


Figure 7: The evolution of $\delta^T(t)\,R\delta(t)$.

5. Conclusions

In this study, the finite-time LFC problem for power systems with actuator fault has been considered. To better reflect the actual demands of practical dynamics, a generalized framework of the actuator constraint has been studied. Given the randomly occurring actuator fault, a homogeneous Markov chain-based actuator fault has been studied. Together with the piecewise Lyapunov theory, sufficient conditions have been attained. In the end, a numerical example has been applied to verify the effectiveness of the developed results.

Data Availability

No data were used to support the current work.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. J. Grainger, W. D. Stevenson, and G. W. Chang, *Power System Analysis*, McGraw-Hill, New York, NY, USA, 1994.

[2] Y. Wang, R. Zhou, and C. Wen, "Robust load-frequency controller design for power systems," *IEEE Proceedings C Generation, Transmission and Distribution*, vol. 140, no. 1, pp. 11–16, 1993.

[3] L. Jiang, W. Yao, Q. H. Wu, J. Y. Wen, and S. J. Cheng, "Delay-dependent stability for load frequency control with constant and time-varying delays," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 932–941, 2012.

[4] M. Li and Y. Chen, "A wide-area dynamic damping controller based on robust H_∞ control for wide-area power systems with random delay and packet dropout," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4026–4037, 2018.

[5] C. Peng and J. Zhang, "Delay-distribution-dependent load frequency control of power systems with probabilistic interval delays," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 3309–3317, 2016.

[6] S. Kuppusamy and Y. H. Joo, "Resilient reliable H_∞ load frequency control of power system with random gain fluctuations," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–9, 2021.

[7] H. Zhang, J. Liu, and S. Xu, "H_∞ load frequency control of networked power systems via an event-triggered scheme," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 7104–7113, 2020.

[8] Q. Zhao and J. Jiang, "Reliable state feedback control system design against actuator failures," *Automatica*, vol. 34, no. 10, pp. 1267–1272, 1998.

[9] K. Zhang, B. Jiang, and M. Staroswiecki, "Dynamic output feedback-fault tolerant controller design for Takagi-Sugeno fuzzy systems with actuator faults," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 1, pp. 194–201, 2010.

[10] W. Zhou, J. Fu, H. Yan, X. Du, Y. Wang, and H. Zhou, "Event-triggered approximate optimal path-following control for unmanned surface vehicles with state constraints," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[11] S. Yin, H. Yang, and O. Kaynak, "Sliding mode observer-based FTC for Markovian jump systems with actuator and sensor faults," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3551–3558, 2017.

[12] J. Tao, R. Lu, P. Shi, H. Su, and Z.-G. Wu, "Dissipativity-based reliable control for fuzzy Markov jump systems with actuator faults," *IEEE Transactions on Cybernetics*, vol. 47, no. 9, pp. 2377–2388, 2017.

[13] Y. Wei, J. Qiu, P. Shi, and L. Wu, "A piecewise-Markovian Lyapunov approach to reliable output feedback control for fuzzy-affine systems with time-delays and actuator faults," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2723–2735, 2018.

[14] J. Cheng, J. H. Park, X. Zhao, H. R. Karimi, and J. Cao, "Quantized nonstationary filtering of networked Markov switching RSNSs: a multiple hierarchical structure strategy," *IEEE Transactions on Automatic Control*, vol. 65, no. 11, pp. 4816–4823, 2020.

[15] Y. Wang, H. Pu, P. Shi, C. K. Ahn, and J. Luo, "Sliding mode control for singularly perturbed Markov jump descriptor systems with nonlinear perturbation," *Automatica*, vol. 127, Article ID 109515, 2021.

[16] J. Cheng, J. H. Park, and Z.-G. Wu, "Observer-based asynchronous control of nonlinear systems with dynamic event-based try-once-discard protocol," *IEEE Transactions on Cybernetics*, pp. 1–11, 2021.

[17] F. Li, W. X. Zheng, and S. Xu, "Finite-time fuzzy control for nonlinear singularly perturbed systems with input constraints," *IEEE Transactions on Fuzzy Systems*, p. 1, 2021.

[18] J. Cheng, D. Zhang, W. Qi, J. Cao, and K. Shi, "Finite-time stabilization of T-S fuzzy semi-Markov switching systems: a coupling memory sampled-data control approach," *Journal of the Franklin Institute*, vol. 357, no. 16, pp. 11256–11280, 2020.

[19] J. Cheng, Y. Wang, J. H. Park, J. Cao, and K. Shi, "Static output feedback quantized control for fuzzy Markovian switching singularly perturbed systems with deception attacks," *IEEE Transactions on Fuzzy Systems*, p. 1, 2021.

[20] R. Vadivel and Y. H. Joo, "Finite-time sampled-data fuzzy control for a non-linear system using passivity and passification approaches and its application," *IET Control Theory & Applications*, vol. 14, no. 8, pp. 1033–1045, 2020.

[21] M. Li and Y. Chen, "Robust tracking control of networked control systems with communication constraints and external disturbance," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 4037–4047, 2017.

[22] J. Cheng, J. H. Park, J. Cao, and W. Qi, "Asynchronous partially mode-dependent filtering of network-based MSRSNSs with quantized measurement," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3731–3739, 2020.

[23] G. Liu, C. Hua, P. X. Liu, and J. H. Park, "Input-to-state stability for time-delay systems with large delays," *IEEE Transactions on Cybernetics*, pp. 1–9, 2021.

[24] L. Hetel, C. Fiter, H. Omran et al., "Recent developments on the stability of systems with aperiodic sampling: an overview," *Automatica*, vol. 76, pp. 309–335, 2017.

[25] D. Zhang, Z. Xu, H. R. Karimi, Q.-G. Wang, and L. Yu, "Distributed H_∞ output-feedback control for consensus of heterogeneous linear multiagent systems with aperiodic sampled-data communications," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4145–4155, 2018.

[26] S. Xu, G. Sun, Z. Li, and H. Zheng, "Finite-time robust fuzzy control for non-linear Markov jump systems under aperiodic sampling and actuator constraints," *IET Control Theory & Applications*, vol. 11, no. 15, pp. 2419–2431, 2017.

[27] P. Park, J. W. Ko, and C. Jeong, "Reciprocally convex approach to stability of systems with time-varying delays," *Automatica*, vol. 47, no. 1, pp. 235–238, 2011.

[28] S. Dong, M. Fang, and S. Chen, "Extended dissipativity asynchronous static output feedback control of Markov jump systems," *Information Sciences*, vol. 514, pp. 375–387, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5154769, 11 pages
https://doi.org/10.1155/2021/5154769

Research Article

An Easy-to-Integrate IP Design of AHB Slave Bus Interface for the Security Chip of IoT

Conggui Yuan,[1] Xin Zheng ,[2] Bo Rao,[2] and Shuting Cai[2]

[1]*School of Electronic Information, Dongguan Polytechnic, Dongguan, China*
[2]*School of Automation, Guangdong University of Technology, Guangzhou, China*

Correspondence should be addressed to Xin Zheng; xinzheng9209@gmail.com

Received 16 August 2021; Revised 29 September 2021; Accepted 18 October 2021; Published 1 November 2021

Academic Editor: Chuan Li

Information security is fundamental to the Internet of things (IoT) devices, in which security chip is an important means. This paper proposes an Advanced High-performance Bus Slave Control IP (AHB-SIP), which applies to cryptographic accelerators in IoT security chips. Composed by four types of function registers and AHB Interface Control Logic (AICL), AHB-SIP has a simple and easy-to-use structure. The System on Chip (SoC) design can be realized by quickly converting the nonstandard interface of the security module to the AHB slave interface. AHB-SIP is applied to the security accelerators of SM2, SM3, and SM4 and random number generator (RNG). Combined with a low-power embedded CPU, TIMER, UART, SPI, IIC, and other communication interfaces, a configurable SoC can be integrated. Moreover, SMIC 110 nm technology is employed to tape out the SoC on a silicon chip. The area of AHB-SIP is 0.072 mm^2, only occupying 6‰ of the chip (3.45*3.45 mm^2), and the power consumption of encryption modules combined with AHB-SIP is lower than that combined with AXI interface, which is decreased up to 61.0% and is ideal for the application of IoT.

1. Introduction

The IoT is a network system that is extended and expanded on the Internet and connects people, devices, and servers. With the popularity of intelligent terminals and the rapid development of artificial intelligence, the majority of intelligent nodes will have the access to the Internet in the future. Regardless of its advantages, IoT technology has caused various security threats, such as the leakage of user privacy information and the attack vulnerability of hard-coded security keys [1, 2]. At the end of 2016, a large number of IoT devices were infected with the Mirai malware. The hackers formed a botnet and launched a DDoS attack against Dyn, a globally DNS provider. Consequently, consumers could not pay on PayPal websites, and users could not log in to social networking sites such as Twitter and Tumblr [3]. Frustaci et al. mention that security is the key issue of IoT [3]. Therefore, low-power, secure, and real-time physical layer SoC security chips play a crucial role in the IoT security domain. Besides, it is of great importance to efficiently design this security chip. Shorter life cycles of products can significantly reduce the time-to-market and rapid simulation capabilities are necessary with the increase of the design space at the early stages of design [4]. In this regard, this study focuses on the design of a highly efficient, low-power, and easy-integrated IP interface and integrated crypto modules.

Five SoC bus standards have been widely used in the design of bus interfaces, including the AMBA Bus [5], the Wishbone Bus [6], the CoreConnect Bus [7], the Avalon bus [8], and the OCP bus [9]. The AMBA is a bus standard for high-performance embedded systems. With many third-party supports, the AMBA has become one of the existing widely supported interconnection standards [5]. The CoreConnect bus is a fully constructed general-purpose solution that can connect high-performance systems such as workstations, but it may be too complex for simple embedded applications [7]. The Wishbone bus and the OCP bus are extensively applied in small embedded systems. The Avalon bus only applies to a series of programmable logic devices

(PLD) [6, 9]. The difference between these SoC buses is the features they provide and the integrity of the specification. According to the reference [10], multiple asynchronous AHB bus interface units were present, which allowed the communication of an OpenGL ES 2.0 vertex shader (VS) processor with other hardware units through the AHB bus in the case of different frequencies. The interface of IDE hard disk, reconfigurable arbiter, and DMA controller were designed with the AHB bus interface in [5, 11, 12], respectively. It is possible to interchangeably adopt the majority of AHB slaves in an AHB-Lite or AHB system. The slave designed for the AHB-Lite system will work in the full AHB and AHB-Lite designs.

In this study, the slave modules are security accelerators. To the best of our knowledge, sensitive information can be protected by utilizing cryptographic algorithms in the proposed solutions. Cryptographic algorithms are classified into three categories: symmetric cryptographic algorithms, asymmetric cryptographic algorithms, and hash algorithm. With the characteristics of high efficiency and low overhead, symmetric cryptographic algorithms (such as AES, SM4, and DES) are suitable for big data encryption. Asymmetric cryptographic algorithms, also known as public-key algorithms (such as RSA, SM2, and ECC), show high security. However, due to the large size of the key, it does not apply to big data encryption. Hash algorithms (such as SHA-1, SHA-256, and SM3) are mainly used to generate a message digest with a fixed length. A configurable SoC with built-in FPGA logic gates that can achieve multiple algorithms for AES and DES is proposed in [13]. A SoC is developed in [14], which can be used in the field of mobile security, but it does not apply to the IoT due to its size and power constraints. A SM3 algorithm integrated into financial IC card is designed in [15], which has low power and small area. In [16], a codesign method is employed to propose an AES-ECC hybrid cryptosystem and an interesting trade-off exist between area occupation and speed. Crypto modules are different in terms of functions and interfaces. Therefore, the traditional method changes the nonstandard interface of specific modules into the AHB slave interface. However, due to the different functions of the slave modules, solving this issue will cost mass manpower and resources, leading to a longer product development cycle and higher costs.

For high-performance synthesizable design, the Advanced High-performance Bus Lite (AHB-Lite), as a part of the AMBA, can be employed in IoT chips. It is a transport interface that supports separate transport and provides excellent data transfer capability.

Compared with a complete AHB master, a transport interface can greatly simplify the interface design if masters are designed based on the AHB-Lite interface specification. All masters designed by the full AHB specification apply to an AHB-Lite system with no modification. Although the AHB bus has been widely used in the SoC, the study of AHB slave interface design for security chips is scarce. By analyzing the advantages and disadvantages of different SoC buses and considering the context of practical applications, this study introduces four functional registers and designs a simple and efficient slave bus controller in combination with the AHB-Lite protocol. Moreover, based on the study of symmetric cryptography, public-key cryptography, and hash algorithms, AHB-SIP for cryptographic accelerators in an IoT security chip is proposed. Even without detailed knowledge of the AHB bus protocol, designers can quickly transform a cryptographic accelerator with a nonstandard interface into an accelerator of the AHB slave interface through AHB-SIP. Therefore, AHB-SIP can improve the design efficiency of implementing an SoC system. Based on the above motivation, we make the following contributions:

(i) Based on the AHB-Lite bus, an easy-to-integrate and fast AHB-SIP IP is proposed, which can quickly convert from a nonstandard interface to an AHB-Lite interface. The slave security modules can be easily integrated into an SoC via the AHB-SIP, and all the slave modules can be configured by software.

(ii) The AHB interface control logic that is the key part of AHB-SIP is proposed, which is equipped with strong data transfer capability and low resource consumption.

(iii) This design is taped out on a silicon chip with SMIC 110 nm process. As the experiment results reveal, the area of AHB-SIP only accounts for 6‰ of the chip, and the security accelerators integrated with AHB-SIP can rapidly achieve the encrypted results.

(iv) We integrate three different security accelerators, which can meet the requirement of IoT devices. Specifically, the power consumption of AHB-SIP-based security accelerators is lower than that of AXI-based security accelerators.

The remainder of this paper is structured as follows: Section 2 introduces the background of the AHB. Section 3 describes the design of AHB-SIP. Implementation and integration of cryptographic accelerators are proposed in Section 4, and the results and analysis are shown in Section 5. Section 6 concludes the whole paper.

2. Background of AMBA AHB

The Advanced Microcontroller Bus Architecture (AMBA) is a high-performance embedded microcontroller on-chip communication standard proposed by ARM [17], which has become one of the most popular on-chip bus systems. The AMBA 2.0 bus standard defines three kinds of buses: the Advanced High-performance Bus (AHB), the Advanced System Bus (ASB), and the Advanced Peripheral Bus (APB) [18]. Figure 1 presents a typical AMBA system structure.

In Figure 1, high-performance and high-throughput modules, such as CPU, DMA, and RAM, are connected by the AHB bus. The ASB bus is a high-performance bus that can connect microprocessors and system peripherals. Compared with the AHB bus, the ASB has smaller data width, and a bidirectional data bus is used. Being simple and easy to use, the APB is generally applied in low-speed modules such as UART and SPI. Among the AMBA systems, the most widely used buses are the AHB and the APB. The AHB-Lite bus is simplified based on the AHB, where the

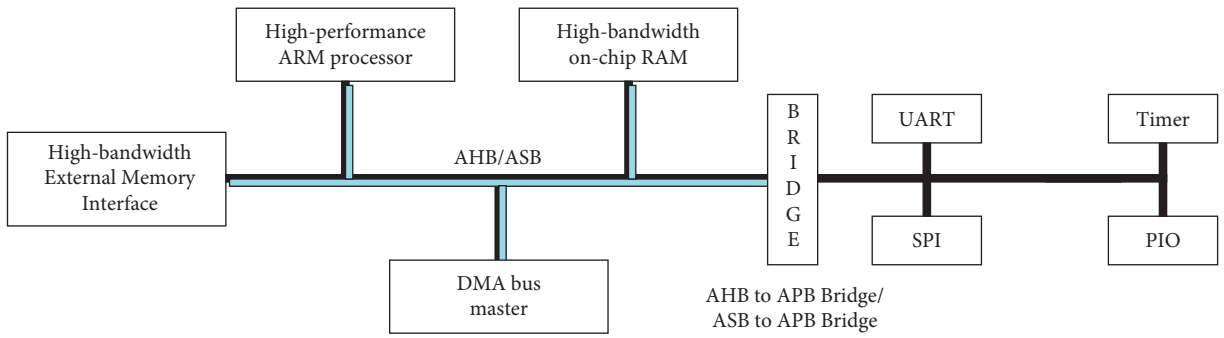


Figure 1: The block diagram of AHB system structure.

AHB supports multiple masters while the AHB-Lite supports only one master. Therefore, it is unnecessary to design an arbiter for the AHB-Lite. Generally, one master is designed in the security chip of IoT, so that the AHB-Lite bus protocol can be considered to use.

The SoC system with the AHB-Lite bus consists of three parts: master, slave, and infrastructure. The master device launches the data transmission, and the slave devices respond after receiving the access request from the master. As shown in Figure 2, an AHB-Lite system is composed of a slave-to-master multiplexor and an address decoder. The address from the master is monitored by the decoder to select the appropriate slave, and the multiplexor routes route the corresponding slave output data back to the master [19]. In our design, the requirements of high-performance synthesizable design are met by using the AHB-Lite.

3. Design of AHB-SIP

AHB-SIP is designed to easily integrate the security units into SoC, which can be configured by software through AHB-SIP, thereby improving the design efficiency of SoC. For the general high-performance computing module, the interface can be classified into four categories: data input, data output, control, and status. Based on different kinds of signals, data interaction is realized by designing four function registers (the status register, the control register, the output register, and the input register), so that the slave modules can be controlled. As is shown in Figure 3, the AHB-SIP consists of four function registers and an AHB Interface Control Logic (AICL) module. In our proposed design, the security units are the slave, and the embedded CPU is the master. The AHB-SIP transfers data between the master and the slave.

3.1. AHB Bus Interface Control Logic. In our design, the control logic is implemented based on the AHB timing. Figure 4 demonstrates the diagram of the AHB protocol sequence in the basic transmission mode.

The control logic is designed according to the AHB bus time sequence, which transfers data between the master and the function registers. The specific functions are divided into the following two aspects:

(1) When the master issues the write request of writing the control value, the data will be written from the master to the corresponding control register. The data will be written to the corresponding input register to write the ordinary data.

(2) When the master issues the read request, if the current status is required to be obtained, the value in the status register will be transmitted to the master. To read the ordinary data, the data in the output register will be sent to the master.

The AICL consists of the slave-to-master multiplexer, the data distributor, the address decoder, and the control logic. Different signals on the AHB bus are read by the control logic, and the control signals will be generated to control the data distributor, address decoder, and multiplexer. In this way, data can be read or written from registers. When the master reads data, the data selector outputs the data to the bus from the specified register based on the address signal and the control signal. When the master writes data, the data distributor will write the data to the corresponding register based on the address decoding result. The AICL module is shown in Figure 5.

3.2. Function Register. The four function registers are mainly adopted for control, calculation, data interaction, and reading status, which can not only realize effective control of the cryptographic accelerators but also obtain their current status for software debugging. Finally, with a 32 bit low-power embedded CPU as the master, the AHB-SIP is employed to integrate the three cryptographic modules and random number generation module into an SoC. The four types of registers are described as follows:

(1) Control register: the control register is utilized to control the start, stop, and working modes of the slave module (such as encryption, decryption, and random number generation)

(2) Input register: the data to be processed by the slave module from the master module are stored by the data input register

(3) Output registers: this type of register can store the data that have been processed by the slave module and are required to be transmitted to the master

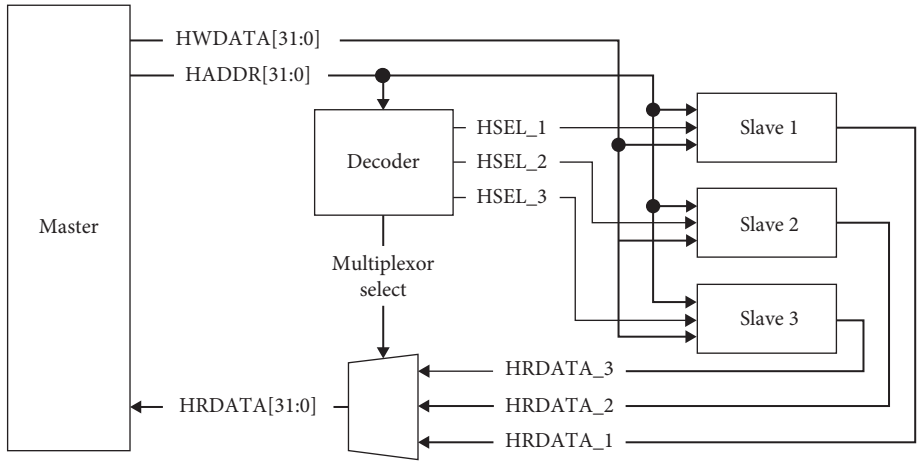


Figure 2: The AHB-Lite block diagram; reprinted from AMBA 5 AHB-Lite protocol [17].

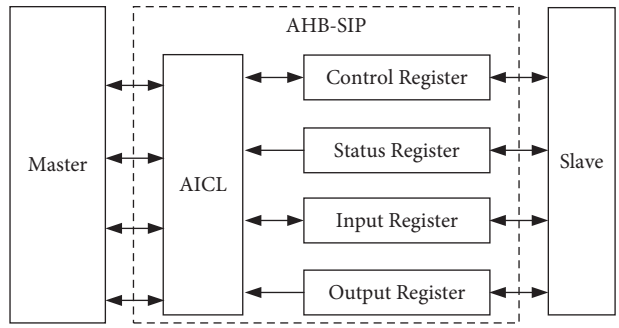


Figure 3: The block diagram of AHB-SIP structure.

(4) Status register: the status of the slave module, such as the mode, the status, and whether the operation is completed, is reflected by the status register

4. Integration of Cryptographic Accelerators

In the SoC, the cryptographic accelerators include the SM2, SM3, SM4, and RNG modules. This section investigates how to integrate these cryptographic accelerators into the SoC quickly by using the AHB-SIP to improve design efficiency. The cryptographic modules are connected with the CPU through AHB. To communicate with disparate IoT devices, the IIC, SPIs, GPIOs, and UARTs are also integrated into the security SoC. Besides, an SRAM is employed to run a real-time operating system (RTOS) and interact with cloud servers. Figure 6 is the architecture of the SoC. This section mainly presents the integration of cryptographic modules.

4.1. Integration of the SM2 Accelerator. SM2 is implemented based on the elliptic curve over GF(p) [20, 21]. The SM2 module is composed of modular operations and scalar multiplication operations. In our design, we utilize the binary extended Euclidean algorithm and the interleaved modular multiplication algorithm to decrease power consumption and chip area [14, 22]. Multiple 256 bit multiplexers, four 256 bit registers, and two 256 bit address are the main hardware overhead of SM2. The structure of the SM2 accelerator can be found in [23]. From the structure, it is observed that the SM2 is a 256 bit ECC. The input data include $(x1, y1)$, $(x2, y2)$, 256 bit key k, and the output data include $(x3, y3)$. Thus, 56 32 bit data registers are needed. The modes of SM2 include point multiplication (PM), multiple point (MP), point addition (PA), modular inverse (MI), modular multiplication (MM), modular subtraction (MS), and modular addition (MA). Therefore, we design a 32 bit status register and a 32 bit control register.

The control register of SM2 is responsible for controlling the computing pattern (enable, disable, or reset). The function of the control register is described in Table 1. For the enable control bit, it will be set to 1 automatically after completing the calculation. The reset control bit must be cleared before writing data. Otherwise, this module is always in reset.

The status register is designed to record the current working status of the SM2 accelerator so that the CPU can achieve the status of this module in real-time. SM2 has four states, which are idle, calculating, finish, and error, respectively.

4.2. Integration of the SM3 Accelerator. To meet the requirement of low power consumption, the proposed SM3 cryptographic accelerator mainly expands and compresses messages that are the most time-consuming parts. The padding and parsing processes are developed by software.

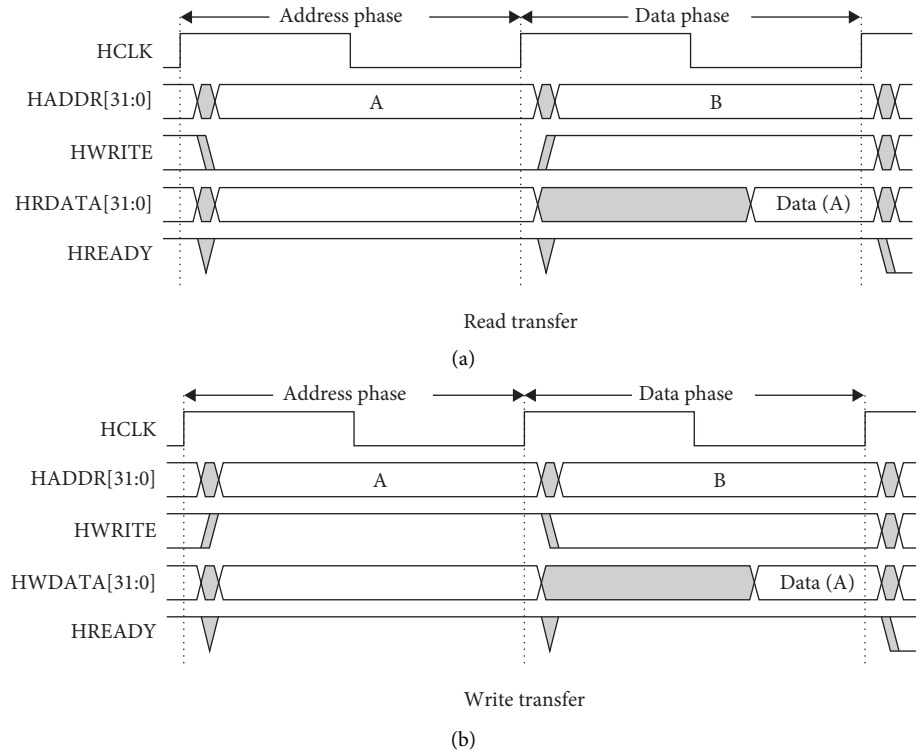


Figure 4: AHB protocol sequence in the basic transmission mode; reprinted from AMBA 5 AHB-lite protocol [17]. (a) Read transfer. (b) Write transfer.

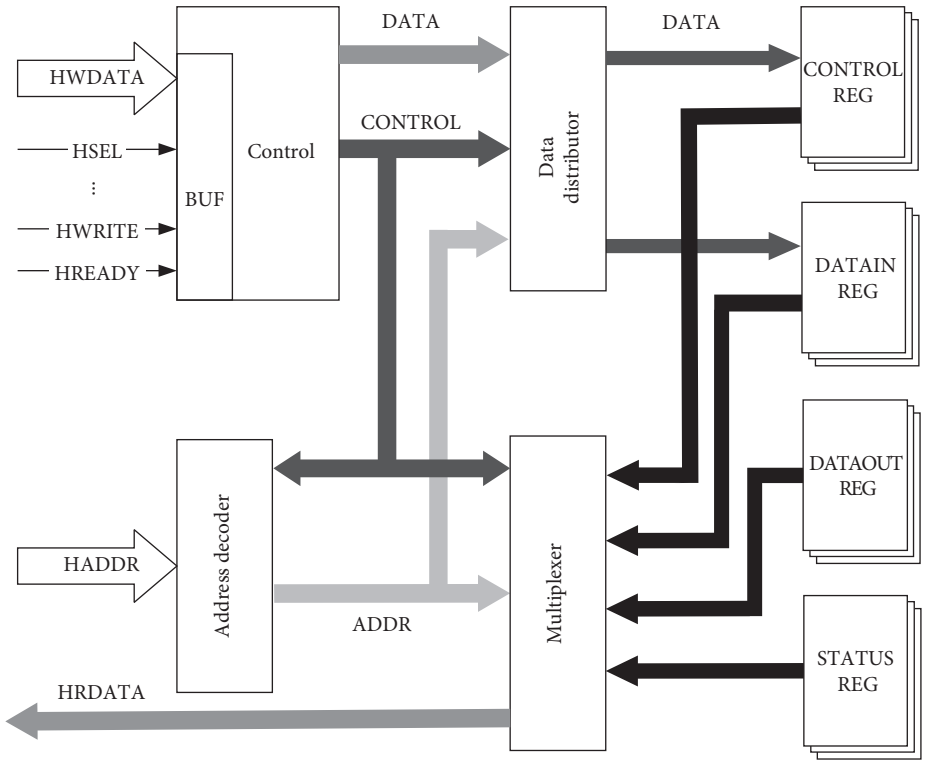


Figure 5: The block diagram of AICL structure.

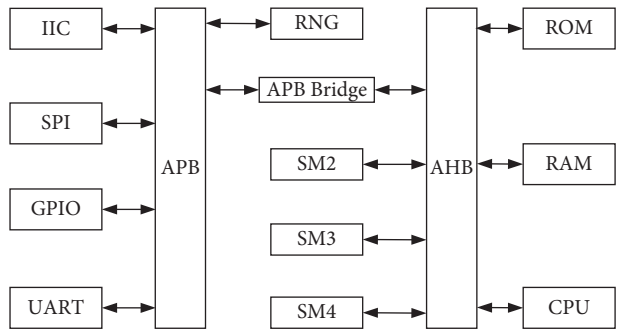


Figure 6: The architecture of the security SoC.

Finally, the 256 bit hash result can be obtained. The detailed codesign procedure can be found in [24].

The input signals of the SM3 accelerator include 512 bit input x, read control signal r, and write control signal w. The output signals include the 256 bit hash value y, the finish signal f, and the state signals. Therefore, this study designs a 32 bit status register, a 32 bit control register, eight 32 bit output registers, and 16 32 bit input registers. The control register of SM3 is responsible for the write/read control, enable/disable, and reset. Before writing to the module, bit5 is set to 0, bit4 and bit3 are set to 1, and then data are written to the input register. After writing the data, bit4 is set to 0, and the module starts the calculation. Once the operation is completed, the result can be read by setting bit4 to 1 and bit3 to 0. If there are several data blocks to be encrypted, bit4 and bit3 are set to 1, and the data are written into the input register until all operations are completed. The function of this control register is described in Table 2.

The status register mainly presents four working states and the exception of SM3. SM3 has four basic states, which are idle, writing, encrypting, and finish, respectively.

4.3. Integration of the SM4 Accelerator. SM4 accelerator contains the round key generation circuit part and the encryption/decryption circuit part. The 128 bits message could be encrypted with 32 clocks. The architecture of SM4 is depicted in [23]. For each group of plaintext M, the ciphertext will be generated after 32 round encryptions. The input signal of the SM4 module consists of 128 bit data input, 128 bit data output, status, and control signal. Therefore, it is necessary to set one 32 bit status register, one 32 bit control register, four 32 bit output registers, and four 32 bit input registers.

The function of the SM4 control register is described in Table 3. First, bit2 and bit3 are set to 1 before data encryption/decryption. Second, the 128 bit key or message is written to the input register. Finally, the corresponding data flag is set so that the module can identify the type of input data. It is worth noting that, since the round key is used in descending order for the encryption process, the message can be directly written to the input register after the key is loaded. For the decryption process, the data to be decrypted cannot be input until the round key has been generated.

The status register of SM4 is designed to present the current work mode, including the encryption mode, the decryption mode, whether the round key is generated, and whether the encryption/decryption process is completed.

In addition, the RNG module in the proposed SoC is an intellectual property depicted in [23]. The ring oscillators are employed to generate pseudo-random numbers or high-speed true random numbers. It consists of an online test module, a postprocessing module, and a high entropy true random source. The standard NIST SP800-22 test is carried out to verify the validity and stability of RNG.

4.4. Overall Steps of the Proposed Method. In this paper, a new method of easy-to-integrate IP design of the AHB slave bus interface for the security chip is proposed, which consists of two steps:

(i) First, the master and slave modules of the system should be determined before designing the interface IP, and the corresponding address space is allocated to these modules through the address decoder.

(ii) Second, the AHB interface control module is designed according to the modules in this security chip and AHB-Lite bus protocol.

(iii) Third, the required function registers for each slave module are designed and integrated with the AHB interface control module.

(iv) Fourth, CPU, memory (RAM and ROM), and security modules (SM2, SM3, and SM4) are integrated into SoC through AHB-SIP, and the functional registers of each security module are designed and configured. CPU is the master module of AHB-Lite bus, while other modules are the slave modules.

(v) Fifth, the RNG module and other low-speed modules are mounted on APB bus through APB bridge that is also the slave module of AHB-Lite bus.

(vi) Finally, the software calls the underlying operation of the hardware security module via CPU. The CPU reads and writes registers using the mode of bus addressing. The encryption/decryption operations are implemented by configuring the function registers of each security module through the CPU, thus realizing the data interaction between software and hardware.

Table 1: The function description of SM2 control register.

Bit	Operation	Type	Description
[3 : 0]	Mode control	R/W	0001 : MM mod N 1001 : MM mod P 0001 : MM mod N 1001 : MM mod P 0010 : MA mod N 1010 : MA mod P 0011 : MS mod N 1011 : MS mod P 0000 : MI mod N 1000 : MI 0101 : PA 0110 : MP 0111 : PM 1111: Idle
4	Enable control		0: Enable 1: Disable
5	Reset control		1: Reset
[31 : 6]	Reserve		

Table 2: The function description of the SM3 control register.

Bit	Operation	Type	Description
[2 : 0]	Reserve	R/W	
3	Data read/write		0: data read 1: data write
4	Enable control		0: enable 1: disable
5	Reset control		1: reset
[31 : 6]	Reserve		

Table 3: The function description of the SM4 control register.

Bit	Operation	Type	Description
[1 : 0]	Data flag	R/W	01: key 10: data to be encrypted 11: data to be decrypted 00: invalid input data
2	Data update		0: no update 1: update
3	Reset control		0: rset
[31 : 4]	Reserve		

5. Experiment Results and Analysis

We first analyze the reasons why we choose the AHB-Lite bus as the SoC bus is that it will be more efficient and save resources. With the rapid development of SoC systems, there are increasing demands for SoC buses. For the widely used SoC bus standards, the AMBA is a bus with complete functions and advanced protocols. In the AMBA, the AHB is an advanced high-performance bus, and AXI focuses on the advanced extensible interface. The bus latency in AHB is lower than that of AXI, and the AHB bus is used more frequently than AXI. As a subset of AHB, the AHB-Lite protocol supports only one master device, and there is no need for the arbiter and the request/authorization protocol. The goal of our design is to develop an efficient and low-power consumption information security chip that can be used for various intelligent hardware platforms and smart home devices. The structure of this chip requires only one master device, which mainly focuses on high-performance and low-power SoC design. To this end, we finally choose the AHB-Lite bus as the SoC system bus.

On the other hand, it is complicated to design a highly dedicated SoC, especially if the structure of the on-chip bus is based on unfamiliar or new protocols. It is difficult to accurately predict the architectural performance via an unfamiliar bus protocol, resulting in the risk of tape-out. Furthermore, the design period is prone to delay because of using a new protocol. The lack of easy-to-use bus interface IP makes the verification environment setup and test vector design more complex. Before communication, it is necessary to ensure that all slave modules have a unified AHB slave interface, or the communication cannot proceed. According to the practical requirements of modules, four functional registers are introduced, and a simple and efficient slave bus controller is designed in combination with the AHB-Lite protocol. Compared with the existing technology, our proposed interface IP and method are featured with the following advantages:

(1) Four functional registers for the communication between the slave and the master modules are introduced to realize the data transfer. Thus, it is unnecessary to know exactly about the AHB-Lite bus protocol.

(2) The AHB-Lite bus can realize the data transmission between the master and the slave modules via simply modifying the four types of function registers.

(3) By converting the nonstandard interface into the AHB slave interface via the AHB-SIP, the SoC system design can be achieved efficiently. The risk of chip tape-out can be reduced, the design period can be shortened, and the performance of SoC can be enhanced.

By utilizing the AHB-SIP and the integration method described in Section 4, we successfully integrate SM2, SM3, SM4 cryptographic accelerators, IIC, SPI, GPIO, UART interfaces, and RNG module into an SoC, accomplishing a low-power IoT security chip. The security chip is taped out with SMIC 110 nm technology process and QFN56 package technology. The system clock frequency is 36 MHz, and the voltage of core and IO are 1.2 V and 3.3 V, respectively. The area of this chip is $3.45 \times 3.45\,\text{mm}^2$. The gates and area of each module are listed in Table 4.

According to Table 4, SM2, SM3, and SM4 cryptographic accelerators have a total area of about $1.0\,\text{mm}^2$. It is noteworthy that the area of AHB-SIP is $0.072\,\text{mm}^2$, only occupying 6‰ of the chip. Since the 128K RAM is applied to the SoC, it occupies about 1/3 of the chip area.

The ASIC layout is shown in Figure 7. The two RAMs are distributed on the right of the layout. 8 KB ROM is in the upper left of the layout, and the CPU is in the lower left of the layout. The rest are SM2, SM3, SM4 cryptographic accelerators, RNG, and other modules. Since the AHB-SIP is

Table 4: The gates and area of each module.

Module	Gates (K)	Area (mm^2)
AHB-SPI	14.199	0.072
SM2/3/4	199.075	1.013
RNG	57.003	0.29
CPU	88.371	0.45
ROM	70.343	0.358
RAM	763.055	3.886
Others	1202.355	5.834
Total chip	2337.398	11.903

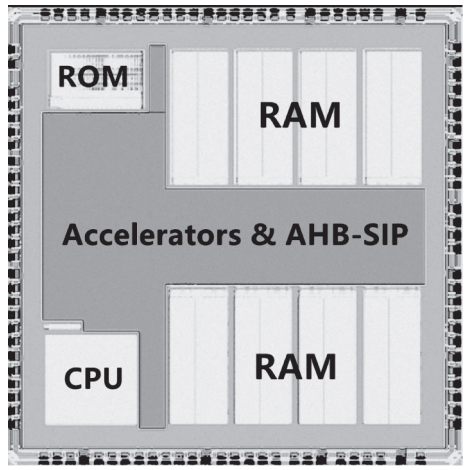


Figure 7: The architecture of the security SoC.

scattered in the layout, the size of AHB-SIP cannot be observed directly from the layout.

Also, to compare with using AXI bus, we experimented to evaluate the total power of crypto accelerator with different SoC bus interfaces. The experiment was implemented on the Xilinx FPGA of Virtex 6 architecture under the frequency of 100 MHz. The ISE Design Suite of Xilinx provides a power simulator XPower Analyzer, which can analyze the power of programable logic devices. By taking the cases of SM2, SM3, and SM4 modules, we found that the power consumption of these encryption modules combined with AHB-SIP was lower than that combined with the AXI interface, which decreased by 61.0%, 49.7%, 48.0%, respectively, as shown in Figure 8. This demonstrates that the fewer hardware resources we used, the lower power is consumed.

For ASIC design, the proposed method is compared with other state-of-the-art schemes to test the performance of the cryptographic accelerators and AHB-SIP. Table 5 lists different implementation methods of the cryptographic accelerators. The results indicate that the proposed method combining the cryptographic accelerators and AHB-SIP provides low power consumption and good performance for the three sorts of cryptographic algorithms.

It can be observed that it is infeasible to compare the results, as technology library, methodologies, and application areas are different. According to Table 5, for the SM2 accelerator, the throughput of PM operation is higher than that in [25], indicating the times of point multiplication per unit time are more than that of [25]. Except for [25], the power consumption of this design is the lowest. Since 40 nm process technology is adopted in [25], no equivalent comparison can be made. In other architectures, the performance of [26] is better than ours, but its area and power consumption are greater. Although the speed of PM operation in [27] is the highest, the area is also the largest. Besides, the number of logic gates is 11.76 times that of the design architecture in our work and approximately triple that of other designs. Considering the cost of developing IoT chips, high power consumption and a large area are inappropriate for IoT chips. For the SM3 accelerator, several implementation methods of the hash algorithm are listed in Table 5. As Table 5 reveals, the implementation method proposed in [24] has high throughput, small area, and high power consumption. Although 886 gates are required in the SHA-3 design in [25], the power consumption and throughput are inferior to our design. At the normalized frequency, the throughput of our design is 13.8 times higher compared with the design in [28]. Compared with the AES architecture implemented in [25], the power consumption of the SM4 accelerator is close to ours at the same frequency, while the efficiency is much higher than [25]. Compared with the architecture implementation in [14], the saved gates with the proposed architecture are approximately 197.5 K. It is evident that, although the throughput in [29] is the highest, it has higher power consumption and a larger area than other architectures. Therefore, it does not apply to IoT security chips.

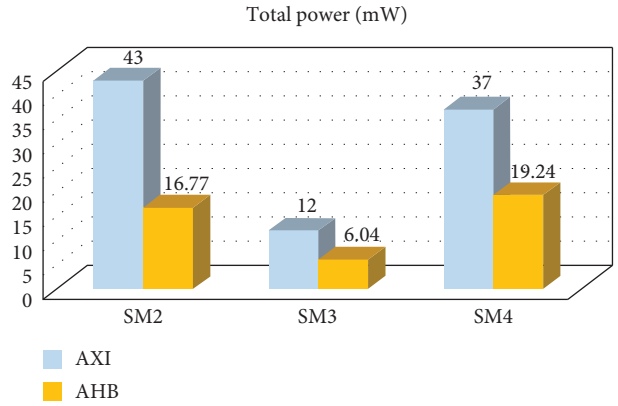


Figure 8: The power comparison among different accelerators combined with different SoC buses of security chip.

Table 5: Performance comparison with other architecture.

Design	library	Frequency (MHz)	Area (gate)	Power (mW)	Throughput
SM2 (our)	0.11-μm	36	56K	3.24	27.64 Kbps
ECC233 [24]	0.04-μm	28.8	-	1.13	8.59 Kbps
SM2 [25]	0.13-μm	214	208K	40.28	1.20 Mbps
SM2 [26]	0.13-μm	163.7	659K	—	12.57 Mbps
SM3(our)	0.11-μm	36	18K	0.18	245.7 Mbps
SM3 [23]	0.13-μm	36	6036	1.24	263 Mbps
SHA-3 [24]	0.04-μm	28.8	886	4.87	14 Mbps
SHA-256 [27]	0.13-μm	102	9036	3.06	47 Mbps
SM4(our)	0.11-μm	36	124K	3.39	115.2 Mbps
AES [24]	0.04-μm	28.8	—	2.8	5.08 Mbps
AES [14]	0.13-μm	200	321.5 K	325	564 Mbps
AES [28]	0.04-μm	1000	9028K	6.17 K	128 Gbps

Table 6: The test result of the RNG.

Statistical Test	Full entropy	High speed
Frequency	990/1000	992/1000
Block frequency	992/1000	995/1000
Runs	984/1000	989/1000
Longest run	988/1000	985/1000
Rank	989/1000	990/1000
FFT	984/1000	989/1000
Cumulative sums	Pass	Pass
Nonoverlapping template	Pass	Pass
Overlapping template	987/1000	987/1000
Universal	992/1000	990/1000
Approximate entropy	990/1000	991/1000
Random excursions	Pass	Pass
Random excursions variant	Pass	Pass
Serial	Pass	Pass
Linear complexity	Pass	Pass

Our designed RNG module in the SoC is tested based on the standard NIST SP800-22. The random numbers to be tested for each set are divided into 1000 groups, with each group containing 1M bit random numbers. According to the NIST standard, if at least 980 of the 1000 random numbers pass a statistical test, it can be considered to pass. We tested a total of five sets. Since the standard of NIST's nonoverlap template matching test is quite strict, if the pass rate is not very poor, it is usually negligible. Thus, the five sets of random numbers are verified to be of quite high quality. Table 6 presents the results of our test.

In conclusion, compared with the above baseline designs, we obtain the following results:

(i) Since the bus latency of AHB is lower than that of AXI and the different structures of the security module design, our proposed method is more efficient than others when using the AHB-SIP.

(ii) We use fewer hardware resources for designing the AHB-SIP, and the area of the chip is smaller. The total power consumption is only 8.4 mW @36 MHz, which is very suitable for IoT devices.

(iii) The results indicate that the balance between the throughput, area, and power consumption of our proposed SoC with AHB-SIP at the normalized frequency is excellent.

6. Conclusion and Future Work

This study proposed a design of AHB-SIP in the field of IoT security, which can easily integrate the security units into an SoC and transform a cryptographic accelerator with a nonstandard interface into an accelerator with the AHB slave interface. Besides, the SM2, SM3, SM4, and RNG security modules are configured by software through AHB-SIP to improve the design efficiency of SoC. Finally, a low-power IoT security chip is realized by using 110 nm process technology. The implementation and test results indicate that the area of AHB-SIP is quite small, the power consumption is lower than AXI-based architecture, and the performance of accelerators is ideal for IoT applications. In the future, it is necessary to study the construction and optimization of AHB-SIP to enhance performance and flexibility.

Data Availability

The Verilog 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 work was supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010145001 and in part by the Science and Technology Planning Project of Guangdong Province of China under Grant 2019B010140002.

References

[1] Y. H. Hwang, "IoT security & privacy: threats and challenges," in *Proceedings of the 1st ACM Workshop on IoT Privacy, Trust, and Security-IoTPTS '15*, vol. 1, ACM Press, Singapore, April 2015.

[2] W. Zhou, "The effect of iot new features on security and privacy: new threats, existing solutions, and challenges yet to be solved," *IEEE Internet of Things Journal*, vol. 6, pp. 1606–1616, 2018.

[3] M. Frustaci, P. Pace, G. Aloi, and G. Fortino, "Evaluating critical security issues of the IoT world: present and future challenges," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2483–2495, 2018.

[4] J. E. Siegel, K. Sumeet, and E. S. Sanjay, "The future internet of things: secure, efficient, and model-based," *IEEE Internet of Things Journal*, vol. 54, pp. 2386–2398, 2017.

[5] G. Ma and He Hu, "Design and implementation of an advanced DMA controller on AMBA-based SoC," in *Proceedings of the 2009 IEEE 8th International Conference on ASIC, IEEE*, pp. 419–422, Changsha, China, October 2009.

[6] A. K. Swain and K. Mahapatra, "Design and verification of WISHBONE bus interface for system-on-chip integration," in *Proceedings of the 2010 Annual IEEE India Conference (INDICON), IEEE*, pp. 1–4, Kolkata, India, December 2010.

[7] R. Hofmann and B. Drerup, "Next generation CoreConnect/spl trade/processor local bus architecture," in *Proceedings of the 15th Annual IEEE International ASIC/SOC Conference*, pp. 221–225, IEEE, Rochester, NY, USA, December 2002.

[8] Q. Zhou, S. Yu-Kun, D.-L. Zhang, and G.-M. Du, A design of multi-core system based on A valon bus," in *Proceedings of the 2011 International Conference on Computer Science and Network Technology*, pp. 1456–1459, IEEE, Harbin, China, December 2011.

[9] C.-Y. Chang, Y.-J. Chang, J.-C. Yeh, S. Y. Lin, and J.-L. Ma, "Design of on-chip bus with OCP interface," in *Proceedings of the 2010 International Symposium on VLSI Design, Automation and Test*, pp. 211–214, IEEE, Hsin Chu, Taiwan, April 2010.

[10] S.-F. Hsiao, C.-G. Lin, P.-H. Wu, and C.-S. Wen, "Asynchronous AHB bus interface designs in a multiple-clock-domain graphics system," in *Proceedings of the 2012 IEEE Asia Pacific Conference on Circuits and Systems*, pp. 408–411, IEEE, Kaohsiung, Taiwan, December 2012.

[11] Z.-Y. Li, S.-B. Liu, and Y. Feng, "Design of an interface between the IDE controller and the AHB bus based on FPGA," *Computer Engineering & Science*, vol. 2, 2017.

[12] A. K. Singh, A. Shrivastava, and G. Tomar, "Design and implementation of high performance AHB reconfigurable arbiter for onchip bus architecture," in *Proceedings of the 2011 International Conference on Communication Systems and Network Technologies*, pp. 455–459, IEEE, Katra, India, June 2011.

[13] S. Xuan, J. Han, Z. Yu, Y. Ren, and X. Zeng, "A configurable SoC design for information security," in *Proceedings of the 2015 IEEE 11th International Conference on ASIC (ASICON)*, pp. 1–4, IEEE, Chengdu, China, November 2015.

[14] W. Huang, J. Han, S. Wang, and X. Zeng, "The design and implement of a mobile security SoC," in *Proceedings of the 2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, pp. 96–98, IEEE, Shanghai, China, November 2010.

[15] Y. Hu, L. Wu, A. Wang, and B. Wang, "Hardware design and implementation of SM3 hash algorithm for financial IC card," in *Proceedings of the 2014 Tenth International Conference on Computational Intelligence and Security*, pp. 514–518, IEEE, Kunming, China, November 2014.

[16] A. Hafsa, N. Alimi, A. Sghaier, M. Zeghid, and M. Machhout, "A hardware/software Co-designed AES-ECC cryptosystem," in *Proceedings of the 2017 International Conference on Advanced Systems and Electric Technologies (IC_ASET)*, pp. 50–54, Hammamet, Tunisia, January 2017.

[17] "ARM AMBA 5 AHB protocol specification AHB5," *AHB-Lite*, vol. 86, 2001.
[18] *AMBA Specification (Rev 2.0)*, AMBA, ARM Limited, UK, 1999.
[19] S. Kante, H. K. Kakarla, and A. Yadlapati, "Design and verification of AMBA AHB-lite protocol using verilog HDL," *International Journal of Engineering and Technology*, vol. 8, pp. 734–741, 2016.
[20] V. S. Miller, "Use of elliptic curves in cryptography," in *Advances in Cryptology*, H. C. Williams, Ed., vol. 218pp. 417–426, 1986.
[21] N. Koblitz, "Elliptic curve cryptosystems," *Mathematics of Computation*, vol. 48, no. 177, p. 203, 1987.
[22] C. H. Wang, C. Y. Lo, and M. S. Lee, "A network security processor design based on an integrated SOC design and test platform," in *Proceedings of the 43rd Annual Design Automation Conference*, pp. 490–495, New York, NY, USA, July 2006.
[23] X. Zheng, C. Xu, X. Hu, Y. Zhang, and X. Xiong, "The software/hardware Co-design and implementation of SM2/3/4 encryption/decryption and digital signature system," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 10, pp. 2055–2066, 2020.
[24] X. Zheng, X. Hu, J. Zhang, J. Yang, S. Cai, and X. Xiong, "An efficient and low-power design of the SM3 hash algorithm for IoT," *Electronics*, vol. 8, no. 9, p. 1033, 2019.
[25] Y. Zhang, L. Xu, Q. Dong, J. Wang, D. Blaauw, and D. Sylvester, "Recryptor: a reconfigurable cryptographic cortex-M0 processor with in-memory and near-memory computing for IoT security," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 995–1005, 2018.
[26] D. Zhang and G. Bai, "Ultra high-performance ASIC implementation of SM2 with SPA resistance," in *Information and Communications Security*, S. Qing, E. Okamoto, K. Kim, and D. Liu, Eds., vol. 9543, pp. 212–219, 2016.
[27] Z. Zhao and G. Bai, "Ultra high-speed SM2 ASIC implementation," in *Proceedings of the 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 182–188, IEEE, Beijing, China, September 2014.
[28] X. Cao, L. Lu, and M. O'Neill, "A compact SHA-256 architecture for RFID tags," in *Proceedings of the 22nd IET Irish Signals and Systems Conference*, ISSC, Ireland City, Dublin, June 2011.
[29] G. Sayilar and D. Chiou, "Cryptoraptor: high throughput reconfigurable cryptographic processor," in *Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, vol. 155, November 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 2636302, 12 pages
https://doi.org/10.1155/2021/2636302

Research Article

Compound Fault Diagnosis of Rolling Bearing Based on ALIF-KELM

Jie Ma, Shitong Liang, Zhengyu Du, and Ming Chen

School of Mechatronics Engineering, Beijing Information Science and Technology University, Beijing, China

Correspondence should be addressed to Jie Ma; mjbeijing@163.com

Received 19 August 2021; Revised 25 September 2021; Accepted 12 October 2021; Published 26 October 2021

Academic Editor: Yong Chen

Aiming at the shortcomings of difficult classification of rolling bearing compound faults and low recognition accuracy, a composite fault diagnosis method of rolling bearing combined with ALIF and KELM is proposed. First, the basic concepts of ALIF and KELM are introduced, and then ALIF is used to decompose the sample data of vibration signals of different bearing states so that each sample can get several IMFs, select the top K IMFs containing the main fault information from each sample, calculate the energy feature and sample entropy of each IMF, and construct a fault feature vector with a dimension of 2K. Finally, the feature vectors of the training set and the test set are input into the KELM model for fault classification. Experimental results show that, compared with EMD-KELM model, ALIF-ELM model, ALIF-BP model, and IFD-KELM model, the rolling bearing composite fault diagnosis method based on the ALIF-KELM model has higher classification accuracy.

1. Introduction

Rolling bearings are one of the basic components and play an important role in various types of industrial equipment. Rolling bearings have been widely used in many engineering fields. However, the actual working environments of rolling bearings are very harsh. After an extended period of operation, these components are prone to failure. In addition to a single failure, the failure types can also easily present as composite failure formed due to simultaneous occurrences of multiple types of failures [1]. Statistical analysis [2] indicates that approximately 30% of all rotating machinery equipment failures are caused by failure of rolling bearings. Consequently, effective monitoring of the integrity health status of rolling bearings and timely elimination of hidden issues play an important role in ensuring safe and reliable equipment operation, reduction in economic and capital losses, and avoiding accidents.

In view of the above situation, most of the methods currently proposed by researchers are based on vibration signal processing composite fault diagnosis technology for rolling bearings, in which the signal decomposition method is one of the effective methods for processing vibration signals. In 1998, Huang et al. [3, 4] proposed an empirical mode decomposition (EMD) algorithm. Ma Xinna and others combined EMD with an adaptive notch filter to realize the adaptive separation and diagnosis of rolling bearing composite faults. However, due to the lack of EMD's strict mathematical theoretical derivation, singular points in the signal easily lead to modal aliasing occurrences. Cubic spline interpolation has either underfitting or overfitting and is unstable under the noise interferences. To effectively resolves rolling bearing problems, researchers have proposed many adaptive mode decomposition methods inspired by the idea of EMD, including local mean decomposition (LMD), empirical wavelet transform (EWT), and variational modal decomposition (VMD) [5–7]. Huang et al. [8] extended the local mean decomposition to a complex local mean decomposition and were successful in applying it to the composite fault diagnosis of rolling bearings. Zhu et al. [9] proposed a parameterized local eigenscale decomposition method for the discontinuity of the first derivative of the local eigenscale decomposition method, applied it to the composite fault simulation signal and the bearing experimental signal, and verified the method's performance. Effectiveness and superiority of the latter method are

demonstrated by a comparative analysis. Hu et al. [10] optimized several important parameters in the variational modal decomposition to improve the decomposition performances. At the same time, they also used the 1.5-dimensional spectrum to suppress noise and enhance the impact signal, combining the two to achieve effective separation of composite faults in the rolling bearings. In order to improve the stability and convergence of the mean function of the upper and lower envelopes under disturbances, Lin et al. [11] proposed an iterative filter (IF) algorithm, which follows the same algorithm framework as EMD and uses low-pass filtering to obtain the upper and lower mean functions of the signal envelope. In 2016, Cicone et al. [12] used the basic solution system of Fokker–Planck (FP) differential equations as the filter function to extend the IF algorithm; they proposed the Adaptive Local Iterative Filter (ALIF) algorithm. ALIF can effectively analyse and process nonlinear and nonstationary signals. At present, the algorithm has been increasingly applied to the field of rotating machinery fault diagnosis, Chen et al. [13] combined ALIF and energy operator demodulation methods to effectively diagnose the fault characteristic frequencies of rolling bearings. Zhang et al. [14] proposed a method based on ALIF and high-order energy operator demodulation and successfully identified weak fault components during the early faults stages of rolling bearings; compared with the low-order energy operator demodulation method, this approach proves to be a superior method.

In recent years, machine learning technology has allowed for better success, through applying intelligent fault diagnosis algorithms [15–18]. Globally, scholars have continued to research intelligent recognition algorithms based on the BP neural network; these applications have been widely applied to the rolling bearing fault diagnosis achieving relatively sound academic resolutions [19–21]. The BP neural network requires performing iterative calculations during the learning process; sometimes, it falls into a local minimum, causing the algorithm to become time intensive, and the generalization ability of the network is very limited [22].

To address the above problems, Huang et al. [23] proposed an extreme learning machine (ELM) based on the single-hidden layer feedforward network (SLFN). The algorithm relies on its own performance and has gradually attracted the attention of scholars in diverse fields, including significance for the development of intelligent diagnosis technology for rolling bearing faults. For the nonstationary characteristics of bearing vibration signals, scholars, locally and abroad, usually use various nonstationary signal processing and analysis methods combined with the ELM algorithm to conduct intelligent diagnosis research on rolling bearing faults. Xu and Ma [24] used a combination of empirical wavelet transform and ELM to apply to the study of intelligent diagnosis of rolling bearing faults and provided bearing experimental data to prove the feasibility of this method. When the intelligent diagnosis model remains unchanged, the construction of the fault feature vector will have an important influence on the diagnosis effect of the intelligent fault diagnosis. KELM is an improved algorithm proposed by Huang et al. [25, 26] and is based on ELM. First, the original algorithm is optimized, and then the kernel function is used to replace the activation function of the hidden layer to make the model stable and universal. The KELM algorithm has improved generalization ability and is more suitable for solving multiclassification problems. This paper proposes a composite fault diagnosis method for rolling bearings that combines both Adaptive Local Iterative Filter (ALIF) and KELM approaches.

2. Adaptive Local Iterative Filter Algorithm

Adaptive Local Iterative Filter (ALIF) is a new type of adaptive mode decomposition method; improvements are due to the iterative filtering algorithm (IF). ALIF mainly constructs a filter function with adaptive characteristics by applying the basic solution system of Fokker–Planck differential equations. Therefore, it is also very necessary to introduce the IF algorithm before introducing the principle of the ALIF algorithm.

2.1. Iterative Filter. IF is similar to the EMD algorithm; it iteratively filters out each eigenmode function (IMF) component. This method convolves the filter functions with the signal to be decomposed to obtain the sliding operator; this process replaces the process of fitting the original data to obtain the mean value of the envelope in the EMD algorithm. IF mainly includes two processes: inner loop and outer loop.

Knowing the preprocessed signal $X(t)$ and the filter function $f(t)$, the sliding operator $\Gamma(X(t))$ is obtained by calculating the convolution of $X(t)$ and $f(t)$:

$$\Gamma(X(t)) = \int_{-h(z)}^{h(z)} X(t+\tau) f(t) \mathrm{d}\tau, \tag{1}$$

where $f(t)$ is the fixed low-pass filter function; $h(z)$ is the filter interval; its calculation expression is as follows:

$$h(z) = 2\left[\frac{N\lambda}{a}\right], \tag{2}$$

where N is the signal length of $X(t)$; λ is the set value; a is the number of extreme points of $X(t)$.

Then calculate the fluctuation operator $K(X(t))$ by preprocessing the difference between the signal $X(t)$ and the sliding operator $\Gamma(X(t))$:

$$K(X(t)) = X(t) - \Gamma(X(t)). \tag{3}$$

Finally, it judges whether the volatility operator $K(X(t))$ meets the conditions of the IMF component, and only the volatility operator that meets the set conditions can be extracted as the IMF component. If not, the volatility operator needs to be screened further, and the specific process is as follows:

(1) Calculate the filtering interval 1 of the preprocessed signal according to formula (2).

(2) Solve the sliding operator $\Gamma(X(t))$ according to formula (1).

(3) Calculate the volatility operator $K(X(t))$ according to formula (3), and the expression of the volatility operator in the screening process is as follows:

$$\mathrm{K}_n(X(t)) = X_n(t) - \Gamma_n(X(t)) = X_{(n+1)}(t). \tag{4}$$

Let

$$\mathrm{IMF}(t) = \lim_{n\longrightarrow\infty} \mathrm{K}_n(X(t)). \tag{5}$$

When $\mathrm{IMF}(t)$ can meet the IMF component conditions, complete the extraction of IMF components; otherwise, continue to repeat steps 1 to 4 until the conditions are met before stopping the screening. However, in actual situations, it is impossible for n to approach infinity, so the screening termination conditions for IMF components can be artificially set as follows:

$$\sigma = \frac{\left\|K_{i,n} - K_{i,n-1}\right\|_2}{\left\|K_{i,n-1}\right\|_2}. \tag{6}$$

That is, when σ is less than a specified threshold, the screening is stopped and $\mathrm{IMF}(t)$ is the filtered IMF component.

The above-mentioned is the inner circulation process, and its main purpose is to extract the qualified IMF components, while the function of the outer circulation process is to stop the inner circulation process. First, the margin after all the effective IMF components of the preprocessed signal are successfully extracted is defined as the residual signal, denoted as $c(t)$:

$$c(t) = X(t) - \mathrm{IMF}(t). \tag{7}$$

When the residual signal $c(t)$ has obvious trend characteristics, that is, there is only one extreme point at most, the entire iterative filtering process is halted. Otherwise, it needs to be used as a fresh preprocessing signal to continue to extract qualified IMF components.

2.2. Adaptive Local Iterative Filter. In IF, in order to reduce the negative impact of noise on it, filter functions are generally set in advance, but some complex signals will lack adaptability when applying IF algorithms and may also cause component waveform distortion. In order to analyse both nonlinear and nonstationary signals more effectively and overcome the shortcomings of the IF algorithm, Cicone et al. were inspired by the diffusion process of partial differential equations and used the solution of the Fokker–Planck equation to construct a filter function; this enabled the filter to be tightly supported in the time domain; its length can be flexibly changed and adaptability is enhanced. Moreover, it can also avoid false components in the iterative filtering process [14]. This allows ALIF to effectively suppress noise sensitivity and modal aliasing in the IF algorithm.

For interval (a,b), there are two differentiable functions $p(x)$ and $q(x)$, and the following two conditions are satisfied:

(1) $p(a) = p(b) = 0$, and $p(x) > 0$ holds for $x \in (a,b)$

(2) $q(a) < 0 < q(b)$

The Fokker–Planck equation is as follows:

$$\frac{\partial}{\partial t} g(x,t) = -\alpha \frac{\partial}{\partial x}\left[p(x,t) g(x,t)\right] + \beta \frac{\partial^2}{\partial x^2} \cdot \left[q^2(x,t) g(x,t)\right], \quad \alpha, \beta > 0. \tag{8}$$

In order to simplify the above formula, convert to the expression of the differential equation:

$$G_t = -\alpha\left(p(x) g\right)_x + \beta\left(q^2(x) g\right)_{xx}, \tag{9}$$

where α and β are called steady-state coefficients and $\alpha, \beta \in (0,1)$.

$(p(x)g)_x$ in equation (9) will have the effect of aggregation so that the solution $g(x)$ of the equation will start from the two endpoints of the interval $[a,b]$ towards the center point. At the same time, $(q^2(x)g)_{xx}$ will produce the effect of diffusion, which causes the solution $g(x)$ of the equation to diffusely move from the center of interval $[a,b]$ to the two endpoints. When the two effects are balanced,

$$-\alpha\left(p(x) g\right)_x + \beta\left(q^2(x) g\right)_{xx} = 0. \tag{10}$$

At this time, the differential equation has a nonzero solution and meets the following conditions:

$$\begin{aligned} &\forall x \in (a,b), \quad g(x) > 0, \\ &\forall x \notin (a,b), \quad g(x) = 0. \end{aligned} \tag{11}$$

The solution $g(x)$ in the Fokker–Planck equation is the filter function $f(t)$ used in iterative filtering. For different intervals $[a,b]$, the solution obtained by the filter function $f(t)$ will also be different, and the function expression will also differ, allowing the ALIF algorithm to solve for filter function adaptively.

3. Kernel-Based Extreme Learning Machine Algorithm

3.1. Extreme Learning Machine. The ELM network structure is shown in Figure 1. The network structure includes three layers: hidden layer, input layer, and output layer [27]. The ELM intelligent learning model needs to provide the number of hidden layer nodes and the type of activation function during the entire learning process, while the input weights and hidden layer thresholds are randomly generated and remain unchanged. Finally, the least square method can be used to solve the output weight under the premise of ensuring that the training error is minimized.

Assuming that there are N existing data samples (x_i, t_i) that are not specific, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^N$ and $t_i = [t_{i1}, t_{i2}, \ldots, t_{in}]^T \in R^N$, the corresponding output expression of SLFN with L hidden layer nodes is as follows:

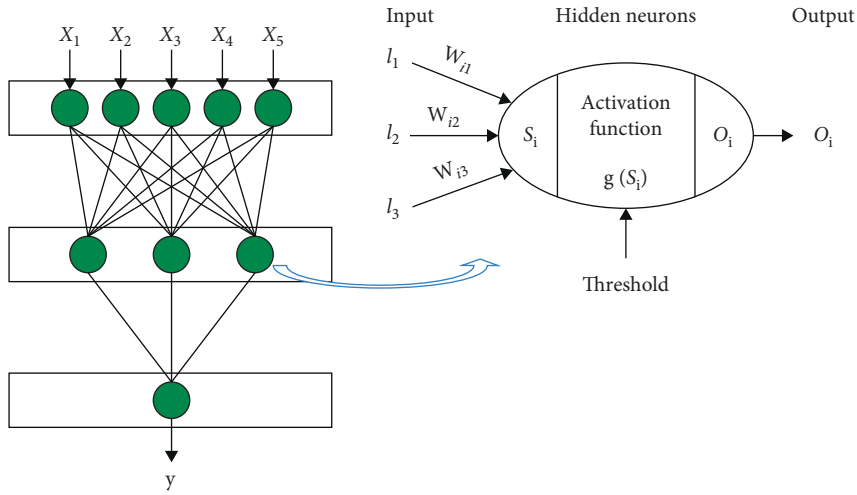


Figure 1: ELM network structure diagram.

$$y = \sum_{i=1}^{L} \beta_i g\left(\omega_i \cdot x_j + b_i\right) = t_j, \quad j = 1, 2, \ldots, N, \tag{12}$$

where β_i represents the connection weight of the output layer and the hidden layer; ω_i represents the connection weight of the input layer and the hidden layer; x_j is the input vector, which also represents all the feature vectors of the j th sample; b_i represents the hidden layer threshold; $g(x)$ represents the hidden layer containing layer activation function; t_j is the output vector, which also represents the class label of the j th sample.

Assuming that the activation function $g(x)$ is infinitely differentiable, then the ultimate goal of ELM learning is to minimize the output error; that is, infinity approaches 0, which can be expressed as follows:

$$\sum_{j=1}^{N} t_j - T_j = 0. \tag{13}$$

Then there are β_i, ω_i, and b_i that make the following formula true:

$$\sum_{i=1}^{L} \beta_i g\left(\omega_i \cdot x_j + b_i\right) - T_j = 0, \quad j = 1, 2, \ldots, N. \tag{14}$$

The abbreviated formula (14) is expressed as a matrix form as follows:

$$H\beta = T, \tag{15}$$

where H is the hidden layer output matrix; T is the expected output matrix; they are represented as follows:

$$H(\omega, x, b) = \begin{bmatrix} g(\omega_1 \cdot x_1 + b_1) & \cdots & g(\omega_L \cdot x_1 + b_L) \\ g(\omega_1 \cdot x_2 + b_1) & \cdots & g(\omega_L \cdot x_2 + b_L) \\ \vdots & \ddots & \vdots \\ g(\omega_1 \cdot x_N + b_1) & \cdots & g(\omega_L \cdot x_N + b_L) \end{bmatrix}_{N \times L},$$
$$\beta = [\beta_1, \beta_2, \ldots, \beta_L]^T_{m \times L}, \tag{16}$$
$$T = [t_1, t_2, \ldots, t_N]^T_{m \times N}.$$

Since the input parameters of the ELM algorithm are randomly generated and remain unchanged, there is no need to adjust during the entire training and learning process. The connection weight 1 of the output layer and the hidden layer under the minimum error can be solved by the following formula:

$$\widehat{\beta} = H^{+}T, \tag{17}$$

where H^{+} represents the Moore–Penrose generalized inverse matrix of H.

3.2. Kernel-Based Extreme Learning Machine. The kernel extreme learning machine is based on the single-hidden layer feedforward neural network extreme learning machine. By introducing the kernel function mapping and regularization theory to optimize the model network, it can improve the accuracy and generalization ability while reducing the complexity and randomness of the network.

The extreme learning machine can be expressed by the following formula through mathematical expression:

$$\begin{aligned} &\text{min: } L_p = \frac{1}{2}\beta_i^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2, \\ &\text{s.t.: } \quad h(x_i)\beta_i = t_i - \xi_i, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{18}$$

where C represents the penalty coefficient; ξ is the training error; $h(x)$ is the output row vector of the hidden layer.

Solving for the above optimization problem, it can be concluded that the improved output function of ELM is as follows:

$$f(x_i) = h(x_i)H^T\left(\frac{1}{C} + HH^T\right)^{-1}T, \tag{19}$$

where H is expressed as follows:

$$H = [h(x_1), \ldots, h(x_N)]^T_{N \times L}. \tag{20}$$

Regarding $h(x_i)$ as the nonlinear mapping of each sample, HH^T represents the inner product form of $h(x_i)$, using the kernel function theory to define the kernel matrix Ω_{ELM} to replace HH^T so as to overcome the fluctuation of the final result of the ELM algorithm due to randomly generated inputs. The kernel matrix definition of KELM is as follows:

$$\Omega_{\text{ELM}} = HH^T, \tag{21}$$

$$\Omega_{\text{ELM}_{ij}} = h(x_i) \cdot h(x_j) = K(x_i, x_j). \tag{22}$$

After finishing formulas (20) ~ (22) and substituting them into formula (19), the new output function e of KELM is obtained as follows:

$$f(x) = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix} \left(\frac{1}{C} + \Omega_{\text{ELM}}\right)^{-1} T, \tag{23}$$

where T represents the label of the data set; Ω_{ELM} is a symmetric matrix with N rows and N columns; $K(x_i, x_j)$ is the kernel function quoted. This paper uses the Gaussian radial basis kernel function; Ω_{ELM} and $K(x_i, x_j)$ are expressed in the following specific forms:

$$\Omega_{\text{ELM}} = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{bmatrix}_{N \times N}, \tag{24}$$

$$K(x_i, x_j) = \exp\left(-\lambda x_i - x_j^2\right), \quad \lambda > 0. \tag{25}$$

where λ represents the nuclear coefficient.

4. Classification Algorithm Based on the Combination of ALIF and KELM

According to the previous introduction and analysis, ALIF can effectively decompose nonlinear and nonstationary vibration signals and further analyse the IMF components obtained after decomposition. ALIF can extract the local characteristics of the fault signal. Compared to traditional neural network algorithms, KELM has a strong generalization learning ability and at the same time has high efficiency and stability. Therefore, this paper combines the two and proposes a diagnostic method for rolling bearing composite faults based on ALIF and KELM. The specific steps are as follows:

(1) The ALIF decomposition is applied to the sample data of the vibration signal of different bearing states, and each sample can get several IMF components and a residual component

(2) Select the first K IMF components containing the main fault information from each sample, calculate the energy characteristics and sample entropy of each component, and fuse them to construct a fault feature vector with a dimension of $2K$

(3) Divide all samples into training samples and test samples in a certain proportion

(4) Select the specific form of the kernel function, and determine the two parameters of the kernel coefficient λ and the penalty coefficient C to complete the initialization of the KELM intelligent diagnosis model

(5) The fault feature vector of each sample is normalized to improve the comparability between data

(6) Let the KELM intelligent diagnosis model continuously learn through the fault feature vector set of the training sample, then test the fault feature vector set of the test sample, and finally, identify different bearing fault types and output the results

The corresponding flowchart of the steps outlined is shown in Figure 2.

5. Classification Experiment

5.1. Experimental Data Processing. The data set analysis and verification in this study were generated from the Xi'an Jiaotong University rolling bearing accelerated life test. The experimental setup is shown in Figure 3 [28]. The sampling frequency is set to 25.6 kHz, the sampling interval is set to 1 min, and each sampling time is 1.28 s, so the number of sampling points for each sample in the data set is 32768. The vibration signal collected in the experiment is all the data of the rolling bearing from normal to failure, including a total of 15 data sets under 3 working conditions. In the following sections, four data sets will be used to analyse and verify the method proposed in this paper. The data description is shown in Table 1. The bearing data used in the subsequent analysis in this paper are based on the failure data intercepted during the whole life cycle.

This paper uses the experimental data set introduced in Table 1 and obtains 102400 sample points of rolling bearing outer ring fault, cage fault, inner ring and outer ring composite fault, inner ring fault, and normal state data from it.

The data of each state of the rolling bearing is divided into 50 samples, a total of 250 samples are obtained from the five states, and each sample contains 2048 sampling points. The procedure is to take one sample from each of the five bearing states and generate their time-domain waveforms as shown in Figure 4.

5.2. Fault Feature Analysis. It is necessary to construct fault feature vectors in advance before KELM performs intelligent fault diagnosis. Selecting appropriate features will help improve the accuracy of fault intelligent diagnosis. Therefore, before proceeding with the method verification, a brief analysis of the fault characteristics used in this paper is given.

Select one sample data in each of the five states of the rolling bearing, set the same parameters for all five samples, and then apply ALIF decomposition to obtain five IMF components and one residual component. Calculate the energy characteristics and sample entropy of the first 4 IMF

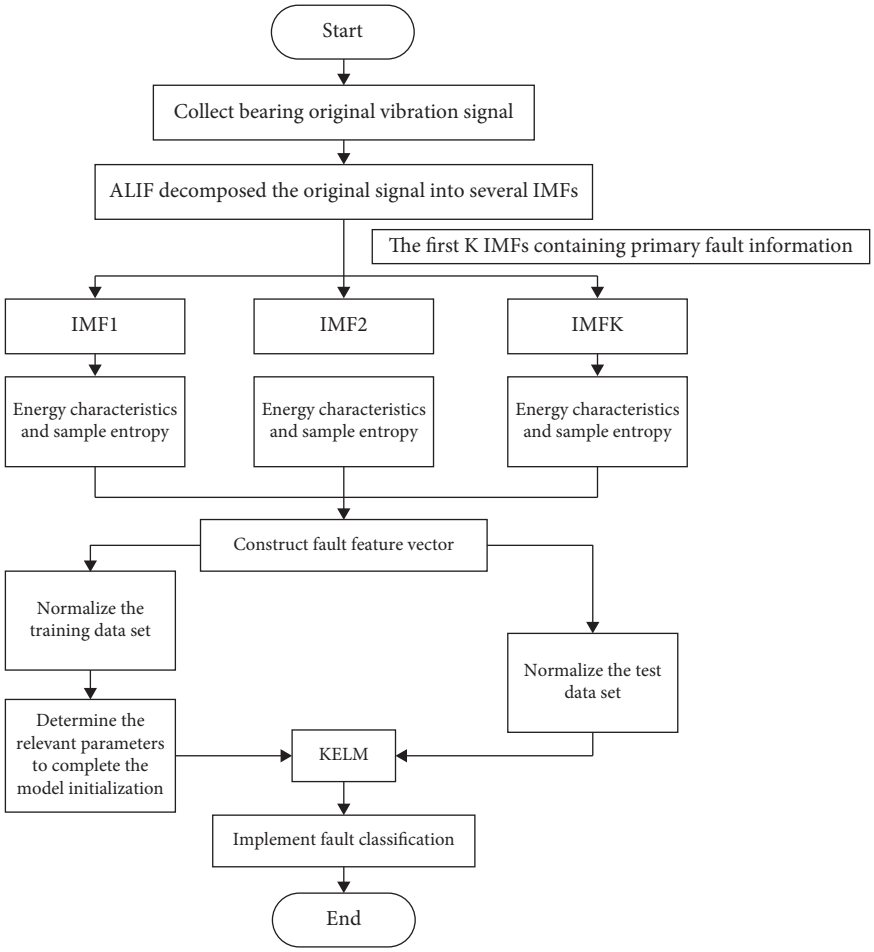


Figure 2: Fault diagnosis flowchart based on ALIF and KELM.

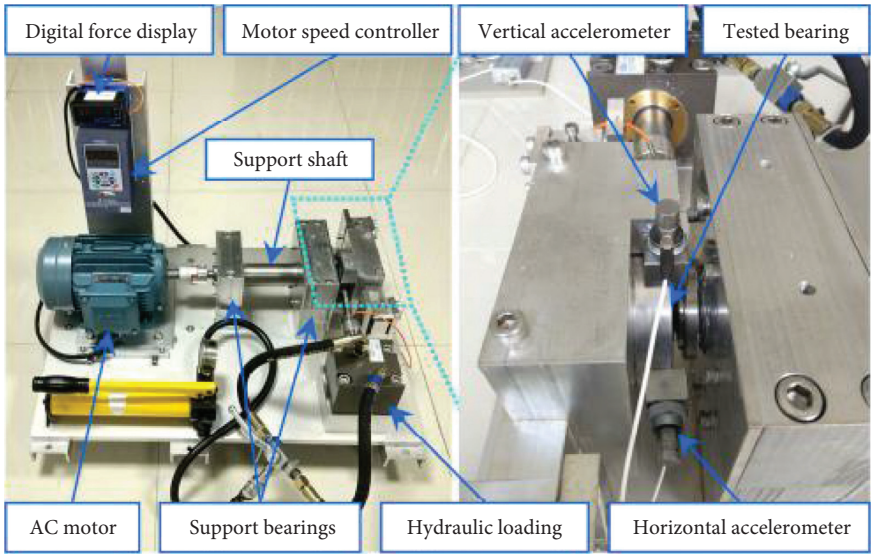


Figure 3: Rolling bearing accelerated life test bench.

components; the results are shown in Figures 5(a) and 5(b). It can be seen from both graphs that the energy characteristics of the first four IMF components and the sample entropy of the first four IMF components decomposed by ALIF are different under different operating conditions. However, if only energy is selected as the feature vector, there are three kinds of bearing state features in the IMF1 component that have obvious aliasing, and there are also two

Table 1: Dataset details.

Dataset	Fault type	Rotating speed (r/min)
Bearing1	Outer ring failure	2100
Bearing2	Compound failure of inner ring and outer ring	2100
Bearing3	Inner ring failure	2250
Bearing4	Cage failure	2250

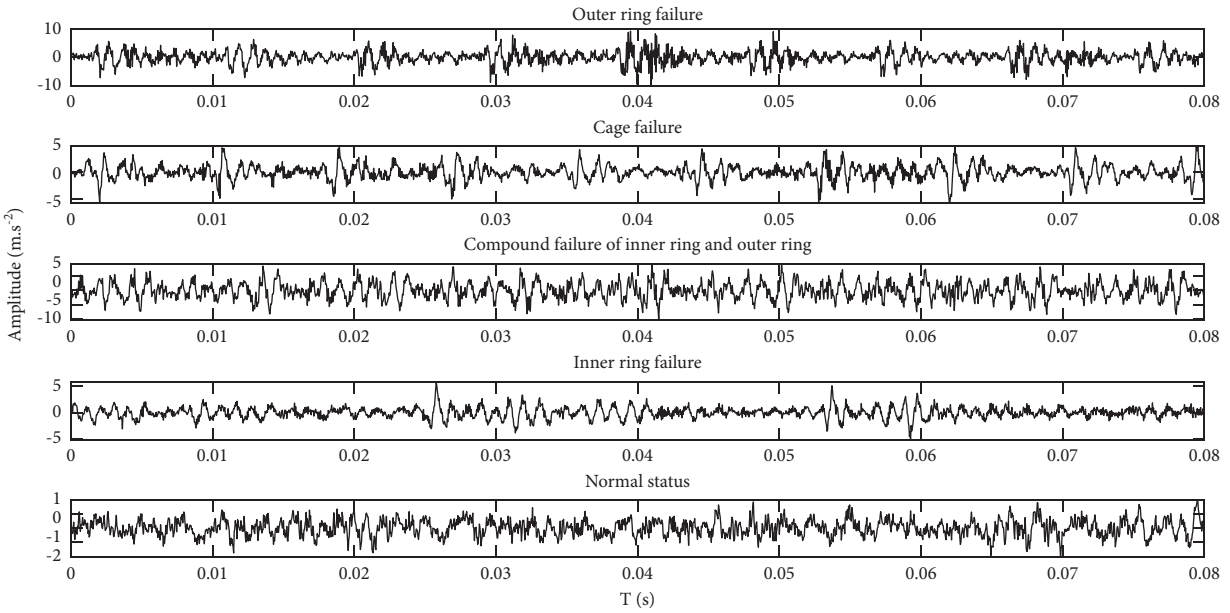


Figure 4: Time-domain waveforms of rolling bearing samples.

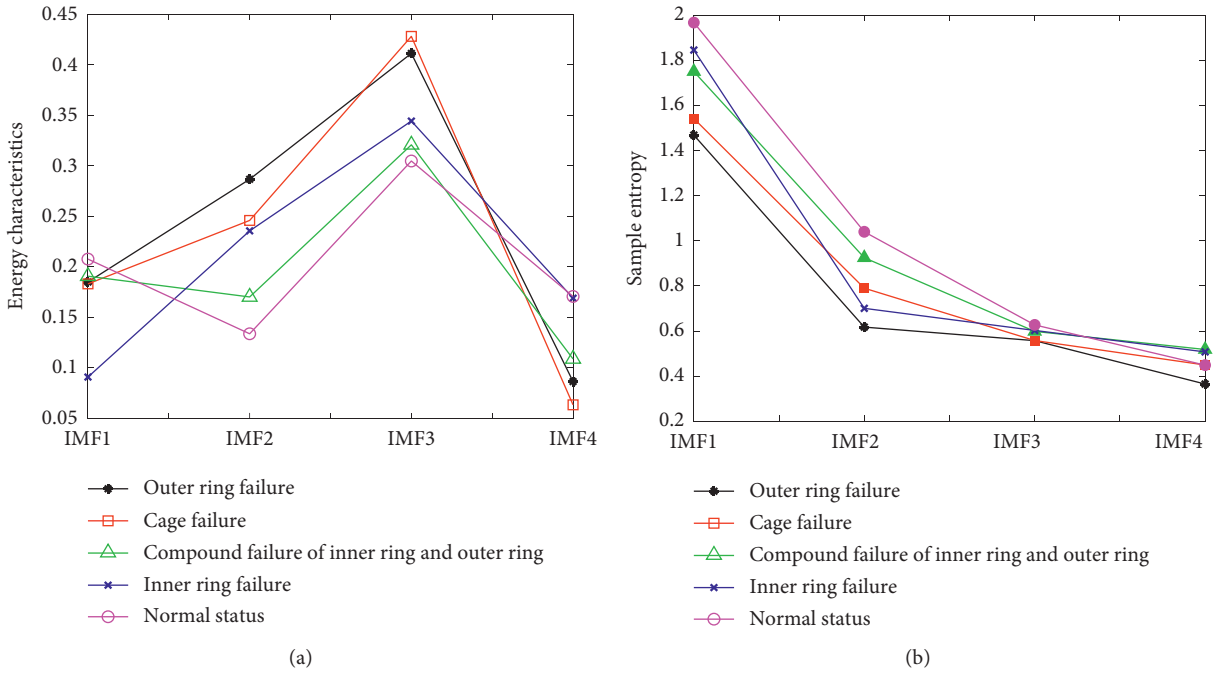


Figure 5: Energy characteristics and sample entropy of the first 4 components. (a) The energy characteristic value of the first 4 components. (b) The value of the sample entropy of the first 4 components.

bearing state features in the IMF4 component that basically overlap. This has been an obstacle to identifying types of bearing faults. When the energy feature and the sample entropy are selected as the fault feature vector, several bearing states where the original energy features are aliased can be effectively distinguished by the sample entropy.

5.3. Experimental Results and Analysis. In order to intuitively distinguish the different operating states of rolling bearings in the subsequent analysis, the sample data of the 4 types of faults and normal states are divided into 5 categories, and the specific category labels are given as shown in Table 2.

First, use the ALIF algorithm to decompose the vibration signals of all samples, make it get 5 IMF components and 1 residual component, calculate the energy characteristics and sample entropy of the first 4 IMF components, and obtain a fault eigenvector matrix with a size of 250×8. Then select 30 samples in each state as the training set and 20 samples as the test set, and normalize the fault feature data set to make the data indicators.

Finally, the Gaussian radial basis kernel function in equation (25) is used as the kernel function of KELM; the kernel coefficient $\lambda = 0.5$ and penalty coefficient $C = 1$ are determined. The initialization of the KELM intelligent diagnosis model is then completed. Input 150 training sample sets into the KELM model for training, and then apply 100 test samples for testing; the result is shown in Figure 6.

The abscissa in Figure 6 represents 100 sets of test samples, and each of the 5 bearing failure categories uses 20 sets of samples as the test; the vertical axis shows the category labels of different failures of rolling bearings, corresponding to Table 2. Judging from the diagnostic results, only one of the 100 test samples was misdiagnosed. The composite fault of the inner ring and the outer ring was misdiagnosed as a cage failure, while the remaining 99 test samples were accurately diagnosed. Taken together, the overall fault diagnosis accuracy rate is 99%, of which the single fault diagnosis accuracy rate is 100%, and the compound fault diagnosis accuracy rate is 95%. Therefore, the effectiveness of the intelligent diagnosis method of rolling bearing composite fault based on ALIF and KELM proposed in the paper can be proved.

However, in practical engineering applications, the sample data available for equipment is usually limited. The next step will be to study whether the method proposed in this paper can achieve higher accuracy fault intelligent diagnosis on test samples with fewer training samples. Setting the number of training samples for each failure type to 30, 25, 20, 15, 10, 5, and 1 in turn, the number of corresponding test samples is 20, 25, 30, 35, 40, 45, and 49; then we can get 7 different results of ALIF-KELM's diagnosis of rolling bearing faults. The relationship between the number of different training samples and the diagnosis accuracy is shown in Figure 7. It can be found that even when the number of training samples is only 1, the overall accuracy of ALIF-KELM can still reach 81.63%. When the number of training samples is 10, the fault intelligent diagnosis accuracy can also be as high as 99%; the specific diagnosis is shown in Figure 8. At this time, in all 200 test samples, only 2 sets of inner and outer ring compound faults were misdiagnosed as cage faults, so the compound fault accuracy rate is still 95%, which is the same as the accuracy rate when the number of training samples is 30. Compared with the

Table 2: Rolling bearing failure type label.

Fault type	Label
Outer ring failure	1
Cage failure	2
Compound failure of inner ring and outer ring	3
Inner ring failure	4
Normal status	5

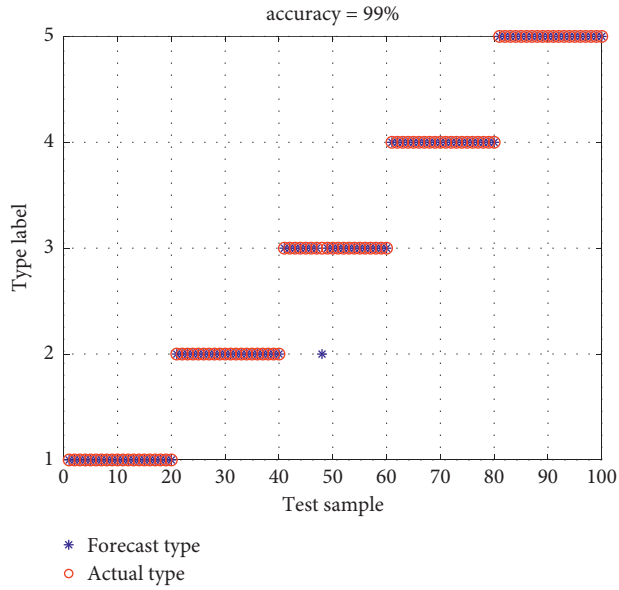


Figure 6: ALIF-KELM fault diagnosis results.

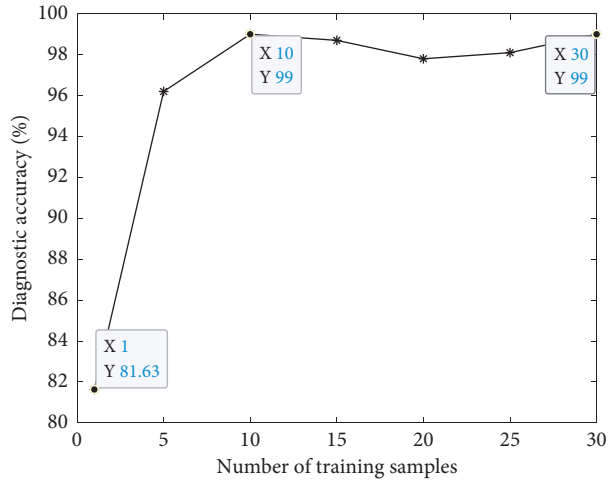


Figure 7: The fault diagnosis accuracy rate of ALIF-KELM under a different number of training samples.

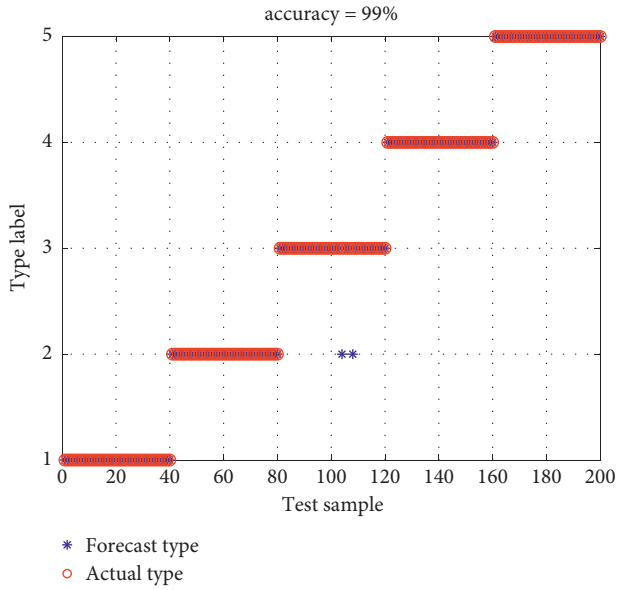


Figure 8: Fault diagnosis result of ALIF-KELM bearing under small sample.

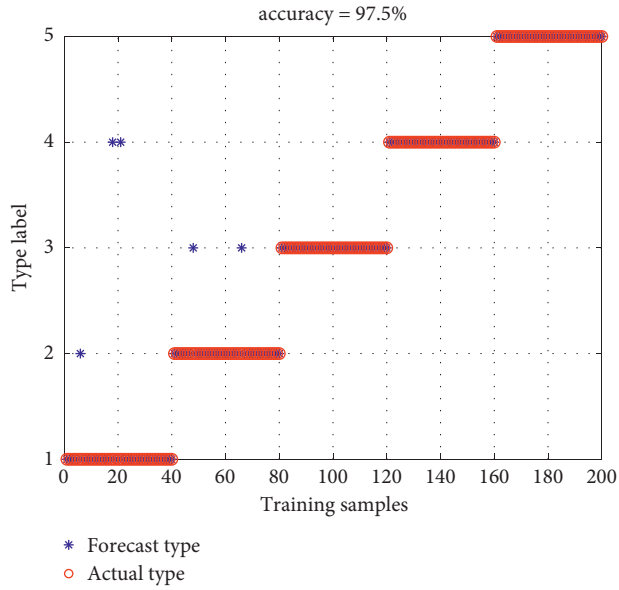


Figure 9: EMD-KELM bearing fault diagnosis results.

IFD-KELM method proposed in [29], when the sample size is 10, the accuracy of the method proposed in this paper is increased by 6.5%. Therefore, it can be further proved that the intelligent diagnosis method of rolling bearing composite fault based on ALIF and KELM proposed in this chapter can still be effective with less sample data.

5.4. Comparative Analysis of Experiments. In order to verify that the energy characteristics and sample entropy of the first four IMF components calculated after the application of ALIF decomposition can more effectively reflect the different fault characteristics of rolling bearings, the ALIF decomposition method is replaced with the traditional decomposition method EMD for comparison. There are 10 training samples and 40 test samples in each bearing state, and the KELM parameters remain the same as before. A total of 50 training sample sets of five bearing states are input into the KELM model for training, and then 200 test samples are used for testing; the result is shown in Figure 9. The figure shows that there are a total of 5 groups of test samples with diagnostic errors. Among them, 3 groups misdiagnosed the outer ring fault as a cage fault and an inner ring fault, and 2 groups misdiagnosed the cage fault as a composite fault of the inner ring and the outer ring; the composite fault test samples are all accurately identified, but the overall fault diagnosis accuracy rate is 97.5%. Figures 7 and 8 show that the overall diagnosis accuracy of ALIF-KELM is 99% in the case of 10 training samples. In contrast, the application of ALIF-KELM has higher accuracy than EMD-KELM's intelligent fault diagnosis method. It is verified that ALIF decomposition is more effective than EMD decomposition. In order to verify the superiority of the performance of the KELM fault diagnosis model, KELM was replaced with the two traditional diagnosis models of ELM and BP neural network. The method of feature extraction and the construction of feature vectors remain unchanged from the original method. The number of training samples in each bearing failure state is still 10, and the number of test samples is 40. Applying these two fault diagnosis models, respectively, the results are shown in Figures 10(a) and 10(b). Figure 10(a) shows that the accuracy of using ELM as a diagnostic model is 78.5%. The accuracy of BP as a diagnostic model is lower; Figure 10(b) shows only 53%. In short, compared with Figure 8, it can be found that the fault diagnosis accuracy of the two traditional diagnosis models is obviously much worse, and there are a lot of misdiagnosis phenomena in 200 test samples.

Since the initial weights of the two diagnostic models, ELM and BP, are randomly generated, the results obtained from each test are usually different. In order to reduce the impact of random fluctuations on the final comparison results, the three diagnostic models KELM, ELM, and BP were retrained and tested 10 times, and the accuracy of fault diagnosis was recorded. The results are shown in Figure 11. In order to be able to study the algorithm efficiency of each model and record the training time and test time consumed by each diagnostic model, the results are shown in Figures 12(a) and 12(b). Figure 10 shows that using KELM as an intelligent diagnosis model to identify fault types has not only the highest test accuracy but also the best stability. It can maintain an accuracy of 99% in 10 tests. Compared with the two diagnostic models of ELM and BP, the accuracy and stability of ELM are significantly better than that of BP. Figures 12(a) and 12(b) show that the KELM model is used for classification, the training time is the lowest, and the test time is basically the lowest, which is not much different from ELM, but the overall efficiency of KELM is the highest. And the algorithm efficiency of the BP model for classification is obviously lower than that of KELM and ELM. In short, ALIF-KELM has higher accuracy and algorithm efficiency than the two fault diagnosis methods ALIF-ELM and ALIF-BP.

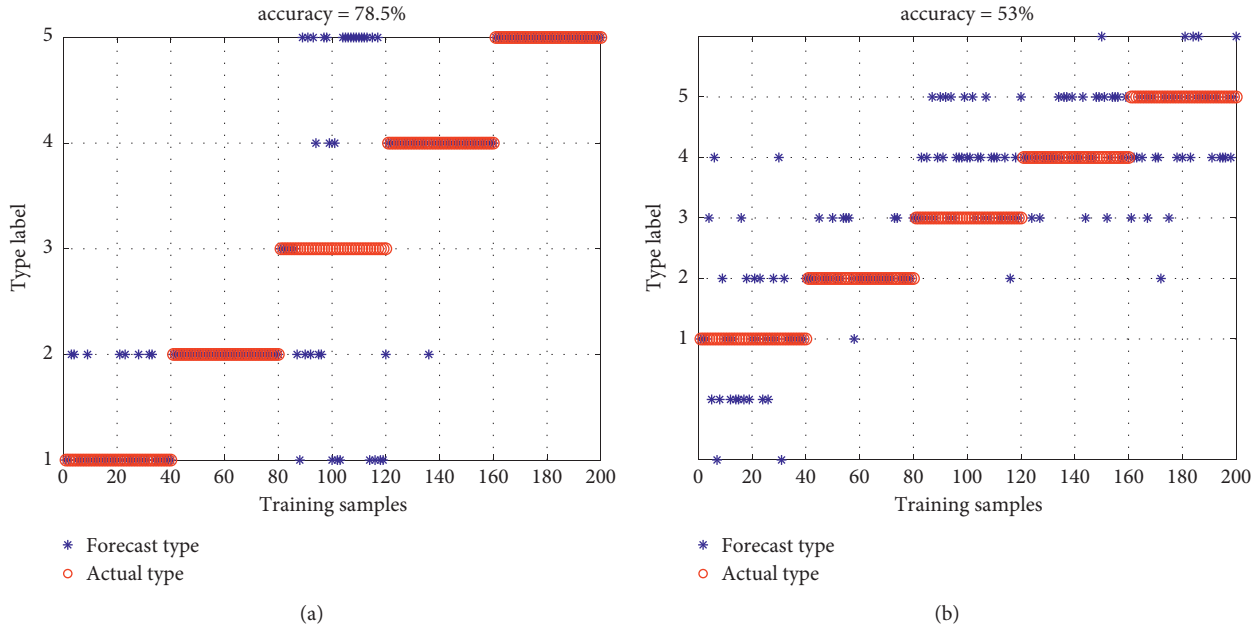


Figure 10: Diagnosis result of traditional diagnosis model. (a) ALIF-ELM fault diagnosis result. (b) EMD-BP fault diagnosis result.

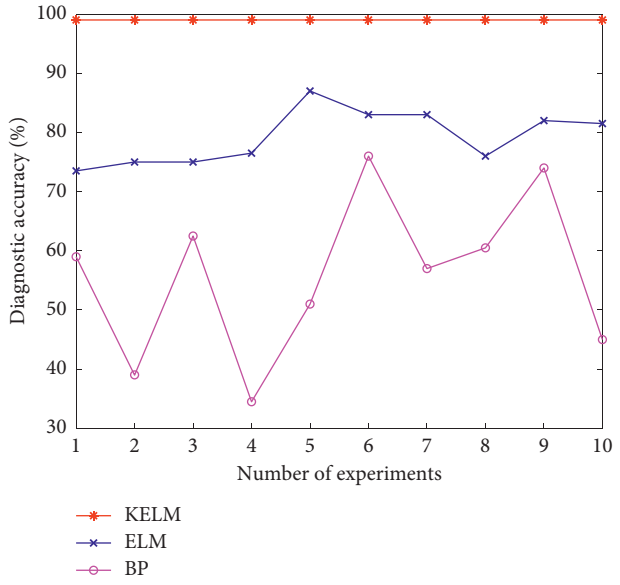


Figure 11: Accuracy of different fault diagnosis models.

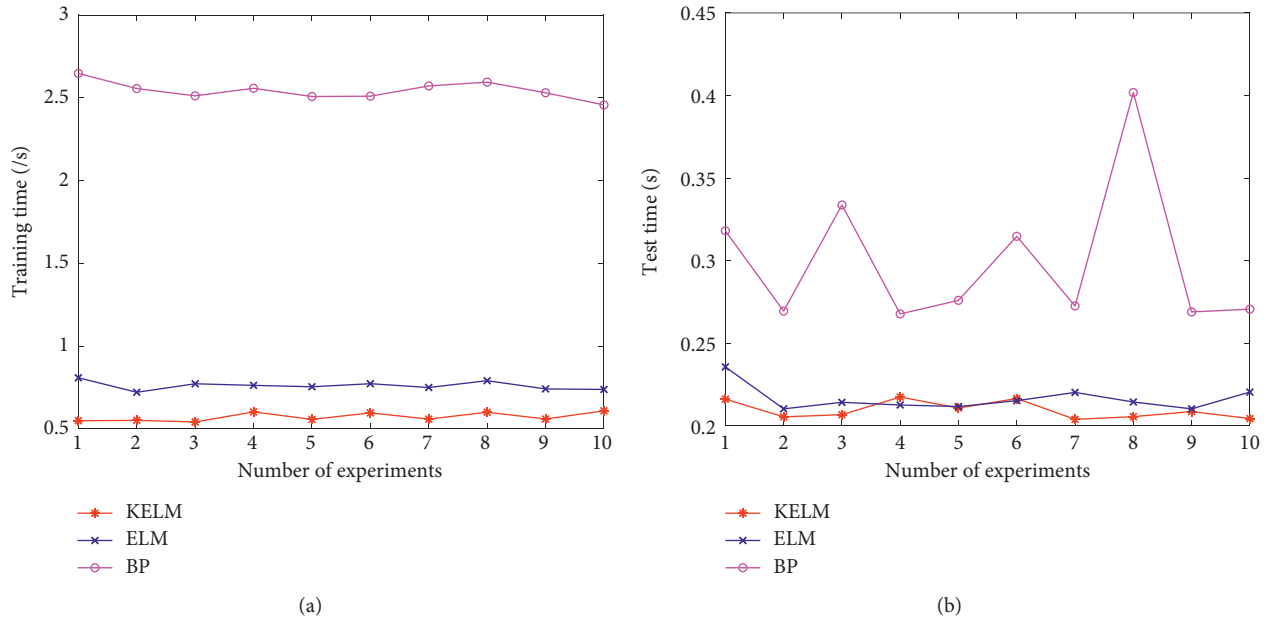


Figure 12: Time comparison of different fault diagnosis models. (a) Training time comparison. (b) Test time comparison.

6. Conclusions

In order to realize the composite fault diagnosis of rolling bearings, this paper proposes a diagnosis method by combining ALIF and KELM. In the case of a small sample, the proposed method is compared with EMD-KELM, ALIF-ELM, ALIF-BP, and IFD-KELM in [29]. The results show that both ALIF and KELM algorithms have certain advantages in the case of small samples, and the method proposed in this paper has high diagnostic accuracy and is suitable for the diagnosis of composite faults of rolling bearings.

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 study.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant no. 2019YFB1705403) and the National Natural Science Foundation of China (Grant no. 61973041).

References

[1] K. Zhang, D. H. Zhou, and Y. Chai, "Review of multiple fault diagnosis methods," *Control Theory & Applications*, vol. 32, no. 9, pp. 1143–1157, 2015.

[2] C. Li, M. Liang, and Z. Q. Chen, *Intelligent Health Management of Rolling Bearings Based on Vibration Signal*, Science Press, Beijing, China, 2018.

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

[4] N. E. Huang, Z. Shen, and S. R. Long, "A new view of nonlinear water waves: the Hilbert spectrum," *Annual Review of Fluid Mechanics*, vol. 31, no. 1, pp. 417–457, 1999.

[5] Z. L. Liu, "Time-frequency representation based on robust local mean decomposition for multicomponent AM-FM signal analysis," *Mechanical Systems and Signal Processing*, vol. 95, pp. 468–487, 2017.

[6] Y. Li, D. Huang, and Z. Qin, "A classification algorithm of fault modes-integrated LSSVM and PSO with parameters' optimization of VMD," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6627367, 12 pages, 2021.

[7] Z. Li, J. Chen, Y. Zi, and J. Pan, "Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive," *Mechanical Systems and Signal Processing*, vol. 85, pp. 512–529, 2017.

[8] C. J. Huang, H. J. Song, and N. Qin, "Fault diagnosis method of rolling bearing based on CLMD and CSES," *Electric Power Automation Equipment*, vol. 40, no. 8, pp. 179–188, 2020.

[9] W. M. Zhu, Y. Yu, and H. F. Xue, "Parameterized local characteristic scale decomposition and its application in composite fault diagnosis," *Journal of Mechanical & Electrical Engineering*, vol. 37, no. 6, pp. 593–599, 2020.

[10] A. J. Hu, Z. R. Bai, and J. Zhao, "Compound fault features separation method of rolling bearing based on parameter optimization VMD and 1.5 dimension spectrum," *Journal of Vibration and Shock*, vol. 39, no. 11, pp. 45–52+62, 2020.

[11] L. Lin, Y. Wang, and H. Zhou, "Iterative filtering as an alternative algorithm for empirical mode decomposition," *Advances in Adaptive Analysis*, vol. 1, no. 4, pp. 543–560, 2009.

[12] A. Cicone, J. Liu, and H. Zhou, "Adaptive local iterative filtering for signal decomposition and instantaneous frequency analysis," *Applied and Computational Harmonic Analysis*, vol. 41, no. 2, pp. 384–411, 2016.

[13] B. J. Chen, X. B. Wang, and C. H. Zhao, "Fault feature extraction of rolling bearing based on ALIF and energy operator demodulation," *Journal of Nanjing University of Science and Technology*, vol. 42, no. 4, pp. 445–452, 2018.

[14] J. B. Zhang, Y. Q. Zhao, and M. Liu, "An improved higher-order analytical energy operator with adaptive local iterative filtering for early fault diagnosis of bearings," *Journal of Vibroengineering*, vol. 22, no. 1, 2020.

[15] J. Ma and J. Xu, "Fault prediction algorithm for multiple mode process based on reconstruction technique," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–8, 2015.

[16] J. Viola, Y. Chen, and J. Wang, "FaultFace: deep convolutional generative adversarial network (DCGAN) based ball-bearing failure detection method," *Information Sciences*, vol. 542, pp. 195–211, 2021.

[17] J. Wang, W. Zhang, H. Wu, and J. Zhou, "Improved bilayer convolution transfer learning neural network for industrial fault detection," *The Canadian Journal of Chemical Engineering*, vol. 32, 2021.

[18] J. Wang, W. Zhang, and J. Zhou, "Fault detection with data imbalance conditions based on the improved bilayer convolutional neural network," *Industrial & Engineering Chemistry Research*, vol. 59, no. 13, pp. 5891–5904, 2020.

[19] L. P. Huang, C. W. Wu, and J. Wang, "Fault pattern recognition of rolling using wavelet package analysis and BP neural network," *Electronic Measurement Technology*, vol. 39, no. 4, pp. 164–168, 2016.

[20] L. Huo, X. Y. Zhang, and H. D. Li, "Bearing Fault diagnosis based on BP neural network," *IOP Conference Series: Earth and Environmental Science*, vol. 208, no. 1, Article ID 012092, 2018.

[21] J. M. Li, X. F. Yao, and X. D. Wang, "Multiscale local features learning based on BP neural network for rolling bearing intelligent fault diagnosis," *Measurement*, vol. 153, Article ID 107419, 2020.

[22] F. Han and D. S. Huang, "Improved extreme learning machine for function approximation by encoding a priori information," *Neurocomputing*, vol. 69, no. 16, 2006.

[23] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, pp. 985–990, Budapest, Hungary, July 2004.

[24] S. Xu and J. Ma, "Rolling bearing fault diagnosis based on EWT and ELM," *Vibroengineering PROCEDIA*, vol. 19, pp. 42–47, 2018.

[25] G.-B. Huang, "An insight into extreme learning machines: random neurons, random features and kernels," *Cognitive Computation*, vol. 6, no. 3, pp. 376–390, 2014.

[26] G. B. Guang-Bin Huang, H. Hongming Zhou, X. Xiaojian Ding, and fnm Rui Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[27] T. Liu, *Research on Intelligent Optimization Based-Extreme Learning Machine with Application to Disease Diagnosis*, Jilin University, Changchun, China, 2020.

[28] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 401–412, 2020.

[29] M. Q. Guo, Y. Y. Huang, and Q. Zhao, "Fault diagnosis method of rolling bearing based on IFD and KELM," *Journal of Fuzhou University (Natural Science Edition)*, vol. 48, no. 3, pp. 341–347, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 3716033, 9 pages
https://doi.org/10.1155/2021/3716033

Research Article

Compound Fault Diagnosis of Gearbox Based on RLMD and SSA-PNN

Shitong Liang and Jie Ma

School of Mechatronics Engineering, Beijing Information Science and Technology University, Beijing, China

Correspondence should be addressed to Jie Ma; mjbeijing@163.com

Received 6 July 2021; Accepted 18 September 2021; Published 29 September 2021

Academic Editor: Yong Chen

In order to solve the difficulty in the classification of gearbox compound faults, a gearbox fault diagnosis method based on the sparrow search algorithm (SSA) improved probabilistic neural network (PNN) is proposed. Firstly, the gearbox fault signal is decomposed into a series of product functions (PFs) by robust local mean decomposition (RLMD). Then, the permutation entropy of PFs, which contains much fault information, is calculated to construct the feature vector and input it into the SSA-PNN model. The experimental results show that compared with the traditional fault diagnosis methods based on EMD-BP and EEMD-PNN, the gearbox fault diagnosis method based on RLMD and SSA-PNN has higher diagnosis accuracy.

1. Introduction

The gearbox is the core component of mechanical equipment, and its running state is closely related to the safe operation of the equipment. Gearbox faults often occur as multiple faults in practical engineering applications and may cause abnormal operation of the equipment system and even lead to significantly reduced service life and degraded property. Therefore, compound fault diagnosis of the gearbox plays an important role in the safety maintenance of the mechanical system. Vibration signal analysis is one of the common analysis methods of gearbox fault diagnosis [1]. Vibration signals can be obtained through the contact sensor installed on the machine shell or base or through the airborne acoustic array sensor. However, in actual working conditions, the environmental background noise is large, and the fault impact characteristics of vibration signals are submerged in the cluttered noise, making it difficult to obtain fault information from original signals with the naked eye. Signal decomposition is one of the effective methods to deal with vibration signals. The ensemble empirical mode decomposition (EEMD) method [2] is widely used for feature extraction of fault signals. By adding Gaussian white noise when dealing with decomposed signal EMD, EEMD uses the binary filter bank characteristics of the EMD filter to fill the whole time-frequency space to reduce mode mixing. However, the added noise may not be completely eliminated and will cause signal reconstruction error [3]. To solve modal aliasing and end effect, Liu proposed a robust local mean decomposition (RLMD) method [4]. Yan [5] reconstructed the PFs obtained from RLMD of the signal and used the *K*-means++ clustering method to cluster the fault features. The effectiveness of this method was verified by simulation and experiments.

In the current fault diagnosis methods [6], the BP neural network is the most widely used, but it also has many shortcomings, such as the tendency to fall into the local extremum because of the dependence on the initial network weight, slow convergence, and so on. Compared with the BP neural network, the probabilistic neural network [7] converges faster and has higher diagnosis accuracy. Wang [8] used multiscale entropy (MSE) to extract fault features from signals and then input them into PNN. The results showed good fault diagnosis ability of the MSE-PNN model. Di [9] used EEMD to decompose signals into multiple IMFs, then took the energy as the feature vector, and inputted it into PNN. It was proved that this method has high recognition accuracy. However, the smoothing factor in PNN can only

be selected by artificial experience without a fixed method. Among swarm intelligence optimization algorithms [10–12], the sparrow search algorithm (SSA) [13] with strong search ability and fast convergence speed [14] is the best, and therefore, it can be used in the adaptive selection of smoothing factors to make them reflect the characteristics of the sample to the maximum extent. Accordingly, a gearbox fault diagnosis method based on RLMD and PNN optimized by SSA is proposed in this paper.

2. Robust Local Mean Decomposition

2.1. Local Mean Decomposition. The local mean decomposition (LMD) method [15] is an adaptive time-frequency representation method through iterative operation and can decompose a signal into a series of product functions (PFs), each of which is the product of the FM signal and the envelope signal. If the given original signal is $x(t)$, the LMD algorithm steps are as follows.

Step 1. All local maxima and minimums of the signal $x(t)$ are obtained. The extreme points are represented by e_w, and the corresponding extremes are marked as $x(e_w)$ with $w = 1, 2, 3, \ldots$

Step 2. The local mean $m^0(t)$ and local amplitude $a^0(t)$ are preprocessed according to formulas (1) and (2), and then, the smoothing algorithm is used to postprocess $m^0(t)$ and $a^0(t)$ for the smoothed local mean $m(t)$ and local amplitude $a(n)$:

$$m^0(t) = \frac{x(e_w) + x(e_{w+1})}{2}, \tag{1}$$

$$a^0(t) = \frac{|x(e_w) - x(e_{w+1})|}{2}. \tag{2}$$

Step 3. The initial local average $x(t)$ is removed from the original signal $m_{11}(t)$, and the estimated zero-mean signal $h_{11}(t)$ is obtained as

$$x(t) - m_{11}(t) = h_{11}(t). \tag{3}$$

Step 4. The estimated FM signal $s_{11}(t)$ is obtained by dividing $h_{11}(t)$ by $a_{11}(t)$, that is,

$$s_{11}(t) = \frac{h_{11}(t)}{a_{11}(t)}. \tag{4}$$

After the LMD, if the signal $s_{11}(t)$ does not meet the requirements, that is, it is not a pure FM signal, and $s_{11}(t)$ will be regarded as a new signal to repeat Steps (1) to (4) P times until the conditions in formula (5) are satisfied:

$$\lim_{P \longrightarrow \infty} a_{1p}(t) = 1. \tag{5}$$

Step 5. When formula (5) is established, the FM signal $s_1(t)$ which meets the requirements can be calculated by formula (6), and the envelope signal $a_1(t)$ can be calculated by formula (7). The first product function $PF_1(t)$ can be obtained by multiplying $s_1(t)$ with $a_1(t)$:

$$s_1(t) = s_{1P}(t), \tag{6}$$

$$a_1(t) = \prod_{j=1}^{P} a_{1j}(t). \tag{7}$$

Step 6. The residual signal $u_1(t)$ is obtained by subtracting $PF_1(t)$ from the original signal. Steps (1) to (5) are repeated Q times until $u_1(t)$ is a constant or nonoscillatory function, and then, the original signal can be represented by the sum of multiple product functions and residual components:

$$x(t) = \sum_{i=1}^{Q} PF_i(t) + u_Q(t). \tag{8}$$

2.2. Robust Local Mean Decomposition. Because of mode aliasing and end effect of LMD, Liu proposed RLMD. The improvements are as follows.

Step 1. Boundary condition optimization: the mirror expansion algorithm [16] is used to find the symmetrical points of the signal with respect to the endpoints at both ends.

Step 2. Envelope estimation: for the value λ^* that needs to be selected by artificial experience, Liu creatively uses a method based on statistical theory:

$$\lambda^* = \mathrm{odd}(\mu_s + 3 \times \delta_s). \tag{9}$$

In the formula, $\mathrm{odd}(\cdot)$ is the nearest odd number of the input, μ_s is the center of the step, and δ_s is the standard deviation of the step.

Step 3. Stop criteria filter: the objective function $f(x)$ is minimized:

$$f(x) = \mathrm{RMS}(z(t) + \mathrm{EK}z(t)), \tag{10}$$

where the zero-baseline envelope signal $z(t) = a(t) - 1$ and $\mathrm{RMS}(\cdot)$ and $\mathrm{EK}(\cdot)$ are given by formulas (11) and (12) as follows:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(z(t))^2}, \tag{11}$$

$$\mathrm{EK} = \frac{(1/N)\sum_{t=1}^{N}(z(t) - \overline{z})^4}{\left((1/N)\sum_{t=1}^{N}(z(t) - \overline{z})^2\right)^2} - 3, \tag{12}$$

where $\overline{z}$ is the average of $z(n)$.

2.3. Simulation Experiment and Analysis. In order to verify the superiority of RLMD, a composite signal is simulated:

$$\begin{cases} x_1(t) = \cos(300\pi t), \\ x_2(t) = \sin(1600\pi t), \\ x_3(t) = 0.5randn(1,n), n = \text{length}(t), \\ x(t) = x_1(t) + x_2(t) + x_3(t). \end{cases} \tag{13}$$

The sampling frequency is 20480 Hz, the number of sampling points is 4096, and the time domain waveforms of each component and composite signal are shown in Figure 1.

The first four PFs and IMFs obtained by simultaneous RLMD and EMD of the signal are shown in Figures 2 and 3.

Figure 2 shows that the PF1 component mainly corresponds to function $x_2(t)$, and the PF2 component mainly corresponds to function $x_1(t)$, while in Figure 3, all components are affected by modal aliasing, resulting in that their significance is not obvious. Therefore, the superiority of the RLMD method is shown.

3. PNN Model Optimized by SSA

3.1. Sparrow Search Algorithm. Sparrow search algorithm is a new swarm intelligence optimization algorithm, which simulates the foraging behavior of the sparrow population, and in this algorithm, the sparrow population is divided into discoverers and participators. The population X with the number of sparrows n is expressed by formula (13):

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix}. \tag{14}$$

In the formula, d is the dimension of the parameter to be optimized and n is the number of populations. Then, the fitness values of all sparrows can be expressed as formula (14):

$$F_X = \begin{bmatrix} f([\, x_{1,1} \;\; x_{1,2} \;\; \cdots \;\; x_{1,d} \,]) \\ f([\, x_{2,1} \;\; x_{2,2} \;\; \cdots \;\; x_{2,d} \,]) \\ \vdots \\ f([\, x_{n,1} \;\; x_{n,2} \;\; \cdots \;\; x_{n,d} \,]) \end{bmatrix}. \tag{15}$$

The iterative formula of the discoverer's position is expressed as formula (15):

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t} \cdot \exp\left(-\dfrac{i}{\alpha \cdot \text{iter}_{\max}}\right), & \text{if } R_2 < ST, \\ X_{i,j}^{t} + Q \cdot L, & \text{if } R_2 \geq ST, \end{cases} \tag{16}$$

where t is the number of iterations, $\text{iter}_{\max}$ is the maximum number of iterations, and Q is a random number, which obeys normal distribution. The maximum value of j is the dimension d of the parameter to be optimized; L is a row matrix with the length d, and the element is 1; R_2 and ST represent early warning value and safety value, respectively.

The iterative formula of the participant's position is expressed as formula (16):

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\dfrac{X_{\text{worst}} - X_{i,j}^{t}}{i^2}\right), & \text{if } i > \dfrac{n}{2}, \\ X_{p}^{t+1} + \left|X_{i,j}^{t} - X_{P}^{t+1}\right| \cdot A^{+} \cdot L, & \text{otherwise}, \end{cases} \tag{17}$$

where X_p and X_{worst} represent the global best location and global worst location found by the discoverer so far, respectively. A is a row matrix with the length d, the element is set as 1 or -1 randomly, and $A^{+} = A^{T}(AA^{T})^{-1}$.

In this paper, the sum of the classification error rate of the training set and the classification error rate of the test set is used as the fitness value.

3.2. Probabilistic Neural Network. A probabilistic neural network [17] is a kind of neural network that can be used for pattern classification, and its essence is a parallel algorithm based on the Bayesian minimum risk criterion. It has the characteristics of a simple learning process, fast training speed, more accurate classification, good fault tolerance, and so on.

Probabilistic neural networks are generally divided into four layers, namely, input layer, pattern layer, summation layer and output layer.

The input layer is used to input the high-dimensional fault feature matrix for analysis. The number of input layers is affected by the dimension of the fault feature matrix.

The pattern layer connects with the input layer through the connection weight. It calculates the matching degree, that is, the similarity between the input feature vector and each pattern in the training set, and then inputs it into the activation function. The result is the output of the pattern layer.

The summation layer is responsible for connecting the pattern layer units of each class. The number of neurons in this layer is consistent with the number of fault types.

The output layer classifies the fault type by outputting the maximum value in the summation layer.

The basic model of PNN is shown in Figure 4.

4. Compound Fault Diagnosis of the Gearbox by RLMD-SSA-PNN

In order to solve the problems of compound fault diagnosis of the gearbox, an improved PNN algorithm based on RLMD and SSA is proposed in this paper, and the specific steps are as follows.

Step 1. The signal is decomposed by RLMD, and a series of PFs are obtained. The number of PFs containing much fault information is judged by the correlation coefficient, and the permutation entropy of each effective PF is calculated. After that, the high-dimensional fault characteristic matrix is constructed.

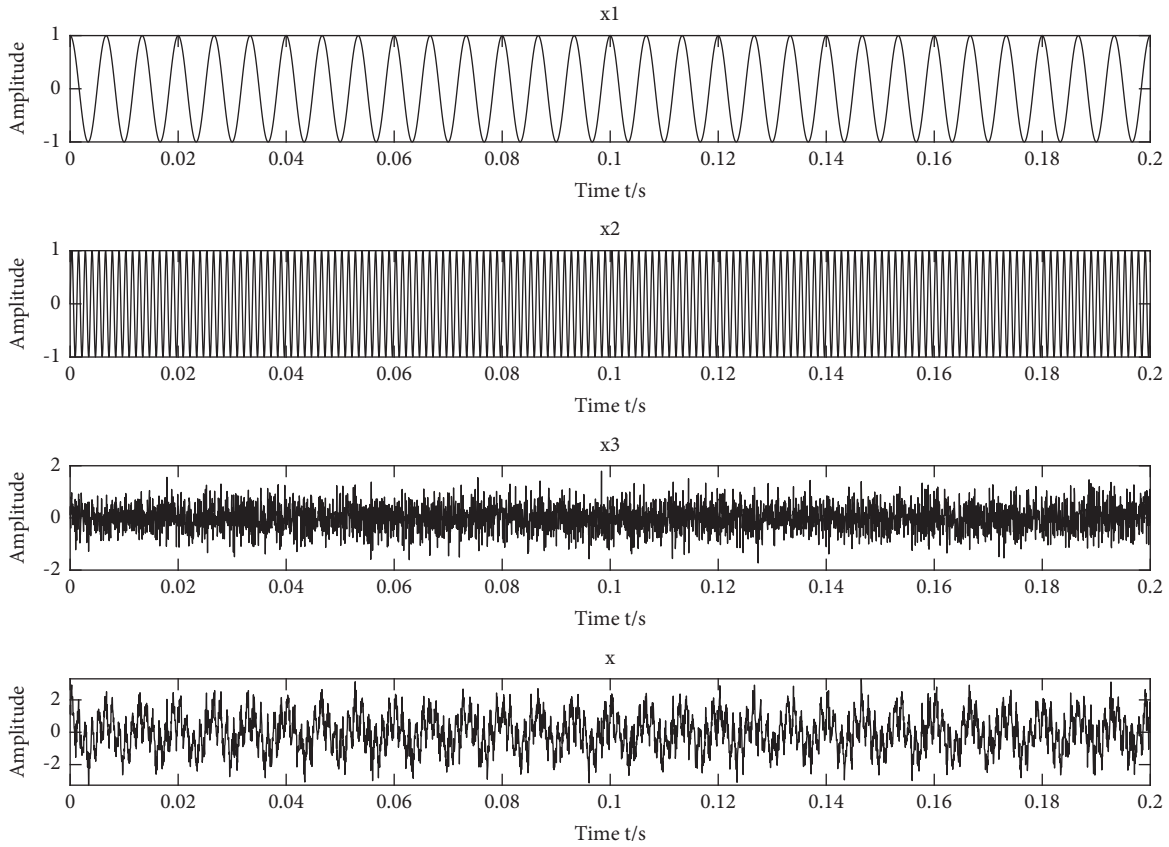


Figure 1: Each component and coincidence signal waveform.

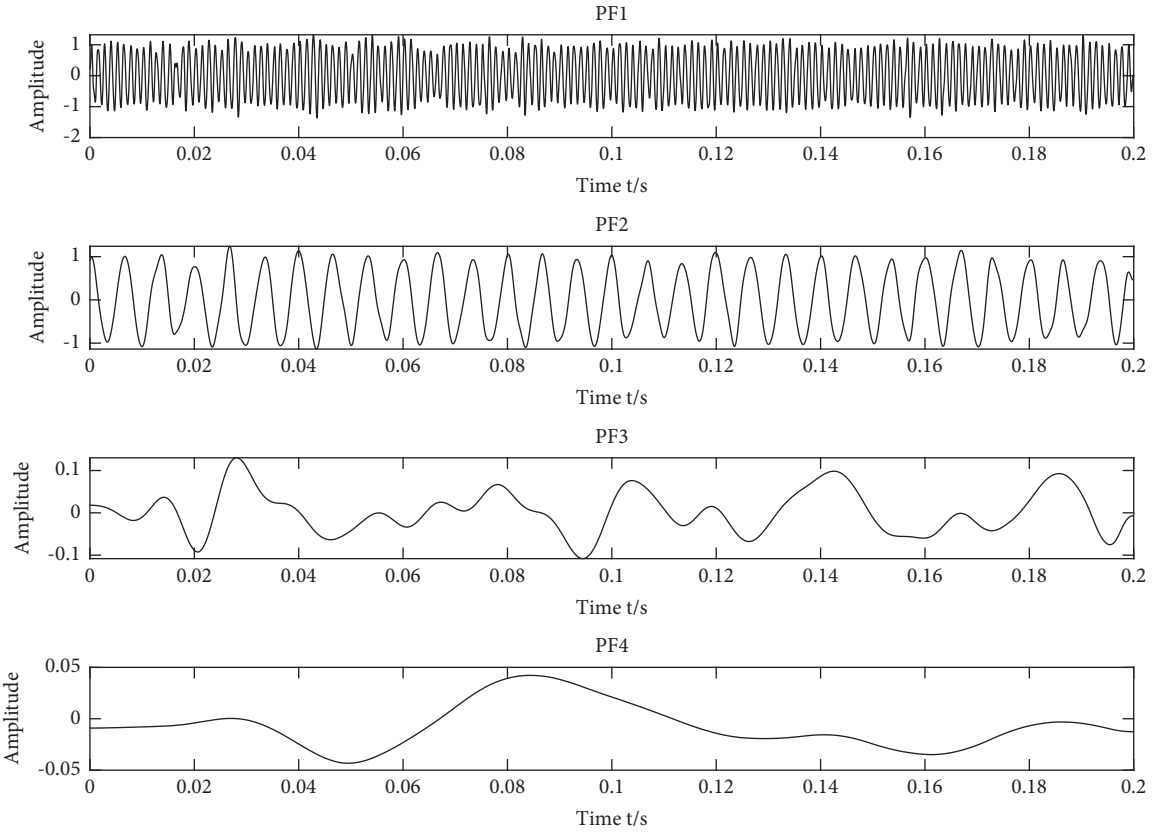


Figure 2: PFs waveform.

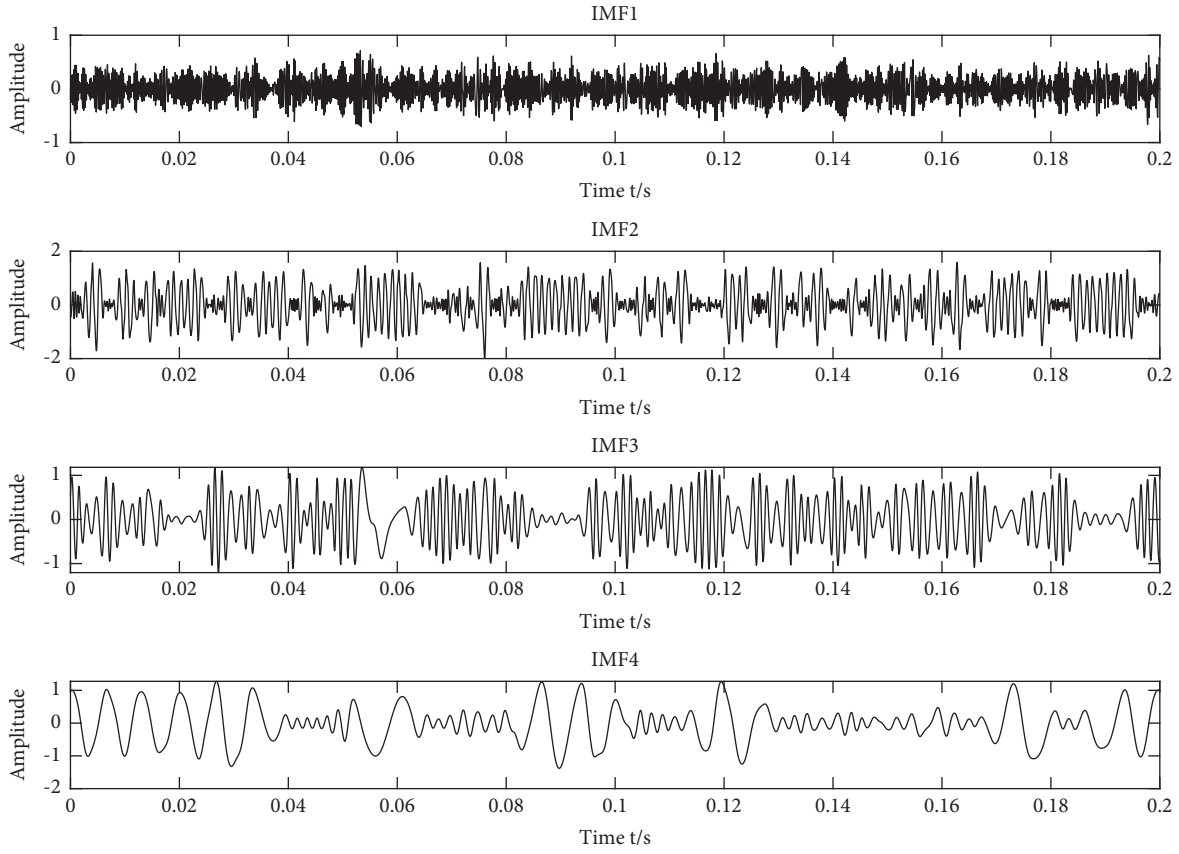


Figure 3: IMFs waveform.

Step 2. The fault characteristic matrix is divided into the training set and test set, which are then labeled.

Step 3. Parameters in the sparrow search algorithm are set, such as the number of populations, maximum number of iterations, and upper and lower boundaries of the search (the range of smoothing factors).

Step 4. The training set and its label are input into the SSA-PNN model for training, making the model find the optimal value of the smoothing factor, and then, the test set and its label are input for the test.

5. Experiment and Analysis

5.1. Data Acquisition. This paper used the gearbox test platform of the Ministry of Education Key Laboratory in Beijing Information Science and Technology University to collect data. The test bench includes an acceleration sensor, conditioning circuit, acquisition instrument, planetary gearbox test platform, and computer.

The range of speed adjustment is 1140–2220 rpm, and the step size is 120 rpm. Five types of data are included, namely, the normal planetary gear, planetary gear tooth fracture, planetary gear tooth surface wear, planetary gear tooth fracture plus tooth surface wear, and planetary gear tooth fracture plus rolling body missing. Each type of data is collected three times, with the sampling frequency of 20480 Hz and the sampling time of each collection being 3s. The length of each collected signal is 61440. The data used in this paper are the fault data under the speed of 1500 r/min, and the time domain diagram of each type is shown in Figure 5.

5.2. Construction of High-Dimensional Fault Characteristic Matrix. Considering the computational cost and efficiency, each collected signal is divided into 30 short samples, with the length of each sample being 2048, and there are a total of 90 short samples for each type of data. Sixty of them are randomly selected as the training set, and the remaining 30 are chosen as the test set. Since there are 5 data types in the experiment, a total of 300 short samples are used as the training set, and 150 samples are selected as the test set. The labels of the normal type, broken tooth type, wear type, wear and broken tooth type, broken tooth type, and rolling body missing type are 1, 2, 3, 4, and 5, respectively.

Taking the fault signal of tooth breaking of the planetary gear as an example, this paper shows the decomposition of a short sample with a length of 2048 by RLMD to get multiple PFs, as shown in Figure 6.

The above six PFs' components are obtained after the fault signal of planetary gear tooth breaking is decomposed by RLMD. In fact, there are still meaningless components in the PFs' components of the signal after RLMD. Taking these components as feature elements will cause interference and reduce the recognition accuracy of the recognition

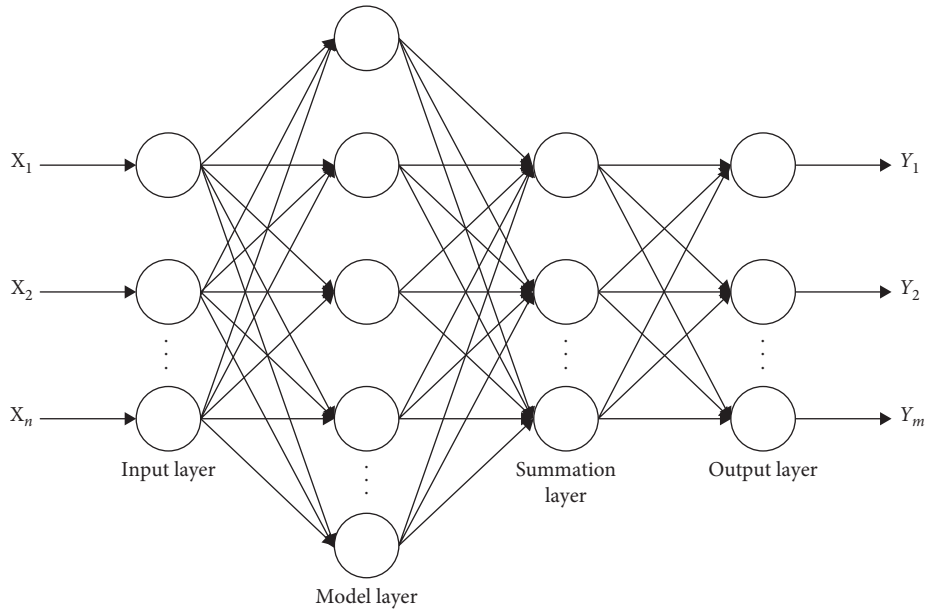


Figure 4: Schematic diagram of the PNN structure.

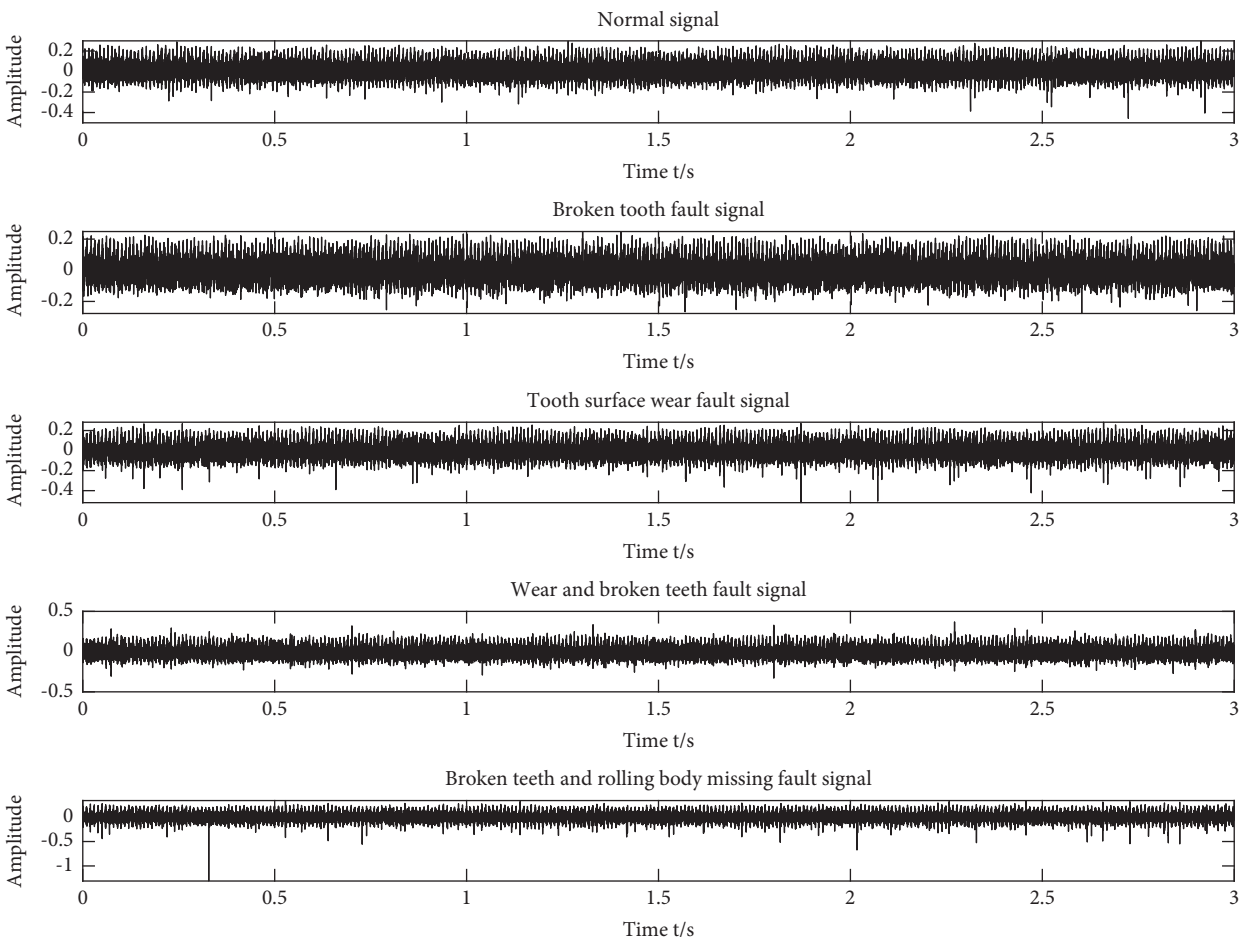


Figure 5: Five types of data waveform.

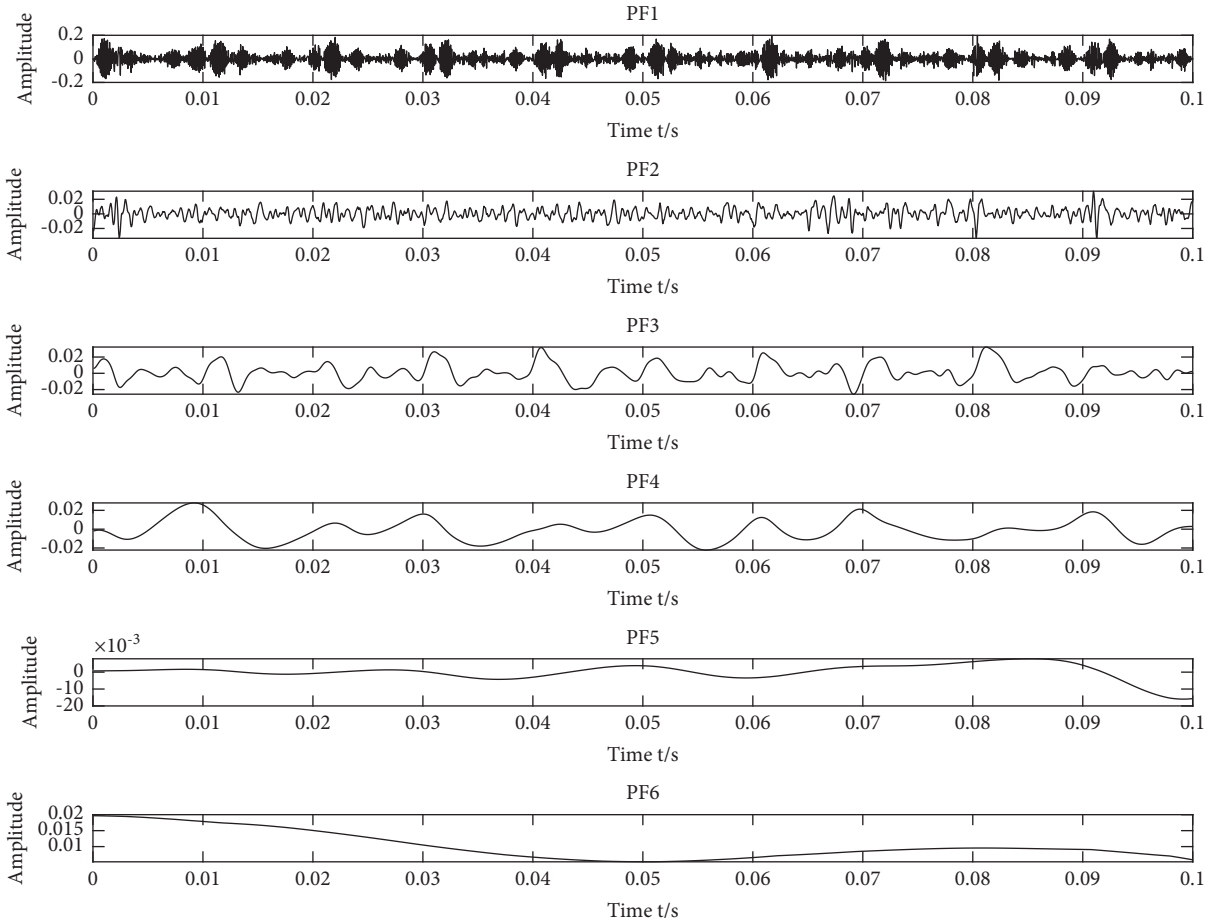


Figure 6: Each PF waveform.

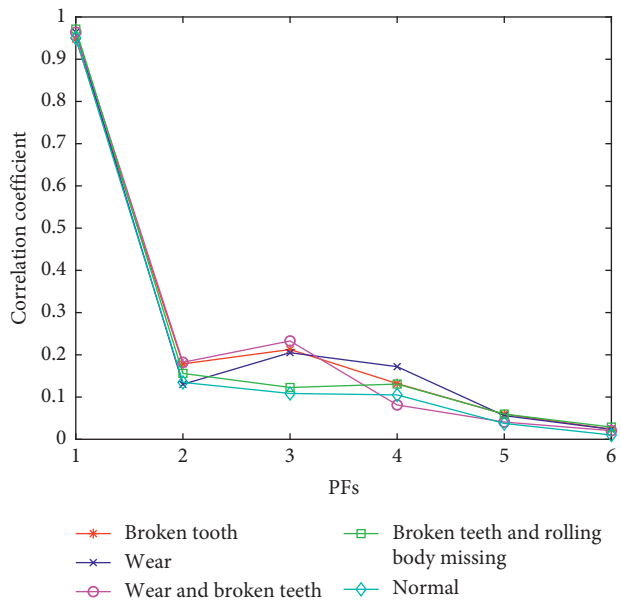


Figure 7: Correlation coefficient between each PF and the original signal.

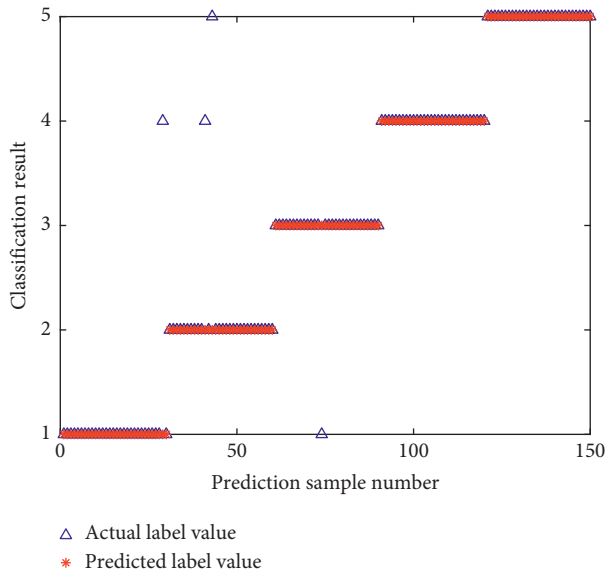


Figure 8: Classification result of the SSA-PNN model.

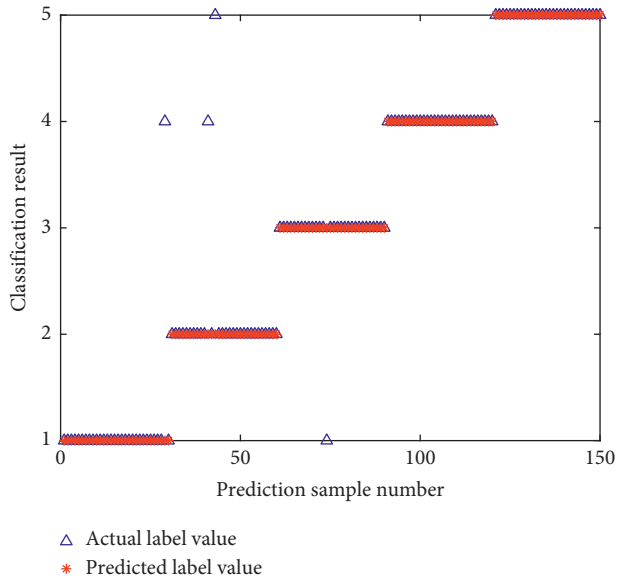


Figure 9: Classification result of the EMD-BP model.

algorithm. Therefore, it is necessary to use the correlation coefficient method to filter out the unimportant PFs components.

The correlation coefficients of PFs' components of five types of data decomposed by RLMD are calculated, as shown in Figure 7. It can be seen that the correlation coefficients of the PFs after PF4 are all lower than 0.1, which indicates that the latter PFs can hardly reflect the fault characteristics of the original signal, and therefore, the first four PFs are retained as effective components. Then, the permutation entropy of these four PFs is calculated, and the fault characteristic matrix of 450×4 is obtained.

5.3. Experimental Results and Analysis. The fault characteristic matrix is inputted into the SSA-PNN model, and the result is shown in Figure 8. It is shown that, in the 150 short samples of the test set, only 4 short samples have classification errors, including the diagnosis of the normal type as wear plus broken teeth, the diagnosis of the broken teeth as wear plus broken teeth, the diagnosis of the broken teeth as rolling body missing plus broken teeth, and the diagnosis of the wear as the normal type, and the overall classification accuracy reaches 97.33%.

In order to verify the superiority of the proposed method, it is compared with other methods. This paper uses

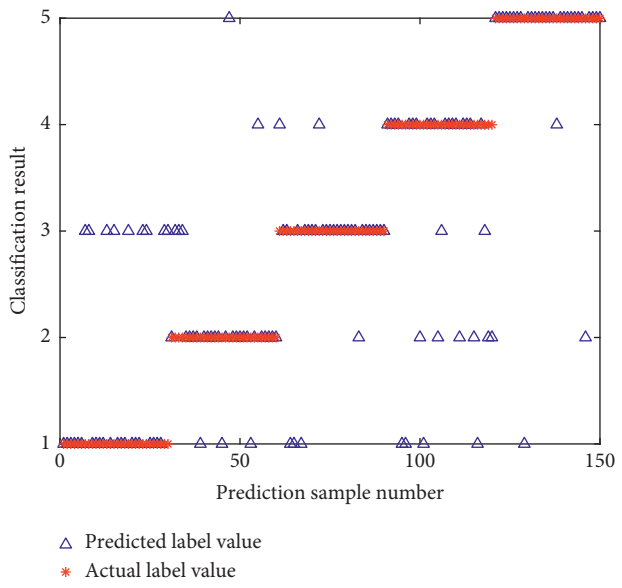


Figure 10: Classification result of the EEMD-PNN model.

EMD to decompose the signal, then takes the first four IMFs as sensitive components, and also uses permutation entropy as the index to construct the feature matrix, which is input into the BP model. The classification results of the EM-BP model are shown in Figure 9. It is shown that there are 38 classification errors in the short samples of the test set, and the classification accuracy is only 74.67%. Similarly, the EEMD-PNN model is used, and its fault classification results are shown in Figure 10. It can be seen that there are 13 classification errors in the short samples of the test set, and the classification accuracy is 91.33%.

6. Conclusions

In this paper, a gearbox fault diagnosis method based on robust mean decomposition and SSA improved PNN model is proposed. This method can effectively adaptively select the smoothing factors in the PNN model so as to achieve a good classification effect. Besides, it has higher diagnosis accuracy than EMD-BP, EEMD-PNN, and other classification methods and is suitable for fault diagnosis of the gearbox.

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

This work was supported by the National Natural Science Foundation of China (Grant no. 61973041) and National Key Research and Development Program of China (Grant no. 2019YFB1705403).

References

[1] G. Yu, T. Lin, Z. Wang, and Y. Li, "Time-reassigned multisynchrosqueezing transform for bearing fault diagnosis of rotating machinery," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1486–1496, 2021.

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

[3] C. Ma, S. H. Wang, and X. L. Xu, "Fault diagnosis for rolling bearing by using acoustic array based on EEMD," *Journal of Electronic Measurement and Instrument*, vol. 31, no. 9, pp. 1379–1384, 2017.

[4] Z. Liu, Y. Jin, M. J. Zuo, and Z. Feng, "Time-frequency representation based on robust local mean decomposition for multicomponent AM-FM signal analysis," *Mechanical Systems and Signal Processing*, vol. 95, pp. 468–487, 2017.

[5] S. T. Yan, Y. G. Zhou, and Y. B. Rrn, "Bearing fault diagnosis method based on RLMD and Kmeans++," *Journal of Mechanical Transmission*, vol. 45, no. 2, pp. 163–170, 2021.

[6] J. Ma and J. A. Xu, "Fault prediction algorithm for multiple mode process based on reconstruction technique," *Mathematical Problems in Engineering*, vol. 2015, Article ID 348729, 8 pages, 2015.

[7] L. X. Yang and Y. L. Zhu, "High voltage circuit breaker fault diagnosis of probabilistic neural network," *Power System Protection and Control*, vol. 43, no. 10, pp. 62–67, 2015.

[8] W. Wang, B. Zhou, and R. M. Zhang, "Fault diagnosis of rolling bearing in mechanical equipment based on MSE and PNN," *China Plant Engineering*, vol. 15, pp. 145–147, 2020.

[9] H. Di, W. L. Sun, and Y. Z. Wu, "Research on fault diagnosis for wind turbine based on EEMD and PNN," *Machinery design and manufacture*, vol. 6, pp. 105–108, 2020.

[10] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, no. 3, pp. 46–61, 2014.

[11] S. Seyedali, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multiobjective problems," *Neural Computing and Applications*, vol. 27, no. 4, pp. 1053–1073, 2016.

[12] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper optimisation algorithm: theory and application," *Advances in Engineering Software*, vol. 105, pp. 30–47, 2017.

[13] J. K. Xue, *Research and Application of A Novel Swarm Intelligence Optimization Technique: Sparrow Search Algorithm*, Donghua University, Shanghai, China, 2020.

[14] Y. L. Li, S. Q. Wang, and Q. R. Chen, "A comparative study of several new swarm intelligence optimization algorithms," *Computer Engineering and Applications*, vol. 56, no. 22, pp. 1–12, 2020.

[15] S. Q. Chen, Z. K. Peng, and P. Zhuo, "Review of signal decomposition theory and its applications in machine fault diagnosis," *Journal of Mechanical Engineering*, vol. 56, no. 17, pp. 91–107, 2020.

[16] G. Rilling, P. Flandrin, and P. Goncalves, "On empirical mode decomposition and its algorithms," in *Proceedings of the 6th IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing*, Grado, Italy, 2003.

[17] F. Specht Donald, "Probabilistic neural networks," *Pergamon*, vol. 3, no. 1, 1990.