

## Article

# A Novel Multi-Source Fault Diagnosis Strategy Based on Knowledge and Data Dual-Drive for a Planetary Gearbox

Hanzhi Yang and Jun Hao \* 

College of Automation, Jiangsu University of Science and Technology, Zhenjiang 212100, China;  
232210305205@stu.just.edu.cn

\* Correspondence: j\_hao4415@tju.edu.cn

## Abstract

Traditional fault diagnosis methods face challenges such as the insufficient utilization of fault information and an imbalance between classification accuracy and computation. To address these issues, a novel multi-source fault diagnosis strategy based on a knowledge and data dual-drive algorithm is proposed. Firstly, a multi-source information correlation matrix (MICM) is designed to enhance the expression of fault information by combining information among time domain, frequency domain and channel correlation features. Then, kernel principal component analysis (KPCA) is used for dimensionality reduction in the MICM. Finally, a novel classifier based on Softmax logical regression (SLR) and K nearest neighbor (KNN) is proposed, where SLR provides an initial pre-classification and KNN is used to achieve accurate classification with less computation. Moreover, the latest planetary gearbox dataset of the wind turbine in a physical experiment is utilized to verify the effectiveness of the proposed MICM-SLR-KNN algorithm, and the experimental results demonstrate the superiority of the algorithm in comparison with other methods.

**Keywords:** knowledge and data dual-drive; multi-source fault diagnosis; multi-source information correlation matrix; planetary gearbox; Softmax logical regression and KNN

## 1. Introduction

In recent decades, wind energy as a form of renewable energy has become a cornerstone of global green energy production [1–3]. However, the fault downtime of wind turbines caused by the planetary gearbox is increasing, which is directly related to wind energy's safety and reliability [4]. Hence, it is of great significance to achieve fault diagnosis of wind turbine planetary gearboxes to reduce power loss and ensure equipment safety and wind turbine reliability. Research into fault diagnosis methods is gradually becoming a hot topic for complex systems.

Fault diagnosis methods are usually divided into knowledge-based and data-driven methods for wind turbine planetary gearboxes. In [5], representative knowledge-based KNN is used to classify the main features and confirm the fault classes, and it demonstrates considerable diagnostic efficacy with remarkable precision. An improved KNN algorithm is introduced in [6], which employs cross-validation to determine the optimal value of K and incorporates Gaussian-based distance weighting for neighboring samples. The method in [7] combines multiple KNNs using different distance metrics and adopts a weighted voting scheme based on the individual precision, which enhances predictive precision and robustness. The aforementioned knowledge-based fault diagnosis methods (KFDMs) have achieved good results. However, these approaches usually have a high computational cost



Academic Editor: Jongmyon Kim

Received: 24 March 2026

Revised: 27 April 2026

Accepted: 5 May 2026

Published: 8 May 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

to achieve fault diagnosis, which results in inefficiency and a difficulty in adapting to large datasets. Reducing online computation while ensuring precision is an urgent problem that needs to be solved for KFDMs.

Recently, with the development of sensors, fault data has become easier to obtain, making data-driven fault diagnosis methods (DFDMs) attract considerable attention [8–10]. Classic data-driven algorithms are often used as benchmarks for building classification models due to their intuitive models and simple implementation. They learn fault patterns from historical data and identify equipment operating conditions. A model of convolutional gated cyclic attention is proposed in [11], with a CNN-BiGRU-CBAM and SLR. A recurrent neural network (RNN) model is constructed for fault diagnosis, which is subsequently enhanced with a fast continuous detection improvement and probability-based correction module [12]. A graph structure is dynamically constructed based on the Transformer multi-head attention mechanism, whose effectiveness is validated using a graph convolutional network (GCN), graph attention network (GAT), and GraphSAGE [13]. Although DFDMs have the advantage of fast online computation, their prediction results have uncertainty, which makes it difficult to ensure the final precision of fault diagnosis. What is more, their performance may tend to degrade when distinguishing similar fault classes [14].

To address the above issues, a novel multi-source fault diagnosis strategy based on knowledge and data dual-drive, namely the MICM-SLR-KNN algorithm, is proposed to achieve intelligent fault diagnosis of the wind turbine planetary gearbox. The proposed algorithm combines the precision of KFDMs and rapidity of DFDMs. Moreover, a novel fault feature extraction matrix, namely MICM, is also designed by combining information among time domain, frequency domain and channel correlation. The main contributions of the paper can be summarized as follows.

- (1) Firstly, a novel hybrid classifier, namely SLR-KNN, is proposed, where the aim of SLR is only to predict fault classes with probability distributions while KNN is utilized to achieve precise classification. The proposed SLR-KNN can avoid the online low efficiency of KFDMs while overcoming the diagnostic uncertainty of DFDMs.
- (2) In the MICM-SLR-KNN algorithm, a novel fault feature extraction matrix, MICM, is constructed to enhance the expression of fault information, including time domain, frequency domain and channel correlation of vibration and speed signals.
- (3) Different from existing standard datasets, the WT-planetary gearbox dataset is collected from the practical experimental platform, including sampling authenticity, data scale, comprehensive coverage of operational frequencies and even interference caused by human assembly.

The article is organized as follows. First, Section 2 introduces the challenges of fault diagnosis for planetary gearbox. Section 3 provides the design of the proposed MICM-SLR-KNN algorithm. Then, Section 4 provides experimental verification and result analysis through the WT-planetary gearbox dataset. Finally, the conclusions are given in Section 5.

## 2. The Challenges of Fault Diagnosis for Planetary Gearbox

The planetary gearbox serves as a critical transmission component in a wind turbine, as it exhibits significantly more complex dynamic behavior than the fixed-axis gearbox. As shown in Figure 1, the planetary gearbox consists of a fixed ring gear, a centrally rotating sun gear and multiple planetary gears. Each planetary gear is mounted on a rolling bearing, enabling simultaneous rotation about its own axis and revolution around the sun gear [15]. In such a configuration, the continuous surface contact between gears would give rise to friction, vibration and heat generation, which results in gear wear, breakage and other faults [9]. In addition, these issues can also be accelerated by harsh operating conditions such as high temperature, speed and load [10]. However, due to the complex structure

and dynamic features of the planetary gearbox, fault data captured by sensors have the following features [10,16,17] that bring difficulties and challenges to the fault diagnosis of a planetary gearbox.

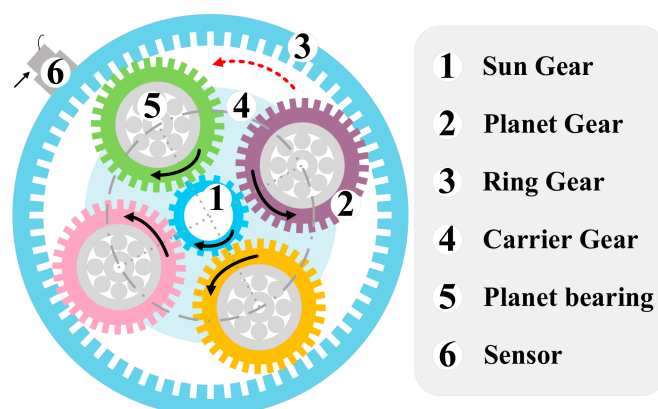


**Figure 1.** Structure of a planetary gearbox.

- (1) A planetary gearbox exhibits multiple meshing coupling points with inherent non-linear signal features. When localized faults occur, the signals manifest a partial imbalance due to the modified dynamic interactions. This is a significant challenge for fault diagnosis to effectively deal with such nonlinear and imbalanced signals.
- (2) The complex mechanical structure of the planetary gearbox leads to a multi-path vibration propagation mechanism. The meshing and propagation paths between planetary gears, ring gears and sun gears are complex. There are multiple paths between faults and sensors. This modulation phenomenon may lead to signal aliasing or attenuation, which increases the difficulty of reliable fault diagnosis.
- (3) The frequency energy associated with bearing fault features is significantly weaker than that generated by gear meshing phenomena. Such an inherent limitation is further exacerbated by the presence of industrial noise in the practical environment. This would make the identification and extraction of bearing fault features particularly challenging.

Given that gearbox faults may lead to downtime loss, significant maintenance costs and even potential safety incidents, it is essential to establish a rapid and efficient fault diagnosis method for planetary gearboxes to improve the stability and safety of wind turbines.

### 3. Design of the MICM-SLR-KNN Algorithm

The flowchart of the proposed MICM-SLR-KNN algorithm based on knowledge and data dual-drive is shown in Figure 2. The design process of the MICM-SLR-KNN algorithm mainly consists of two stages, namely offline training and online implementation.

In the offline training stage, the original fault data collected from multiple sensors undergo denoising by Kalman filter and normalization, which reduces the impact of Gaussian noise and dimensional inconsistency. Then, the MICM is established by time domain, frequency domain, and channel correlation analyses on the aforementioned normalized data samples. Afterwards, KPCA is adopted for MICM to reduce redundancy and capture fault features comprehensively. The feature reconstruction loss is defined to determine the output dimension of KPCA. Finally, these low-dimensional features are used to train SLR. After being given by SLR, the class probability would be transferred to KNN to achieve precise fault diagnosis and classification.

In the online implementation stage, the online-acquired data from sensors are processed through a Kalman filter with normalization. Then, the MICM is formed based on

normalized data. Then, the low-dimensional features can be obtained via the configured KPCA and fed into SLR-KNN to achieve precise fault diagnosis.

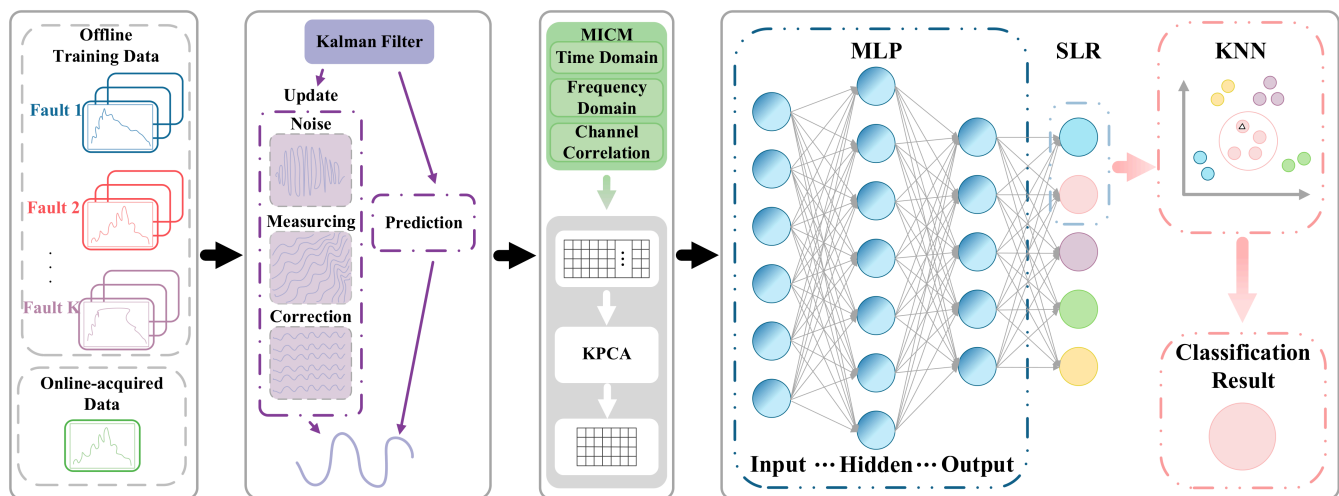


**Figure 2.** Flowchart of the MICM-SLR-KNN algorithm.

In the MICM-SLR-KNN algorithm, the MICM is established based on knowledge of fault features, while the novel hybrid classifier is designed via knowledge and data dual-drive. The specific descriptions are as follows.

### 3.1. Establishment of the MICM

The original fault data exhibit limited feature diversity, substantial redundant information and Gaussian noise. Consequently, the data are unsuitable to be used directly as fault features in the fault diagnosis. Aimed at comprehensively characterizing the fault information embedded within the fault data, domain knowledge is integrated with feature extraction to construct the MICM, including time domain, frequency domain and channel correlation. Hence, the MICM respectively reveals the variation pattern, energy distribution and information coupling of the fault data.

Time domain features in the MICM reflect the gear meshing dynamics and impact event, which are described as

$$X_p = \max(|x_i|), \quad (1)$$

$$C_p = \frac{X_p}{rms}, \quad (2)$$

$$C_{pm} = \frac{X_p}{\frac{1}{N} \sum_{i=1}^N |x_i|}, \quad (3)$$

$$C_{ps} = \frac{X_p}{\left(\frac{1}{N} \sum_{i=1}^N \sqrt{|x_i|}\right)^2}, \quad (4)$$

$$kurtosis = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4}, \quad (5)$$

$$skewness = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}, \quad (6)$$

where  $x_i$  is the channel fault data for feature extraction with root mean square  $rms$ , mean  $\mu$ , and standard deviation value  $\sigma$ , respectively. The peak value  $X_p$  captures the maximum impact amplitude. The crest factor  $C_p$  reflects the severity of impulsive components. The impulse factor  $C_{pm}$  can effectively detect sudden impulses even under very small impact conditions. The margin factor  $C_{ps}$  reflects how prominently the impulsive peaks stand out relative to the overall vibration level. The kurtosis calculates the presence of heavy-tailed and impact-dominated vibration. The skewness measures the asymmetry of the waveform.

The frequency domain features incorporate mechanical knowledge of meshing dynamics, where the energy distribution reveals fault features. The selected segment of the sensor fault data is transformed as the input data using FFT to obtain its frequency spectrum  $X(f_k)$ . The corresponding frequency sequence  $f_k$  is defined as

$$f_k = \frac{k f_s}{N}, k = 0, 1, \dots, \frac{N}{2} - 1, \quad (7)$$

where  $f_s$  is the sampling frequency and  $N$  is the data length.

The power spectral density (PSD) is used to describe the power distribution and is calculated as

$$P(f_k) = \frac{|X(f_k)|^2}{N}. \quad (8)$$

The frequency features considered in the paper are presented as

$$FC = \frac{\sum_{k=0}^{N/2-1} f_k P(f_k)}{\sum_{k=0}^{N/2-1} P(f_k)}, \quad (9)$$

$$MSF = \frac{\sum_{k=0}^{N/2-1} f_k^2 P(f_k)}{\sum_{k=0}^{N/2-1} P(f_k)}, \quad (10)$$

$$VF = \frac{\sum_{k=0}^{N/2-1} (f_k - FC)^2 P(f_k)}{\sum_{k=0}^{N/2-1} P(f_k)}, \quad (11)$$

where the frequency center  $FC$  describes the weighted average frequency of the spectrum. The mean square frequency  $MSF$  measures the concentration of energy in higher frequency bands. The frequency variance  $VF$  quantifies how widely the spectral energy is distributed around the average frequency.

During the operation of the wind turbine planetary gearbox, periodic impact vibration signals are generated when the fault occurs. They contain distinct pulse structures and characteristic frequencies in the vibration waveforms. However, due to the complex operating environment and strong Gaussian noise [18], the periodic features of the original fault data are often difficult to identify directly. The autocorrelation can capture periodic features and effectively suppress Gaussian noise [19], which can be calculated as

$$R_{xx}(\tau) = \sum_t (x(t) - \bar{x})(x(t + \tau) - \bar{x}). \quad (12)$$

To leverage the property, the autocorrelation peak is adopted as a fault feature and calculated as

$$P_{auto} = \max(R_{xx}(\tau)), \tau > 0. \quad (13)$$

The multi-channel fault data are inherently coupled [20]; therefore, the cross-correlation peak is selected as a key feature to exploit similarity between data [21], which can be obtained as

$$R_{xy}(\tau) = \sum_t (x(t) - \bar{x})(y(t + \tau) - \bar{y}), \quad (14)$$

$$P_{cross} = \max(R_{xy}(t)), \quad (15)$$

where  $x, y$  are fault data from different sensor channels.

In summary, the MICM contains rich fault features and provides a solid foundation for subsequent dimensionality reduction and classification.

### 3.2. Design of Hybrid Classifier Based on Knowledge and Data Dual-Drive

When the low-dimensional features are obtained after KPCA, reliable and effective classifiers become crucial. A novel hybrid classifier SLR-KNN is proposed here, where SLR is trained and established using offline training to achieve pre-classification with probability distributions, while KNN achieves precise classification through similarity calculation between samples and nearest neighbor discrimination rules.

#### 3.2.1. Pre-Classification of SLR

As the initial stage of the hybrid classifier, SLR partitions all input samples into probabilistically assigned subsets. The process establishes refined sample domains and reduces the search space for precise classification by KNN.

The SLR is implemented using a multi-layer perceptron (MLP). Given the low-dimensional feature vector  $m \in \mathbb{R}^d$ , the hidden layer applies a nonlinear transformation

$$h = \phi(W_1 m + b_1), \quad (16)$$

where  $\phi(\cdot)$  denotes an element-wise nonlinear mapping.  $h$  is the feature representation of the hidden layer.  $W$  and  $b$  are weight and bias, respectively.  $W_1$  and  $b_1$  are initialized by Kaiming.  $\phi(\cdot)$  employs RELU here.

The output layer then calculates the class logits

$$o = W_2 h + b_2, \quad (17)$$

where  $W_2$  and  $b_2$  are initialized by Xavier.

Let us substitute (17) into (18):

$$q(o_i) = \frac{e^{o_i}}{\sum_{i=1}^C e^{o_i}}, \quad (18)$$

where  $q(\cdot)$  transforms outputs of the output layer into probabilities between 0 and 1.  $C$  is the class number.

To measure the discrepancy between the predicted class and the truth class, the cross-entropy loss is adopted as the training objective

$$L(\Theta) = -\frac{1}{M} \sum_{i=0}^{M-1} \sum_{k=0}^{C-1} \varphi_k \log q(o_k), \quad (19)$$

where  $M$  is the number of fault samples.  $\varphi_k = 1$  when  $k$  is the true class and  $\varphi_k = 0$  otherwise.

During training, the algorithm parameters are iteratively updated using the AdamW optimizer to minimize (19). For each batch, parameter gradients are obtained via back-propagation, and the update rule is expressed as

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \hat{g}^{(t)}, \quad (20)$$

where  $\eta$  is the learning rate and  $\hat{g}^{(t)}$  denotes the bias-corrected gradient estimate provided by the optimizer.

The trained SLR outputs a probability vector for each

$$\hat{n} = \arg \max(q), \quad (21)$$

where  $\hat{n}$  is the class corresponding to the higher probability.

For each predicted class, the corresponding training subset

$$D_k = \{(m, z) | \hat{n} = k\} \quad (22)$$

is constructed to provide a more localized search space for KNN, where  $z$  is the true class of  $m$ .

A separate KNN is constructed for each  $D_k$  to store and manages its corresponding feature vectors.

### 3.2.2. Precise Classification by KNN

During online implementation for fault diagnosis, an online-acquired sample  $m^*$  is sent to the hybrid classifier after data preprocessing and feature extraction. It first passes through the trained SLR to obtain  $\hat{n}_k$  with a higher possibility. The sample is then sent into KNN and the Euclidean distance is calculated by:

$$d(m^*, m) = \|m^* - m\|_2, m \in D_k. \quad (23)$$

KNN prediction is achieved by:

$$\hat{z} = \arg \max_k \sum_{m \in \mathbb{N}(m)} 1(z = k) \quad (24)$$

where  $\mathbb{N}$  denotes the set of the  $K$  nearest neighbors of  $m^*$ .  $1(\cdot)$  is the indicator function returning 1 if true, and 0 otherwise.

If  $\hat{z} = \hat{n}_k$ , the final decision is  $\hat{z}$ . Otherwise,  $\hat{n}_{k'}$  corresponding to the second higher probability of SLR is obtained. The above steps are repeated to get a new prediction  $\hat{z}'$ . If  $\hat{z}' = \hat{n}_{k'}$ ,  $\hat{z}'$  is taken as the final output; otherwise,  $m^*$  is defined as an outlier.

The pseudocode for the MICM-SLR-KNN algorithm [Algorithm 1] is as follows.

---

#### Algorithm 1 Outline of SLR-KNN hybrid classifier for planetary gearbox fault diagnosis

---

##### Offline SLR training and separate KNN construction based on class probabilities

1. The input training fault samples  $[m, z]$ , class number  $C$ , epoch number  $E$  and learning rate  $\eta$  are obtained.
  2. construct Dataloader for training fault samples
  3. initialize MLP network using (16), (17)
  4. initialize AdamW with  $\eta$
  5.  $bestP \leftarrow 0$
  6. **for**  $epoch = 1$  to  $E$
  7.      $q$  is calculated by (18)
  8.      $q$  is fed into (19)
  9.     update parameters using (20)
  10.     $\hat{n}$  is obtained by (21)
  11.    compute validation precision  $P$
  12.    **if**  $P > bestP$
  13.        $bestP \leftarrow P$
  14.    **for** each training sample
  15.       save  $m$ ,  $\hat{n}$  and  $z$
  16.    **end**
-

**Algorithm 1** *Cont.*


---

```

17.           for  $k = 0$  to  $C - 1$ 
18.               if  $\hat{n} == k$ 
19.                   store  $[m, z]$  into  $D_k$ 
20.               endif
21.           end
22.       endif
23.   end
24.   construct separate KNN for  $D_k$ 
25.   The trained SLR and the separate KNN are the outputs.

```

---

**Online** Real-time classification of online-acquired fault sample using SLR-KNN hybrid classifier

```

26.   The online prediction receives an online-acquired fault sample  $[m^*, z]$ . The
      trained SLR and separate KNN for  $D_k$  are also obtained.
27.    $m^*$  undergoes the data preprocessing and feature extraction
28.    $\hat{n}$  is obtained by (16), (17), (18), (21)
29.    $m^*$  and  $m \in D_k$  are fed into (23)
30.    $\hat{z}$  is obtained by (24)
31.   if  $\hat{z} == \hat{n}_k$ 
32.       the final prediction is  $\hat{z}$ 
33.   elseif
34.        $\hat{n}'$  is obtained
35.       do steps 29 to 30 again and  $\hat{z}'$  is obtained
36.       if  $\hat{z}' == \hat{n}_{k'}$ 
37.           the final prediction is  $\hat{z}'$ 
38.       elseif
39.            $m^*$  is defined as an outlier
40.       endif
41.   endif
42.   The final classification result of  $m^*$  is output.

```

---

#### 4. Experimental Verification and Result Analysis

Different from existing standard fault datasets, many practical factors are accounted for in the WT-planetary gearbox dataset [22], such as sampling authenticity, data scale, comprehensive coverage of operational frequencies and interference caused by human assembly. Hence, such a WT-planetary gearbox dataset is used as the data source of the MICM-SLR-KNN algorithm, which can fully verify the reliability and effectiveness of the proposed algorithm in the practical experimental platform. It is openly available in GitHub at <https://github.com/Liudd-BJUT/WT-planetary-gearbox-dataset> (accessed on 23 July 2025) [22].

The WT-planetary gearbox platform is shown in Figure 3 [22]. Five typical pieces of fault information are obtained by sensors, including healthy state (N), broken tooth (B), worn tooth (W), root crack (R), and missing tooth (M). The multi-source information is recorded, including vibration signals and speed pulse signals. The gearbox operates at eight rotational speeds ranging from 20 to 55 Hz under two different load conditions, so as to construct experimental conditions with rich operating condition coverage. Figure 4 shows one period of vibration signals for fault B and W at a rotational speed of 55 under the same load condition. Moreover, the sampling time of each operating condition is more

than five minutes. Not only does this ensure sufficient signal size, it also preserves the integrity of the signal in the construction of the MICM.

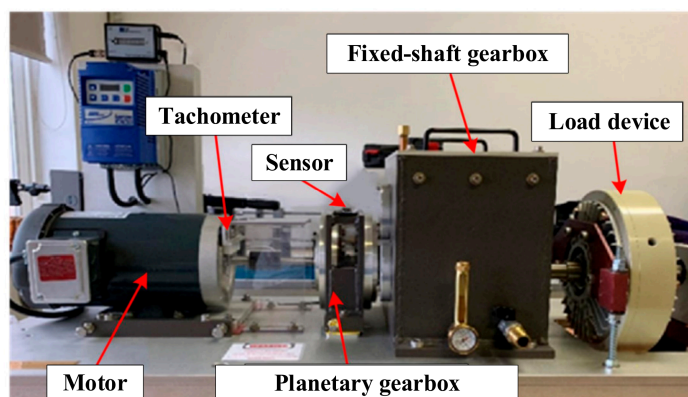


Figure 3. Experimental platform.

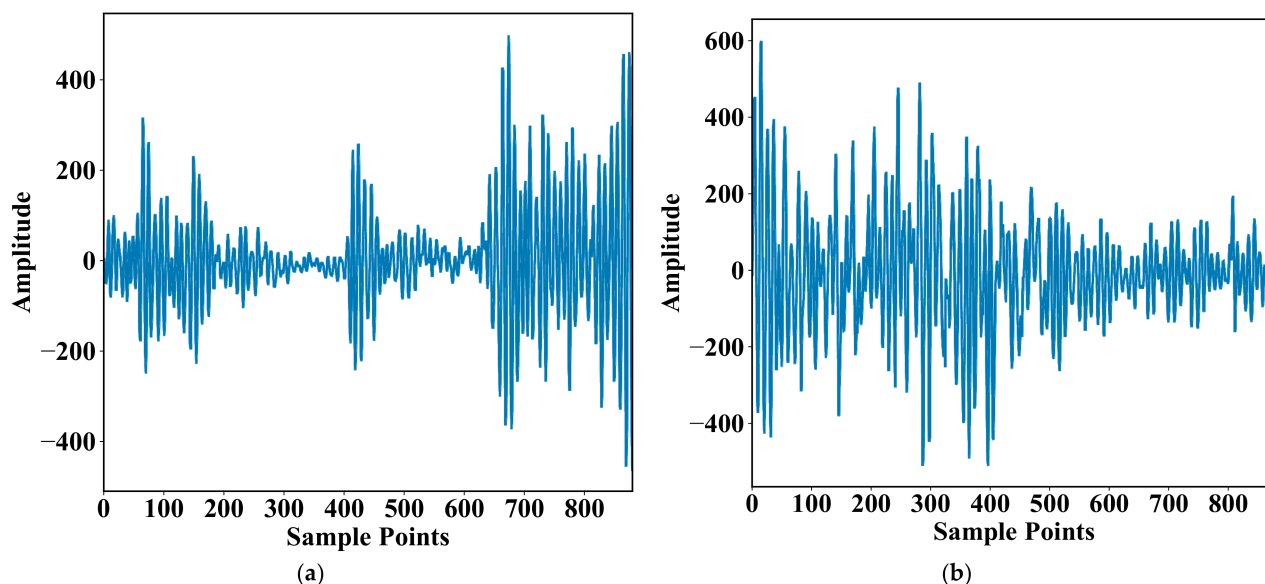


Figure 4. Representative vibration signals of different faults: (a) vibration signals of fault B; (b) vibration signals of fault W.

Two typical algorithms, namely PSAL-SLR [23] and DTW-KNN [24], are used for comparison to demonstrate the superiority of the proposed MICM-SLR-KNN algorithm. The precision  $P$  and recall  $R$  are adopted as criteria, and their expressions are defined as

$$P = \frac{TP}{TP + FP} \times 100\%, \quad (25)$$

$$R = \frac{TP}{TP + FN} \times 100\%, \quad (26)$$

where  $TP$ ,  $FP$ , and  $FN$  represent true positive, false positive and false negative, respectively. Their standard deviations,  $\sigma_P$  and  $\sigma_R$ , are also used as criteria to assess the stability of the method. The expression is expressed as

$$\sigma = \sqrt{\frac{1}{K-1} \sum_{i=1}^K (a_i - \bar{a})^2}, \quad (27)$$

where  $a_i$  is the  $P$  or  $R$  for each experimental operation.  $K$  represents the number of experimental operations.

To accurately analyze the contribution of features from different domains to the  $P$  and  $R$  of the MICM-SLR-KNN algorithm, ablation experiments are also conducted using only single domain features in the paper.

#### 4.1. Experimental Results of a Single-Operation Condition

Firstly, the experimental results are validated on a single-operation condition, where the rotational speed is 55 Hz. In this case, the process noise and measurement noise are set to 0.1 and 5, based on empirical observations, to balance noise suppression and fault information fidelity. The output dimension of KPCA is determined based on the feature reconstruction loss,

$$\varepsilon = 1 - \frac{\sum_{i=1}^s \lambda_i}{\sum_{i=1}^S \lambda_i}, \quad (28)$$

where  $\lambda$  is the eigenvalue.  $s$  is the output dimension of KPCA, and  $S$  is the original dimension of the MICM. It is set to 20, due to the fact that  $\varepsilon$  is limited within 0.03. The parameters of neural networks are selected through a trial and error method, such as the size of hidden layer and the learning rate. The number of neurons in the hidden layer is 16 and the learning rate is 0.1. A similar case is adopted for  $K$  in KNN, and the  $K$  in KNN is 1. In the PSAL-SLR algorithm, the size of the hidden layer and the learning rate are set to 128 and 0.00004, respectively.  $K$  in the DTW-KNN algorithm is set to 3. The experimental results of the MICM-SLR-KNN algorithm are shown in Figures 5 and 6. Experimental results of different algorithms are listed in Table 1. The relevant results of the ablation experiment are shown in Figure 7.

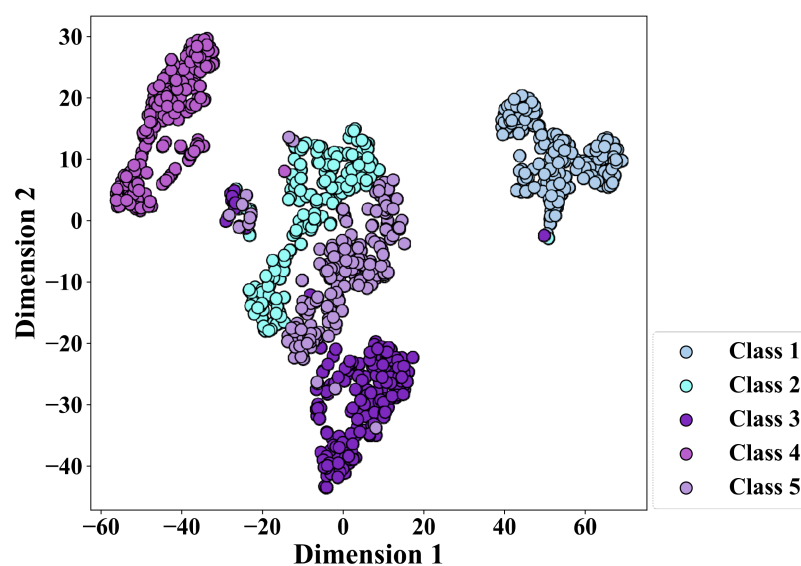


**Figure 5.** T-SNE visualization of low-dimensional features under the single-operation condition.

The t-SNE visualization of the features extracted by MICM-KPCA is shown in Figure 5.

From Figure 5, one can see that the feature space exhibits a clear separation among the five fault types.

The classification result of the MICM-SLR-KNN algorithm on the online-acquired data is shown in Figure 6.

From Figure 6, the MICM-SLR-KNN algorithm has good fault diagnosis ability. Most fault samples are predicted accurately.

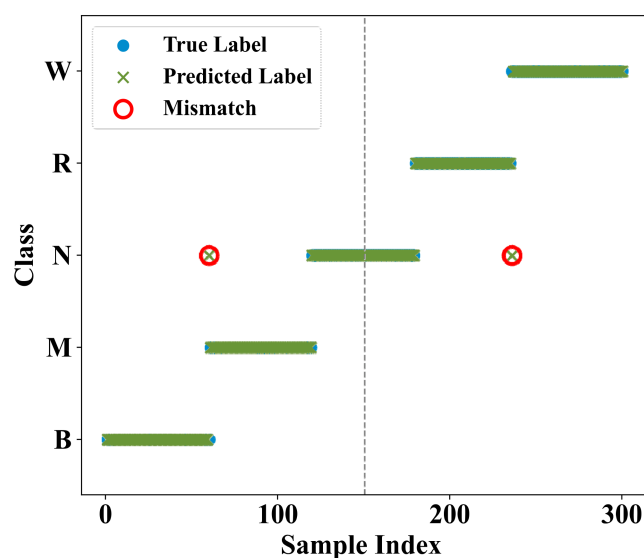


**Figure 6.** True and predicted labels for the online-acquired data ordered by class under the single-operation condition.

**Table 1.** Comparison of different algorithms under the single-operation condition.

|            | MICM-SLR-KNN | PSAL-SLR [23] | DTW-KNN [24]    |
|------------|--------------|---------------|-----------------|
| $T$        | 0.12 s       | 0.11 s        | 23.00 s         |
| $P$        | 99.53%       | 90.92%        | 92.00%          |
| $\sigma_P$ | 0.17%        | 4.18%         | -- <sup>1</sup> |
| $R$        | 99.53%       | 91.45%        | 92.73%          |
| $\sigma_R$ | 0.17%        | 4.40%         | -- <sup>1</sup> |

<sup>1</sup> DTW-KNN is a deterministic algorithm. When run multiple times with the same fault data,  $P$  and  $R$  are exactly the same. Its standard deviation is undefined and is 0.

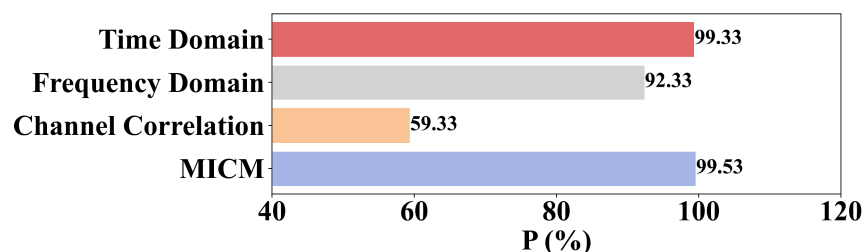


**Figure 7.** Ablation experiment of different domain features under the single-operation condition.

To fully demonstrate the effectiveness of the MICM-SLR-KNN algorithm and avoid randomness of the experimental results, the average  $P$  and  $R$  of the fifteen experimental operations serve as the final criteria and are listed in Table 1. The standard deviations  $\sigma_P$  and  $\sigma_R$  are also listed.  $T$  represents the time of a single period.

From Table 1, the proposed MICM-SLR-KNN algorithm demonstrates superior performance to the other two algorithms in terms of both computation and diagnostic effectiveness. Furthermore, the standard deviations of the precision and recall are relatively low, indicating the stability of the proposed algorithm.

The  $P$  of the MICM-SLR-KNN algorithm based on different single domain features and the MICM is shown in Figure 7, whose condition is single-operation.

From Figure 7, one can see that time domain and frequency domain features contribute significantly to the experimental results. Channel correlation features contribute moderately.

#### 4.2. Experimental Results of Multiple-Operation Condition

Due to the multiple operating frequencies and complex fault information, the fault diagnosis of the multiple-operation condition is more challenging. The rotational speeds are from 20 to 55 Hz. In this case, according to the method of setting parameters under the single-operation condition, the process noise and measurement noise are 0.1 and 50, respectively. The output dimension of KPCA is 70. The number of neurons in the hidden layer is 200 and the learning rate is 0.05. K is set to 1. The size of the hidden layer and the learning rate in the PSAL-SLR algorithm are set to 128 and 0.00004, respectively. K in the DTW-KNN algorithm is set to 3.

The relevant experimental results are presented in Figures 8 and 9. The comparison results of different algorithms are listed in Table 2. The relevant results of the ablation experiment are shown in Figure 10.

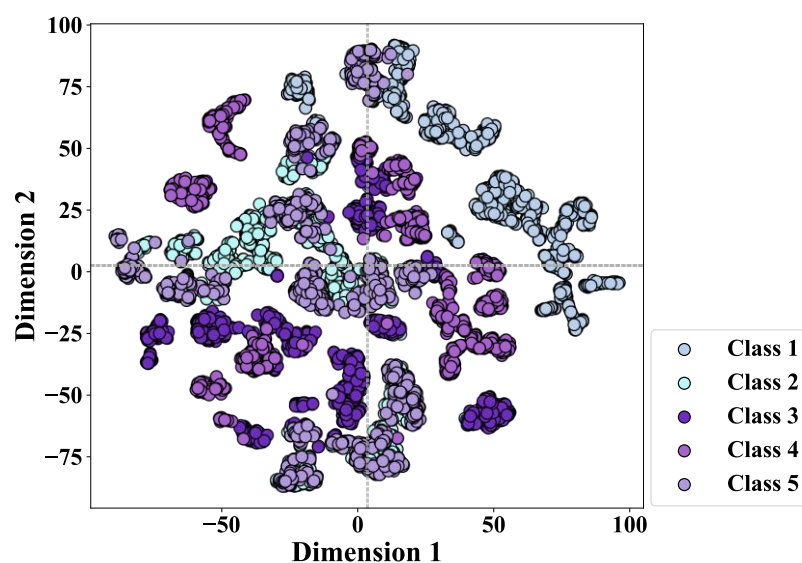


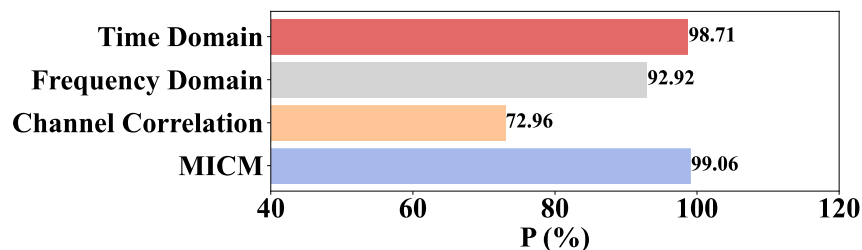
Figure 8. T-SNE visualization of low-dimensional features under multiple-operation conditions.

| True Label | B     | M               | N     | R      | W     |   |
|------------|-------|-----------------|-------|--------|-------|---|
|            | 99.6% | 0.0%            | 0.0%  | 0.0%   | 0.4%  |   |
|            | 0.2%  | 98.6%           | 0.0%  | 0.0%   | 1.2%  |   |
|            | 0.0%  | 0.2%            | 99.2% | 0.0%   | 0.6%  |   |
|            | 0.0%  | 0.0%            | 0.0%  | 100.0% | 0.0%  |   |
|            | 0.0%  | 1.9%            | 0.0%  | 0.0%   | 98.1% |   |
|            |       | B               | M     | N      | R     | W |
|            |       | Predicted Label |       |        |       |   |

Figure 9. Confusion matrix for the MICM-SLR-KNN algorithm under the multiple-operation condition.

**Table 2.** Comparison of different algorithms under multiple-operation conditions.

|            | MICM-SLR-KNN | PSAL-SLR [23] | DTW-KNN [24] |
|------------|--------------|---------------|--------------|
| $T$        | 0.38 s       | 0.31 s        | 184.24 s     |
| $P$        | 99.06%       | 88.39%        | 81.04%       |
| $\sigma_P$ | 0.17%        | 4.41%         | --           |
| $R$        | 99.06%       | 88.88%        | 81.53%       |
| $\sigma_R$ | 0.17%        | 4.43%         | --           |

**Figure 10.** Ablation experiment of different domain features under multiple-operation conditions.

The t-SNE visualization of the low-dimensional features is shown in Figure 8.

From Figure 8, it can be seen that the MICM-KPCA can still effectively group fault samples from identical faults within similar operating conditions, revealing its sensitivity to the condition and consistency of clustering.

The classification performance of the MICM-SLR-KNN algorithm is shown in Figure 9.

From Figure 9, the MICM-SLR-KNN algorithm has a good classification performance for each class of fault. The average results and standard deviations of fifteen experimental operations for three algorithms are listed in Table 2. The time to make an online prediction is also listed.

From Table 2, the MICM-SLR-KNN algorithm demonstrates superior fault diagnosis performance, with notably higher precision and recall compared to the PSAL-SLR and DTW-KNN algorithm. Additionally, it has significant advantages in online computation speed, which ensures online applicability performance. The low standard deviation indicates that the algorithm is stable, avoiding the randomness of the results.

Figure 10 shows the  $P$  of the MICM-SLR-KNN algorithm using single domain features under multiple-operation conditions.

From Figure 10, the time domain and frequency domain features can significantly impact the experimental results, while channel correlation features have only a moderate effect.

#### 4.3. Result Analysis

The time domain, frequency domain and channel correlation features are extracted from multi-source vibration and speed signals to construct the MICM. It comprehensively expresses multi-source information and allows for the effective separation of fault samples from different fault classes within the low-dimensional feature space of KPCA. The t-SNE visualizations of features extracted by MICM-KPCA are also shown in Figures 5 and 8.

In terms of the classifier, a novel SLR-KNN hybrid classifier is proposed to solve the uncertainty issue of SLR and the low efficiency of KNN. In the SLR-KNN, SLR is used to obtain fault classification with probability distributions, which provides the criterion of classification priority for KNN to reduce the time consumption of traversal searching. Meanwhile, KNN, as an effective classification algorithm, can be regarded as double-checked validation of SLR, which solves the uncertainty of SLR classification results.

The results of fault diagnosis for proposed MICM-SLR-KNN algorithm can be seen in Figures 6, 7, 9 and 10, and Tables 1 and 2.

The online computation time of the MICM-SLR-KNN algorithm is slightly increased in comparison with existing DFDMs. However, it would not affect the online applicability of the MICM-SLR-KNN algorithm, and the related results as shown in Tables 1 and 2.

In the experimental simulation section, a typical dataset from experiments under single-operation conditions and multiple-operation conditions, from the practical WT-planetary gearbox shown in Figure 3, is assembled to further validate the superiority of the MICM-SLR-KNN algorithm. Such a dataset contains a rich variety of fault types, all of which are practically and directly collected by sensors. Hence, the practical WT-planetary gearbox dataset is also crucial for verifying the authenticity and effectiveness of the algorithms. Compared to the PSAL-SLR and DTW-KNN algorithms, the proposed MICM-SLR-KNN algorithm achieves better results, which further proves the superiority of the MICM-SLR-KNN algorithm in practical engineering.

## 5. Conclusions

In this manuscript, a novel MICM-SLR-KNN algorithm based on knowledge and data dual-drive is proposed to achieve the precise fault diagnosis of a practical WT-planetary gearbox. Unlike existing KFDMs and DFDMs, SLR of the proposed algorithm only lists possible classifications, while KNN is used to achieve precise classification by verifying two classes with the higher possibility. The proposed algorithm effectively mitigates the uncertainty of SLR accuracy while overcoming the inefficiency of KNN. Moreover, the fault information enhancement matrix, MICM, is designed to further improve the extraction of optimal fault features. Experimental simulations have verified the superiority and online applicability of the MICM-SLR-KNN algorithm. In the future, the MICM-SLR-KNN algorithm would be validated on larger and more extensive datasets and developed for practical systems.

**Author Contributions:** Methodology, J.H.; software, H.Y.; validation, J.H.; formal analysis, H.Y.; investigation, H.Y.; resources, J.H.; data curation, H.Y.; writing—original draft, H.Y.; writing—review and editing, J.H.; visualization, H.Y.; project administration, J.H.; and funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Research Foundation Project of Jiangsu University of Science and Technology, grant number 1032932204. The APC was funded by Jiangsu University of Science and Technology.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original data presented in the study are openly available in GitHub at <https://github.com/Liudd-BJUT/WT-planetary-gearbox-dataset> (accessed on 23 July 2025).

**Acknowledgments:** This work was supported by the Research Foundation Project of Jiangsu University of Science and Technology. The authors express gratitude to Jiangsu University of Science and Technology.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|      |                                             |
|------|---------------------------------------------|
| MICM | Multi-source information correlation matrix |
| KPCA | Kernel principal component analysis         |

|      |                                        |
|------|----------------------------------------|
| SLR  | Softmax logical regression             |
| KNN  | K nearest neighbor                     |
| KFDM | Knowledge-based fault diagnosis method |
| DFDM | Data-driven fault diagnosis method     |
| RNN  | Recurrent neural network               |
| GCN  | Graph convolutional network            |
| GAT  | Graph attention network                |
| PSD  | Power spectral density                 |
| MLP  | Multi-layer perceptron                 |
| N    | Healthy state                          |
| B    | Broken tooth                           |
| W    | Worn tooth                             |
| R    | Root crack                             |
| M    | Missing tooth                          |

## References

1. Paul, S.; Dey, T.; Saha, P.; Dey, S.; Sen, R. Review on the Development Scenario of Renewable Energy in Different Country. In Proceedings of the 2021 Innovations in Energy Management and Renewable Resources, Kolkata, India, 5–7 February 2021.
2. Tiismus, H.; Maask, V.; Astapov, V.; Korōtko, T.; Rosin, A. State-of-the-Art Review of Emerging Trends in Renewable Energy Generation Technologies. *IEEE Access* **2025**, *13*, 10820–10843. [\[CrossRef\]](#)
3. Musa-Basheer, H.O.; Salami, M.J.; Monye, S.I.; Avwunuketa, A.A. Renewable Wind Energy: A Review of Control Strategies for Renewable Wind Energy. In Proceedings of the 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals, Omu-Aran, Nigeria, 2–4 April 2024.
4. Taherian-Fard, E.; Sahebi, R.; Niknam, T.; Izadian, A.; Shasadeghi, M. Wind Turbine Drivetrain Technologies. *IEEE Trans. Ind. Appl.* **2020**, *56*, 1729–1741. [\[CrossRef\]](#)
5. Zhao, L.; Yan, D.; Wang, M.; Tang, J.; Chai, T. Recognition of Mill Load with KPCA and KNN Based on Shell Vibration Signals. In Proceedings of the 2011 3rd International Workshop on Intelligent Systems and Applications, Wuhan, China, 28–29 May 2011.
6. Zhao, S.; Song, Q.; Wang, M.; Huang, X.; Cao, D.; Zhang, Q. An Early Fault Diagnosis Method of Rolling Element Bearings Based on MED, DFA, and Improved KNN. In Proceedings of the 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology, Shenzhen, China, 13–15 November 2020.
7. Pawlovsky, A.P. An Ensemble Based on Distances for a kNN Method for Heart Disease Diagnosis. In Proceedings of the 2018 International Conference on Electronics, Information, and Communication, Honolulu, HI, USA, 24–27 January 2018.
8. Han, P.; Liu, Z.; He, X.; Ding, S.X.; Zhou, D. Multi-Condition Fault Diagnosis of Dynamic Systems: A Survey, Insights, and Prospects. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 15728–15744. [\[CrossRef\]](#)
9. Basangar, S.; Tripathi, B.N. Literature Review on Fault Detection of Equipment using Machine Learning Techniques. In Proceedings of the 2020 International Conference on Computation, Automation and Knowledge Management, Dubai, United Arab Emirates, 9–10 January 2020.
10. Shao, Z.; Zhang, T.; Kosasih, B. Compound Faults Diagnosis in Wind Turbine Gearbox Based on Deep Learning Methods: A Review. In Proceedings of the 2024 Global Reliability and Prognostics and Health Management Conference, Beijing, China, 11–13 October 2024.
11. Liu, X.; Zhang, X.; Suo, Y.; Yang, M. Fault Diagnosis of Small Sample Wind Turbine Gearbox Based On T-CBC. In Proceedings of the 2024 10th International Conference on Systems and Informatics, Shanghai, China, 14–16 December 2024.
12. Li, W.; Liu, Y.; Li, Y.; Guo, F. Series Arc Fault Diagnosis and Line Selection Method Based on Recurrent Neural Network. *IEEE Access* **2020**, *8*, 177815–177822. [\[CrossRef\]](#)
13. Dong, X.; Dong, L.; Luo, Z.; Liu, S.; Liao, Z. Study on Graph Structure Construction Method Based on Transformer Multi-Head Attention Mechanism. In Proceedings of the 2025 5th International Conference on Mechanical, Electronics and Electrical and Automation Control, Chongqing, China, 9–11 May 2025.
14. Lv, Q.; Geng, L.; Cao, Z.; Cao, M.; Li, S.; Li, W.; Fu, G. Adaptive Sparse Softmax: An Effective and Efficient Softmax Variant. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 3148–3159. [\[CrossRef\]](#)
15. Liu, Y.; Gao, T.; Wang, D.; Wu, W. A Review of Order-Ratio-Based Fault Feature Extraction Techniques for Planetary Gearboxes. *IET Conf. Proc.* **2025**, *2024*, 1255–1263. [\[CrossRef\]](#)
16. Zhang, X.; Wang, L.; Miao, Q. Fault Diagnosis Techniques for Planetary Gearboxes Under Variable Conditions: A Review. In Proceedings of the 2016 Prognostics and System Health Management Conference, Chengdu, China, 19–21 October 2016.

17. Hu, M.; Wang, G.; Ma, K. Identification of Wind Turbine Gearbox Weak Compound Fault Based on Optimal Empirical Wavelet Transform. *Meas. Sci. Technol.* **2023**, *34*, 045003. [\[CrossRef\]](#)
18. Lu, S.-D.; Jiang, Y.-H.; Wu, C.-C.; Sian, H.-W. A High-Accuracy and Robust Diagnostic Tool for Gearbox Faults in Wind Turbines. *J. Eng.* **2024**, *2024*, e12411. [\[CrossRef\]](#)
19. Zhou, X.; Xiong, X. An Improved Auto-Correlation Algorithm and Its Application in Fetal Heart Rate Detection. In Proceedings of the 2009 2nd International Conference on Biomedical Engineering and Informatics, Tianjin, China, 17–19 October 2009.
20. Lei, Y.; Han, D.; Lin, J.; He, Z. Planetary Gearbox Fault Diagnosis Using an Adaptive Stochastic Resonance Method. *Mech. Syst. Signal Process.* **2013**, *38*, 113–124. [\[CrossRef\]](#)
21. Zeng, Y.; Zhang, R.; Zhang, H.; Qiao, S.; Huang, F.; Tian, Q.; Peng, Y. GCCNet: A Novel Network Leveraging Gated Cross-Correlation for Multi-View Classification. *IEEE Trans. Multimed.* **2025**, *27*, 1086–1099. [\[CrossRef\]](#)
22. Liu, D.; Cui, L.; Cheng, W. A Review on Deep Learning in Planetary Gearbox Health State Recognition: Methods, Applications, and Dataset Publication. *Meas. Sci. Technol.* **2023**, *35*, 012002. [\[CrossRef\]](#)
23. An, Y.; Zhang, K.; Liu, Q.; Chai, Y.; Huang, X. Rolling Bearing Fault Diagnosis Method Base on Periodic Sparse Attention and LSTM. *IEEE Sens. J.* **2022**, *22*, 12044–12053. [\[CrossRef\]](#)
24. Cai, Y.; Hao, L.; Zhou, Y.; Chen, J.; Hu, Q.; Duan, X.; Wang, G. Rotating Rectifier Fault Diagnosis of Nuclear Multiphase Brushless Excitation System Based on DTW Metric and kNN Classifier. *IEEE Trans. Power Electron.* **2023**, *38*, 10329–10343. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.