An Interpretable Wavelet Kolmogorov–Arnold Convolutional LSTM for Spatial-temporal Feature Extraction and Intelligent Fault Diagnosis

Junfan Chen[1], Tianfu Li[1,*], Jiang He[1] and Tao Liu[1]

1 Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, 650550, China

*Corresponding author: tianfu.li@kust.edu.cn

As industrial systems become increasingly complex, the significant research interest has been devoted to intelligent fault diagnosis approaches leveraging deep learning. However, existing methods still face two critical challenges in practical applications: 1) the extracted features often fail to maintain robustness in nonstationary conditions. 2) deep neural networks generally exhibit a black box nature, offering limited interpretability in their feature extraction process. To solve the above issues, an interpretable wavelet Kolmogorov–Arnold convolutional LSTM (WKAConvLSTM) is proposed, which mainly consists of two key components: 1) a wavelet Kolmogorov–Arnold kernel (WKAK) with learnable scale and translation parameters is designed and then embedded into convolutional layers to enable the extracted spatial features interpretable. 2) a multi-head attention enhanced LSTM (MA-LSTM) is proposed to effectively capture crucial temporal dependencies in sequential data. In order to verify its effectiveness, the proposed model is tested on bearing and gearbox datasets under complex conditions, including noise interference, nonstationary operating conditions, and data class imbalance. The experimental data demonstrate that it not only achieves superior diagnostic accuracy compared with the advanced baselines models, but also enhances interpretability of the extracted features.

Keywords: Intelligent fault diagnosis, Interpretability, Kolmogorov-Arnold Networks, LSTM

1. Introduction

With the fast-paced growth of industrial systems, how to efficiently supervise and diagnose equipment condition has gained significant importance. [1]. Prognostics and Health Management (PHM) system [2], as a key component in achieving intelligent operation and maintenance of equipment, can effectively reduce losses caused by equipment failures and shutdowns through the application of advanced technologies such as signal processing and artificial intelligence [3-5]. Therefore, how to improve the accuracy of intelligent fault diagnosis (IFD) is one of the research focuses in PHM systems [6, 7].

Currently, existing IFD methods are typically grouped into machine learning (ML)-driven methods [8] and deep learning (DL)-driven methods [9, 10]. The former belongs to shallow models, its performance depends heavily on feature engineering, making it difficult to apply to industrial big data mining, such as SVM [11] and random Forest [12]. The latter, through end-to-end learning, can adaptively extract fault features from industrial big data, and therefore has become the mainstream IFD method, e.g., graph neural networks and convolutional neural networks [13, 14].

Although deep learning-based IFD methods have demonstrated significant effectiveness, it still faces two major challenges: 1) the complex operating conditions of the equipment, with continuously changing speed and load, hinder the IFD method to effectively extract robust fault features, resulting in poor actual diagnostic results; 2) The black box property of deep neural networks hinders understanding of how the model extracts features, resulting in a lack of credibility in the diagnostic results.

To achieve robust fault feature extraction, some works propose extracting the spatial-temporal features of monitoring signals from both the temporal and spatial dimensions [15]. For example, Bao et al. [16] utilized a graph convolutional network to model the spatial features in the multivariate data, followed by a sliding window for temporal feature extraction. Li et al. [17] leveraged kernel principal component analysis for modeling spatial features and employed complementary ensemble empirical mode decomposition to characterize temporal dependencies. However, the above methods still belong to the black box model and cannot solve the problem of poor interpretability in IFD methods.

In recent years, scholars in the field of IFD have been committed to making the decision-making mechanism of deep learning models transparent, and gradually formed the idea of using fault diagnosis domain knowledge to guide the construction of deep networks [18]. The pioneer work is WaveletKernelNet [19], which combined the knowledge of continuous wavelet transform into convolutional neural network, thus designing a specific wavelet kernel convolution for multi-scale feature extraction. Wang et al. [20] Wang et al. combined the discrete wavelet transform with a deep neural network for hierarchical frequency analysis and interpretable feature extraction. The above research offers a potential avenue for enhancing model interpretability, however, how to extract interpretable robust fault features remains an open topic.

To solve these issues, this paper proposes an interpretable wavelet Kolmogorov–Arnold convolutional LSTM (WKAConvLSTM) for spatial-temporal feature extraction and intelligent fault diagnosis. The proposed WKAConvLSTM mainly consists of two parts, that is, the wavelet Kolmogorov–Arnold convolutional layer (WKAConv) for interpretable spatial feature extraction, and the multi-head attention enhanced LSTM (MA-LSTM) for temporal feature extraction. In WKAConv layer, the wavelet basis function is constructed into a wavelet Kolmogorov–Arnold kernel (WKAK) with learnable scale factor and translation factor through the Kolmogorov–Arnold representation theorem, and then embedded into the traditional convolution layer for interpretable feature extraction. MA-LSTM leverages a multi-head attention mechanism to perform weighted fusion on the features extracted by LSTM to fully mine the important temporal features in sequence data. The proposed approach is assessed using three datasets, with the findings demonstrating its effectiveness and superiority, and the primary contributions of this study are summarized as follows:

1) The WKAK with learnable scale and translation parameters is designed and embedded into the conventional convolution layer for interpretable spatial feature extraction.

2) MA-LSTM is proposed to effectively capture crucial temporal features in sequential data by performing weighted fusion of LSTM outputs across multiple attention subspaces.

3) WKAConvLSTM is constructed to extract robust spatial-temporal features through fusing interpretable spatial feature extraction with attention-guided temporal modeling.

The remainder of this paper is laid out as follows: Section 2 briefly reviews the related works. Section 3 elaborates the proposed methods and its key components. Section 4 describes the experiments. Section 5 discusses the effects of each model module and the choice of mother wavelet on performance. And the propose method's interpretability is analyzed in Section 6. Finally, Section 7 concludes the paper.

2. Related Works

2.1. Spatial-temporal Feature Extraction Methods

At present, the spatial-temporal feature extraction methods used for robust feature mining are mostly two-stage methods [21]. That is, in the first stage, the convolutional operation or graph convolutional operation is leveraged to capture the spatial features of the monitoring signal, Then, in the second stage, recurrent neural networks, such as LSTM, GRU and other models, are used to model the temporal features. Such as Zhao et al. [22] used an adaptive multiscale CNN to capture intricate spatial feature, and a highway LSTM to model global temporal dependencies. Singh et al. [23] leveraged a graph attention network to extract spatial features of multisensory data, and then used a LSTM to models temporal patterns. Although this two-stage spatiotemporal feature extraction can effectively obtain robust fault features, the interpretability of the extracted features is still unclear.

2.2. Interpretable Intelligent Fault Diagnosis Methods

Interpretable IFD methods are currently generally distinguished as ante-hoc and post-hoc interpretable approaches [18]. The ante-hoc interpretable approaches introduce physical models or other priors to constrain the IFD model's construction and learning process, thereby improving the model's self-interpretability. Such as, WPConvNet [24] incorporates the wavelet packet transform to the convolutional layer, while some research constructed a earnable wavelet operator [25] for fault feature extraction. LGSC-Net [26] embeds the sparse coding optimization algorithm into the network structure to achieve ante-hoc interpretability and noise-robustness. In the contrary, the post-hoc interpretable methods require constructing additional models or techniques to explain the learning process of the IFD model, such as Grad-CAM [27], Shapley-value [28]. Although interpretable techniques can help open the black box of IFD methods, how to extract robust fault features with interpretability requires further research.

3. Proposed Method

As depicted in Fig. 1, WKAConvLSTM integrates a WKAConv layer to extract spatial features and an MA-LSTM layer to model temporal dependencies. In the following subsections, the WKAConv layer and MA-LSTM module will be introduced in detail.

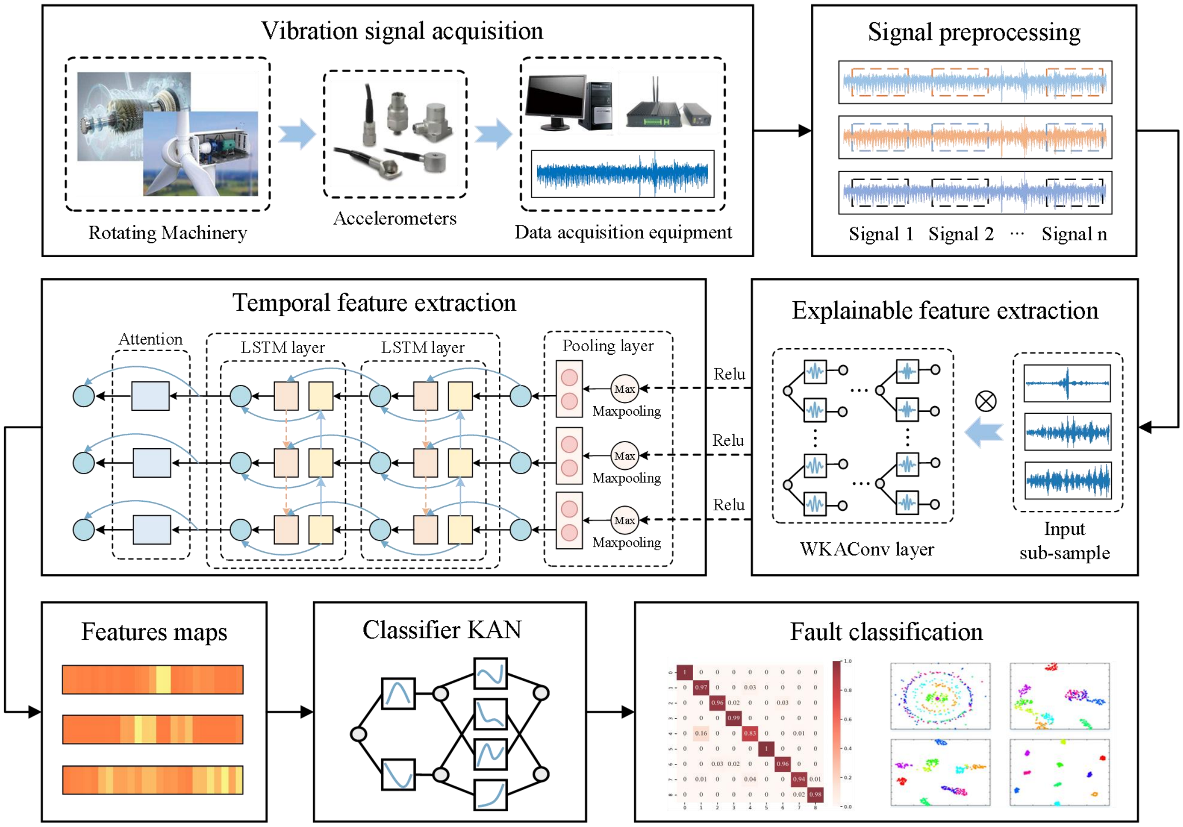


Fig. 1. The proposed WKAConvLSTM for intelligent fault diagnosis.

3.1. Wavelet Kolmogorov–Arnold Convolutional Layer

The traditional convolution operation can be defined as:

$$h = w * x + b \qquad (1)$$

where w is convolutional kernel, x is input, b is bias and h is the output.

If we understand convolution operations from the perspective of inner product matching, we can conclude that the reason why traditional convolution operations lack interpretability is because the randomly initialized convolution kernels lack physical meaning, which makes it difficult to extract interpretable fault features from vibration signals [19].

In order to overcome the above difficulties, we use the Kolmogorov–Arnold representation theorem to construct the traditional wavelet basis function into a wavelet Kolmogorov–Arnold kernel (WKAK) with learnable a scale factor and b translation factors [29]. Where the wavelet basis function ψ can be expand to the wavelet dictionary with the scale factor a and translation factor b, that is

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \qquad (2)$$

where $\psi_{a,\,b}$ denotes the wavelet dictionary, a determines the dilation or compression of the wavelet, thereby affecting its frequency resolution, while b controls the central position of the wavelet on the time axis, which is used to locate signal's local features.

After that, with the help of Kolmogorov–Arnold representation theorem, every continuous function may be approximated by an inner function and an outer function, which is defined as

$$f(t)=\sum_{q=1}^{2n+1}\Phi_q\left(\sum_{p=1}^{n}\phi_{q,p}\left(t_p\right)\right) \tag{3}$$

where Φ_q is the outer function and $\phi_{q,p}$ is the inner function. Furthermore, as depicted in Fig. 2, if we replace the inner function and the outer function with one learnable function $\psi_{a,b}$, thereby, the WKAK can be constructed, that is:

$$\mathrm{WKAK}(t)=\psi_L\circ\psi_{L-1}\circ\cdots\circ\psi_1(t) \tag{4}$$

where ψ_L denotes the L-th layer learnable wavelet dictionary, which means we can use several wavelet dictionaries to approximate the WKAK.

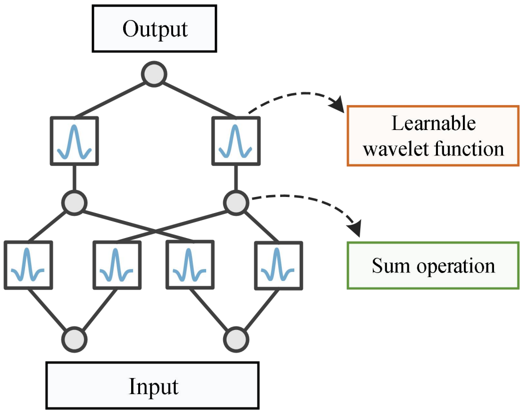


Fig. 2. Detailed structure of WKAK.

Then, by replacing the traditional convolution kernel with the learnable WKAK, the wavelet Kolmogorov–Arnold convolutional (WKAConv) layer can be defined as

$$h=\psi_L\circ\psi_{L-1}\circ\cdots\circ\psi_1 * x+b \tag{5}$$

As can be seen in (5), when the number of learnable functions grows, the WKAConv layer experiences a sharp rise in computational burden, therefore, only two learnable wavelet dictionaries are used.

3.2. MA-LSTM Layer

In a standard LSTM structure, information flow is regulated through three gates, namely the forget, input, and output gates [30], and its core calculation can be briefly expressed as

$$o_t=\sigma\left(W_o\cdot\left[h_{t-1},x_t\right]+b_o\right) \tag{6}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\tanh(W_C\cdot\left[h_{t-1},x_t\right]+b_C) \tag{7}$$

$$h_t=o_t\odot\tanh(c_t) \tag{8}$$

where x_t is the current input, while h_t and c_t denotes the hidden state and the memory cell, respectively. $\sigma(\cdot)$ indicates the sigmoid activation function, and $\odot$ denotes point-by-point multiplication. Although LSTM effectively captures temporal dependencies, it struggles to distinguish the relative importance of information across different time steps.

To overcome the above issue, the conventional LSTM is augmented with a multi-head attention module, thereby improving its ability to capture important features within time series. The central idea is to enable the model focus on diverse representation subspaces in parallel through multiple attention heads, defined as:

$$\mathrm{Attn}(Q,K,V)=\mathrm{softmax}\left(\frac{QK}{\sqrt{d_k}}\right)V \tag{9}$$

where Q, K, V denote query, key, and value matrices projected from input sequence, respectively. Then, the multi-head attention expands the above process to k parallel heads, that is:

$$\text{head}_i = \text{Attn}\left(HW_i^Q, HW_i^K, HW_i^V\right) \quad (10)$$

$$\text{MAttn}(H) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_k\right)W^O \quad (11)$$

where $H = [h_1, h_2, \ldots, h_3]$ denotes the input sequence of the multi-head attention, W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d \times d_k}$ are learnable weights. In MA-LSTM layer, the hidden state M of LSTM is dynamically weighted through the multi-head attention mechanism to strengthen the discriminative contribution of key time steps in the overall sequence representation, and the weighted results are fused with the original hidden state to improve the feature expression ability. The above process can be defined as:

$$M_{\text{attn}} = \text{MAttn}(M) \odot M \quad (12)$$

$$M_{\text{final}} = \text{Concat}(M, M_{\text{attn}}) \quad (13)$$

During the training phase, the cross-entropy loss $L(r, p)$ is adopted to quantify the discrepancy of the predicted probability distribution p from the ground truth labels r, formulated as:

$$L(r, p) = -\sum_{i=1}^{n} r(x_i) \log p(x_i) \quad (14)$$

where i denotes the number of categories.

3.3. WKAConvLSTM based Intelligent Fault Diagnosis

The intelligent fault diagnosis framework based on WKAConvLSTM as depicted in Fig. 1. To ensure comparability across different measurements, the vibration data obtained from sensors are first normalized, mapping their amplitude values into the interval of [0,1]. Subsequently, the continuous vibration signal is divided into fixed-length segments via a non-overlapping sliding window, and these sub-samples are provided as input to the model. In the training phase, spatial-temporal features are extracted, and then supplied to a classification module for fault recognition and categorization. The above process is summarized into Algorithm I.

Algorithm I: WKAConvLSTM for Fault Diagnosis

```
Input: vibration signal $X = [x_1, x_2, \ldots, x_n]$, fault label $y_1, y_2, \ldots, y_n$
Output: Machine fault type $\bar{y}$ and its diagnostic accuracy.
Data preparation:
  1) Min-Max normalization.
  2) Subsample generation.
  3) Random data splitting strategy.
Model training:
  1) $\bar{y}_{\text{train}} \leftarrow$ WKAConvLSTM ($X_{\text{train}}$).
  2) Updated iteratively by using the
     Adam optimizer to minimize (14).
Model validation:
  $\bar{y}_{\text{test}} \leftarrow$ WKAConvLSTM($X_{\text{test}}$)
```

4. Experiments

To assess the effectiveness of the proposed WKAConvLSTM, a series of experiments are performed across three widely used benchmark datasets, where the first and third dataset is obtained from a bearing fault simulation experiment and the second dataset is collected from a gearbox system operating under variable conditions. All experiments are implemented on Windows 11 with an AMD R9 7945HX CPU and an RTX4060 GPU.

4.1. Fault Diagnosis under Noisy Conditions

In this experiment, rolling bearing dataset is leveraged for model verification which is collected from a fault simulation test rig for rotating machines operated by Shandong University of Science and Technology [31], as depicted in Fig. 3. The dataset is primarily composed of ten bearing states: normal condition (NC), inner-race fault (IF), outer-race fault (OF), and rolling element fault (RF), where the last three fault categories is categorized into three severity levels (slight, medium, and severe) to reflect different damage scales, as shown in Table 1.

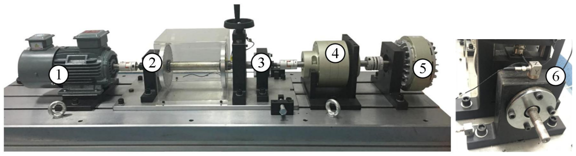


Fig. 3. Planetary gearbox fault simulation test rig.

Table. 1. Detail description of SDUST bearing dataset.

Fault	Description	Label
NC	No fault in rolling bearing	0
IF	Inner-race defect (0.2 mm)	1
	Inner-race defect (0.4 mm)	2
	Inner-race defect (0.6 mm)	3
OF	Outer-race defect (0.2mm)	4
	Outer-race defect (0.4mm)	5
	Outer-race defect (0.6mm)	6
RF	Fault located on the rolling element (0.2 mm)	7
	Fault located on the rolling element (0.4 mm)	8
	Fault located on the rolling element (0.6 mm)	9

For data preparation, Gaussian noise with signal-to-noise ratios (SNR) between 0 and -5 dB is injected into the original signal to simulate different degrees of noise interference, which can be defined:

$$SNR = 10\lg\left(\frac{P_{signal}}{P_{noise}}\right) \tag{15}$$

where P_{signal} denotes the signal power, P_{noise} is the noise power. As the SNR levels decreases, the noise power gradually increases. When the SNR drops to -5 dB, the noise amplitude exceeds three times that of the signal.

After that, the noise signals are mapped to the interval of [0, 1] through using the min-max normalization method, and then a sliding window is applied without overlap to divide the signals into segments of 1024 points each. For every health status, 820 sub-samples are generated, with 80% (656 samples) allocated to the training datasets and the remaining 20% (164 samples) to the testing datasets. This results in 6,560 training samples and 1,640 test samples in total.

To confirm the performance of WKAConvLSTM, five existing advanced baseline models are leveraged for comparison in our research, including CNN [32], KANConv [33], xLSTM [34], CNN-LSTM [21] and ConvFormer [35]. For fairness, all comparison models are designed with comparable network depth, and their detail description are depicted in Table 2.

Table 2. The detail description of the comparison models.

Model	Description
CNN	Extract fault features using convolution operations.
KANConv	Use learnable functions instead of fixed kernels to extract nonlinear features.
xLSTM	An LSTM variant, can enhance the performance of sequence modeling.
CNN-LSTM	Combine CNN for feature extraction and LSTM for sequence modeling.
ConvFormer	Extract local features

through convolution and integrate global information via the Transformer encoder.

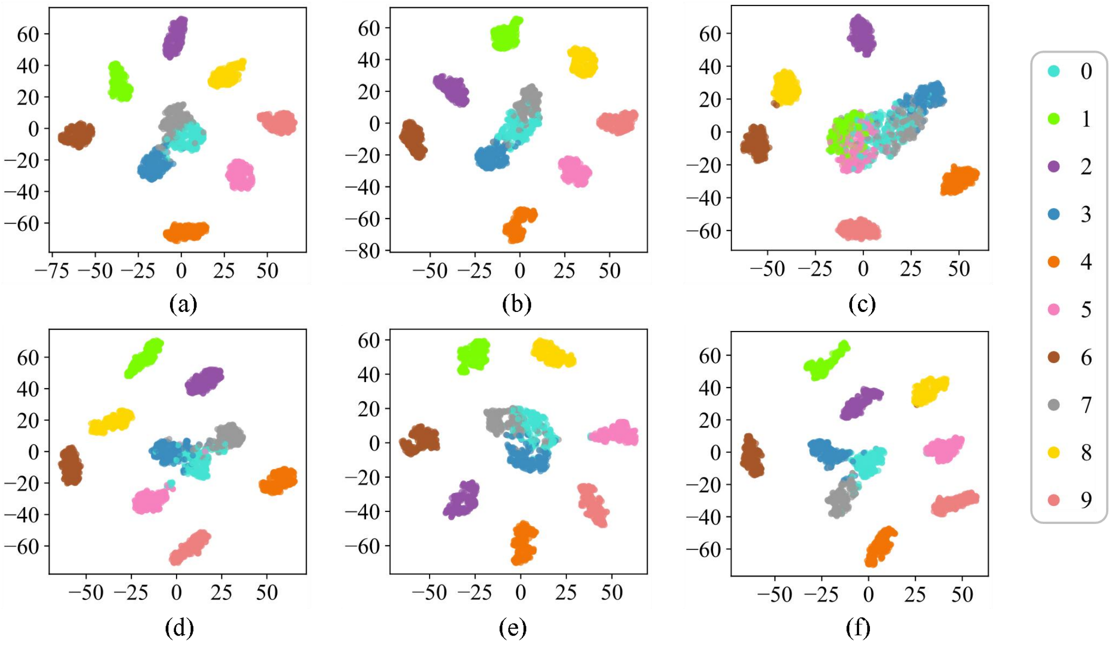


Fig. 4. The t-SNE visualizations of all models under the condition of SNR = -5dB. (a) CNN. (b) KANConv. (c) xLSTM. (d) CNN-LSTM. (e) ConvFormer. (f) WKAConvLSTM.

Table 3. The average accuracy (%) under Gaussian noise.

Models	SNR Level (dB)						Noise free
	-5	-4	-3	-2	-1	0	
CNN	95.99	97.83	98.57	98.60	98.83	99.09	99.81
KANConv	96.30	97.93	98.37	98.81	99.16	99.45	99.87
xLSTM	73.68	75.43	78.69	79.56	81.10	83.56	97.42
CNN-LSTM	94.61	96.61	97.83	98.28	98.65	99.37	99.74
ConvFormer	95.17	96.93	97.32	97.11	97.41	97.92	99.71
WKAConvLSTM	**97.42**	**98.81**	**99.33**	**99.54**	**99.74**	**99.83**	**99.95**

During the experiment, all models are trained for 100 epochs with Adam optimization, initialized at a learning rate of 0.001. To mitigate the influence of randomness, the training and evaluation are repeated five times for each model, and the accuracy of the final epoch in each run is recorded. The average accuracy over the five runs is reported as the final evaluation metric. The experimental performance evaluation of all baseline models at various SNR levels are provided in Table 3.

As shown by the experiments, the WKAConvLSTM consistently outperforms baseline models, exhibiting higher classification accuracy and greater robustness under diverse SNR scenarios. Notably, under moderate noise levels (e.g., -1dB, -3dB and -5dB), the model maintains performance comparable to that in noise-free environments. Even in the presence of severe Gaussian noise at -5 dB, the classification accuracy drops by only 2.53%, while still exceeding the lowest-performing baseline by a substantial margin of 23.74%.

To intuitively demonstrate feature extraction differences between the proposed method and baseline models, t-SNE is leveraged to display the feature distribution of the test dataset from the last classification layer. It can be seen from Fig. 4 that the WKAConvLSTM reaches the most compact clustering among all baseline models, indicating its superior capability in noise-robust and discriminative feature learning. In contrast, other models exhibit varying feature overlap, with xLSTM showing the most severe class confusion. While the majority of categories are well separated by the proposed method, boundaries between a few classes (i.e. #0, #3, and #7) appear relatively close. This may be attributed to Gaussian noise reducing inter-class distinctions, causing confusion in distinguishing these specific signals.

4.2. Fault Diagnosis under Nonstationary Condition

The MCC5-THU gearbox dataset [36] is collected by Tsinghua University under non-stationary conditions, with time-varying speed and load, as illustrated in Fig. 5.

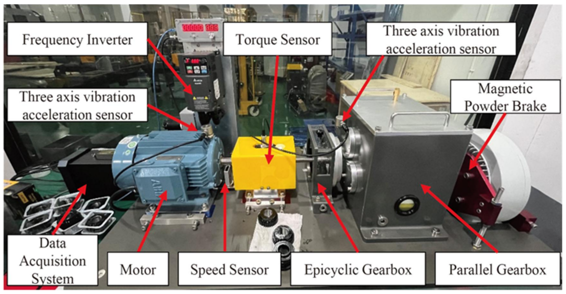


Fig. 5. The gearbox test rig.

This dataset provides vibration signals measured along the X, Y, and Z axes, covering normal condition, tooth-related defects (missing, worn, pitted, cracked, or broken) and compound fault of gear fracture combined with inner or outer ring bearing failure. Except for the normal condition and broken tooth fault, the remaining four single faults and two compound faults are classified into three levels of severity (slight, medium, and severe), with a total of 20 classification tasks. In this experiment, the signal in X directions with a rotation speed of 3000rpm and a load of 0~10A is used to evaluate the proposed model's adaptability and robustness in complex operating conditions.

Table 4. The diagnostic accuracy (%) of Gearbox under nonstationary conditions.

Model	Min	Max	Avg
CNN	85.28	90.29	87.68
KANConv	86.35	92.62	90.13
xLSTM	78.00	81.71	80.61
CNN-LSTM	82.58	85.35	84.25
ConvFormer	81.74	86.65	83.93
WKAConvLSTM	**90.35**	**92.76**	**91.74**

In this experiment, the data preprocessing method, hyperparameters and comparison models as in the previous experiments are adopted. From the dataset, 11,984 sub-samples are generated, with 9,587 assigned to the training dataset and 2,397 to the testing dataset. The outcomes of the experiments are presented in Table 4 and the accuracy of each trial of the six models as depicted in Fig. 6.

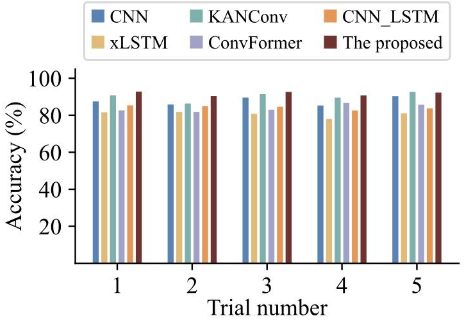
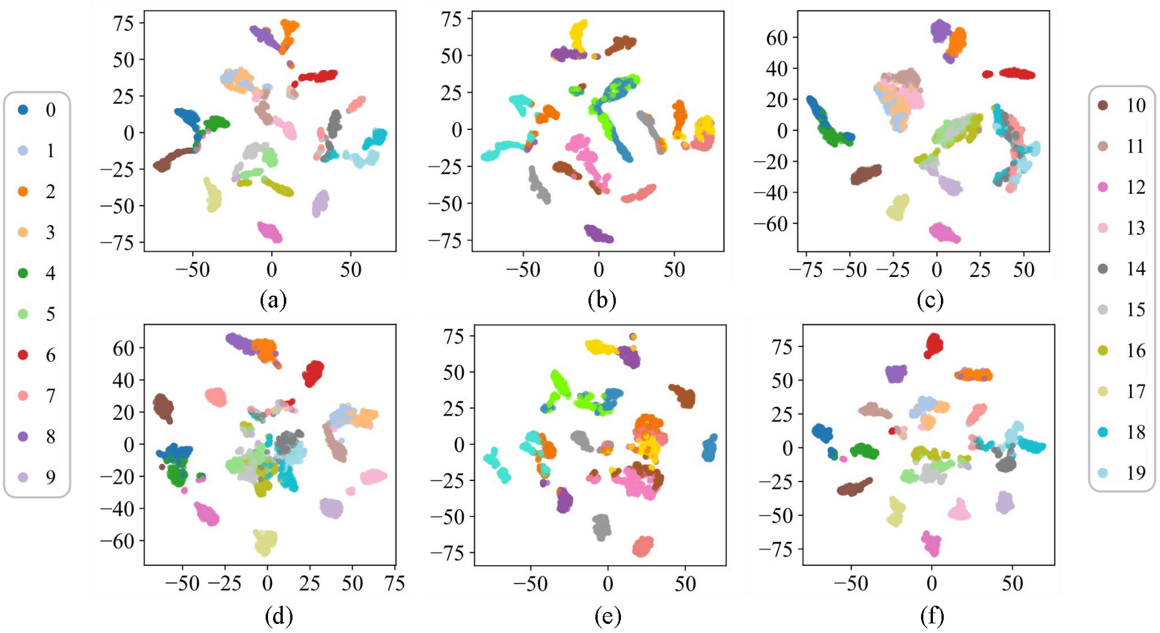


Fig. 6. Diagnostic accuracy of gearbox of each trial.



(a) (b) (c)

(d) (e) (f)

Fig. 7 The t-SNE visualizations of all models for gearbox fault diagnosis. (a) CNN. (b) KANConv. (c) xLSTM. (d) CNN-LSTM. (e) ConvFormer. (f) WKAConvLSTM.

Table 5. The details of the three datasets.

Health status	Label	Proportion of training samples			Proportion of testing samples
		Dataset 1	Dataset 2	Dataset 3	
Normal	0	50%	50%	50%	50%
Severe inner	1	50%	30%	25%	50%
Severe outer	2	50%	30%	20%	50%
Severe ball	3	50%	20%	15%	50%
Severe combo	4	50%	20%	10%	50%
Medium outer	5	50%	15%	7.5%	50%
Medium inner	6	50%	10%	5%	50%
Medium ball	7	50%	7.5%	2.5%	50%
Medium combo	8	50%	5%	1%	50%

The analysis of the results indicates that the WKAConvLSTM consistently reaches the highest diagnostic accuracy across all experiments, with an average of 91.74%, and observed extremes between 90.35% and 92.76%. This narrow range demonstrates strong robustness and stability under variable operating conditions. In contrast, the conventional approaches (e.g., CNN and xLSTM) achieve significantly lower average accuracies of 87.68% and 80.61%, respectively, indicating their limited generalization capability in nonstationary environments. Although KANConv reaches a comparable average accuracy to the proposed method, its stability is slightly worse.

To provide an intuitive view, t-SNE is employed on the post-classification features, projecting them into a lower-dimensional space for visualization. This approach facilitates an intuitive examination of the

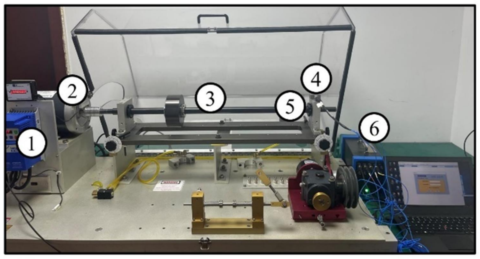
Table 6. Diagnostic results (%) under data imbalance conditions.

Dataset	Models	Min-acc	Max-acc	Avg-acc ± Std
Dataset A	CNN	96.44	99.08	98.13 ± 1.23
	KANConv	98.35	98.70	98.56 ± 0.15
	xLSTM	65.00	71.53	67.50 ± 2.89
	CNN-LSTM	96.44	98.18	97.53 ± 0.69
	ConvFormer	98.18	98.87	98.56 ± 0.29
	WKAConvLSTM	98.70	99.22	**98.94** ± **0.23**
Dataset B	CNN	94.10	94.39	94.28 ± 0.13
	KANConv	91.41	93.84	92.00 ± 1.04
	xLSTM	53.09	59.72	57.45 ± 2.70
	CNN-LSTM	95.75	96.61	96.11 ± 0.32
	ConvFormer	94.18	96.88	95.99 ± 1.10
	WKAConvLSTM	97.92	98.52	**98.16** ± **0.24**
Dataset C	CNN	83.78	84.20	84.04 ± 0.16
	KANConv	78.73	80.38	79.93 ± 0.68
	xLSTM	47.83	48.87	48.33 ± 0.40
	CNN-LSTM	89.41	92.62	91.75 ± 1.34
	ConvFormer	90.19	91.93	91.04 ± 0.66
	WKAConvLSTM	96.09	87.66	**95.43** ± **0.64**

spatial distribution and grouping of various categories, highlighting the model's ability to distinguish between different classes. It can be found form Fig. 7 that WKAConvLSTM can effectively separate the most fault categories compared with other baseline models. However, the dynamically changing load causes the signal to present highly complex nonlinear characteristics, which increases the fuzziness of feature boundaries and the overlap between categories, thereby inevitably leading to classification errors on minority samples for the model.

4.3. Fault Diagnosis under Data Imbalance Conditions

In practical industrial applications, fault samples often exhibit significant imbalanced distributions, which mainly attributed to the complexity of operating conditions and differences in fault occurrence probabilities. To evaluate the robustness and diagnostic performance of WKAConvLSTM under such imbalanced conditions, the time-varying data provided by HUSTBearing dataset is leveraged for experimental validation [37]. This dataset simulates inner-race defects, outer-race defects, rolling element defects, and their combined defects of rolling bearing based on the Spectra-Quest test rig. Each fault type includes two severity levels: medium and severe, as depicted in Fig.8 and Fig. 9.



1: Speed control, 2:Motor, 3: Shaft, 4: Acceleration sensor, 5: Bearing, 6: Data acquisition board

Fig. 8. Test rig of HUSTBearing dataset.

For this study, two datasets with different imbalance degrees are designed based on HUSTbearing dataset. The details are shown in Table 5. It can be observed that the Dataset 1 is balanced and designed to serve as a baseline for the experiments, with an equal number of samples in both the training and testing datasets. Datasets 2 and 3 maintain the same number of testing datasets configuration as Dataset 1. However, their training datasets are constructed by randomly sampling each fault category according to a specified proportion to introduce data class imbalance.

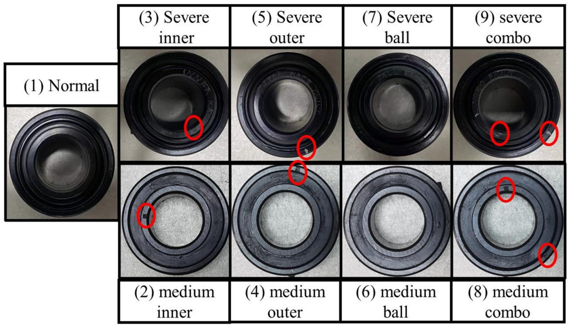


Fig. 9. The fault types of rolling bearing.

Furthermore, the baseline model, hyperparameters, and evaluation metrics are maintained as in the previous experiments. The outcomes of these experiments are summarized in Table 6 and Fig. 10. For a clearer illustration of the WKAConvLSTM's performance, the confusion matrices and t-SNE visualizations of the learned feature representations for all three datasets are illustrated in Fig. 11 and Fig. 12.

The analysis of the results indicates that Analysis of the results indicates that the WKAConvLSTM attains superior average accuracy compared with all benchmark models across the three datasets. Notably, as the degree of class imbalance increases, the accuracy of all model decreased to varying extents, among which WKAConvLSTM shows the smallest performance degradation, with only a 3.51% drop even under highly imbalanced conditions, fully demonstrating its robustness in feature extraction. However, as shown in Fig. 11 and 12, the model still

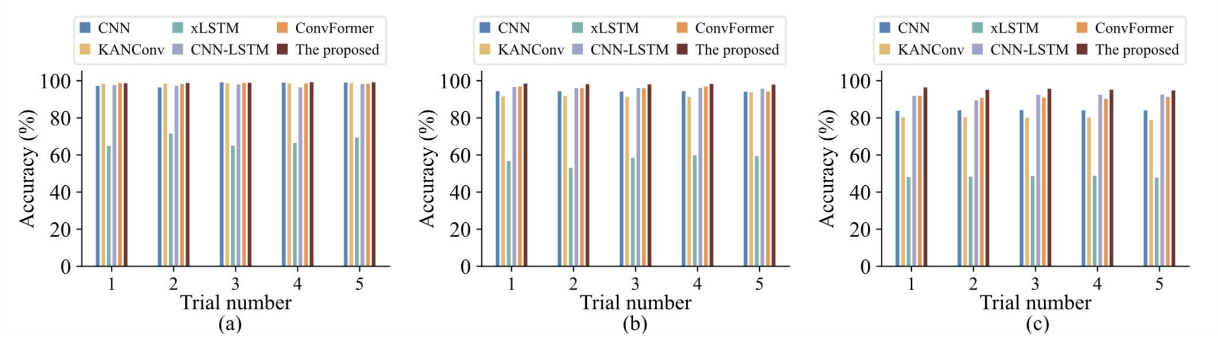


Fig. 10. Diagnostic accuracy under data imbalance conditions. (a) Dataset 1. (b) Dataset 2. (C) Dataset 3.

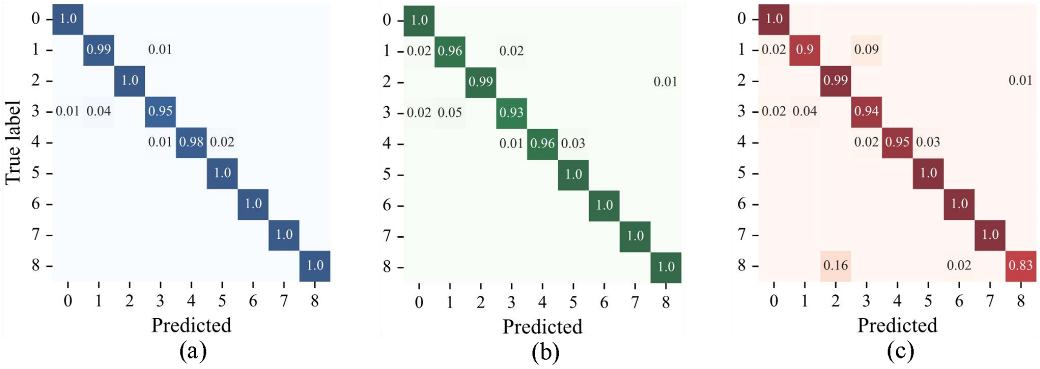


Fig. 11. Confusion matrix of WKAConvLSTM. (a) Dataset 1. (b) Dataset 2. (C) Dataset 3.

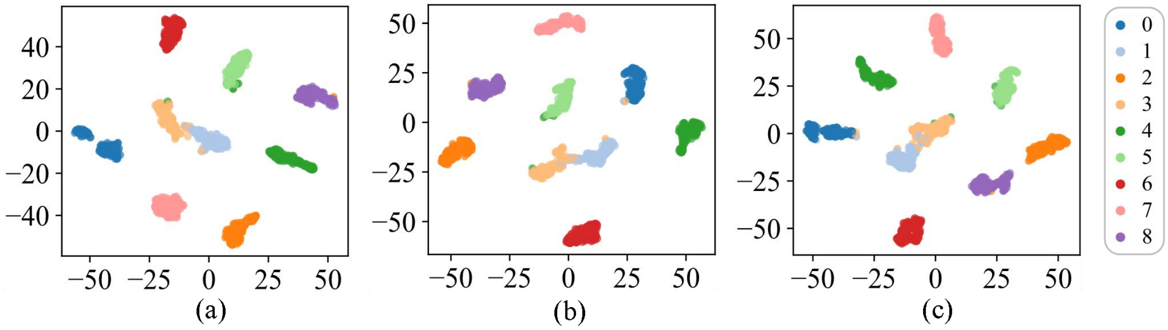


Fig. 12. T-SNE visualization of feature extracted by WKAConvLSTM under the data imbalance conditions. (a) Dataset 1. (b) Dataset 2. (C) Dataset 3.

Table 7. Diagnostic accuracy (%) of ablation experiments.

Model	WKAConv	MA	MCC5-THU	HUSTBearing
Model 1	√	×	90.20 ± 1.28	97.66 ± 0.22
Model 2	×	√	88.38 ± 2.76	98.40 ± 1.40
Model 3	×	×	87.33 ± 1.75	95.52 ± 2.04
WKAConvLSTM	√	√	91.74 ± 1.11	98.94 ± 0.23

exhibits a certain degree of misclassification for several fault categories (e.g., #1 vs. #3, #2 vs. #8), which can be mainly attributed to the extreme data imbalance resulting in a severe shortage of training samples for some classes, thereby limiting the model's discriminative capability for minority categories.

5. Further discussion

5.1. Ablation Study

To assess the impact of individual components within WKAConvLSTM on overall performance, an ablation study is performed in this part. Based on the WKAConvLSTM, three comparative models are additionally constructed for verification: 1) Model 1: remove the multi-head attention mechanism; 2) Model 2: replace the WKAConv layer with a conventional convolutional layer; 3) Model 3: remove both the multi-head attention mechanism and the WKAConv layer, making the model degenerate into CNN-LSTM. The outcomes of the experiments are presented in Table 7.

It can be found from these experimental results that removing the multi-head attention mechanism or replacing the WKAConv layers with traditional convolutional layers results in lower accuracy on both datasets than the complete WKAConvLSTM, but still significantly outperforms Model 3. This indicates that both the multi-head attention mechanism and the WKAConv layer can effectively enhance the model's capabilities of feature extraction and classification, while their synergy is crucial and irreplaceable to the overall performance of the WKAConvLSTM.

5.2. The Influence of Mother Wavelet

To investigate how the selection of different mother wavelets (MWs) impacts the effectiveness of WKAConvLSTM, three variants are constructed using the Mexican

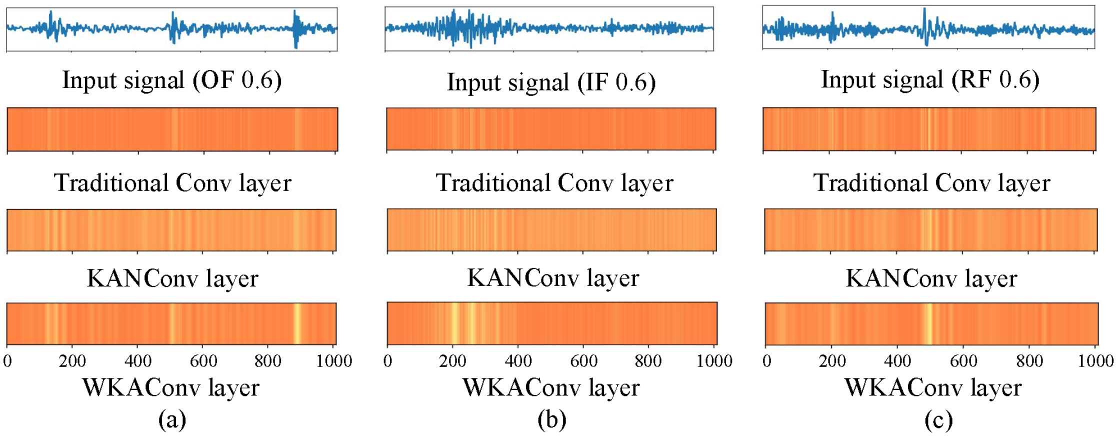


Fig. 13. Feature map of traditional convolutional layer, KANConv layer and WKAConv layer.

Table 8. The formula of four mother wavelet.

MW	Formula
Morlet	$\psi\left(t\right)=Ce^{-\frac{t^2}{2}}\cos\left(5t\right)$
Mexhat	$\psi\left(t\right)=\frac{2}{\sqrt{3\sigma}\pi^{1/4}}\left(1-\frac{t^2}{\sigma^2}\right)e^{-\frac{t^2}{2\sigma^2}}$
Laplace	$\psi(t)=Ce^{\frac{-\xi}{\sqrt{1-\xi^2}}\omega(t-u)}\sin(2\pi f(t-u))$
	d t^2

Hat (Mexhat), Laplace, and Derivative of Gaussian (DoG) wavelets shown in Table 8, denoted as Model_M, Model_L, and Model_D, respectively. Their performance is compared with the standard WKAConvLSTM based on the Morlet wavelet, and the outcomes of the experiments are presented in Table 9.

The examination of the outcomes reveals that the choice of mother wavelet has a significant impact on the performance of the WKAConvLSTM, yet some wavelet functions lack universal applicability. For instance, the Laplace wavelet performs well on the SDUSTBearing and HUSTBearing datasets, but its performance deteriorates significantly under nonstationary conditions (MCC5-THU gearbox dataset), reaching only 39.65%. Moreover, the Mexhat and DoG wavelets consistently underperform compared to the Morlet wavelet across all three datasets. In summary, the WKAConvLSTM based on the Morlet wavelet can achieves the best overall performance.

6. The Interpretability of the Proposed Method

To further discuss the interpretability of the WKAConvLSTM, three vibration signals with distinct impact components corresponding to outer-race defects, inner-race defects, and rolling element defects are

Table 9. Diagnostic accuracy (%) of ablation experiments.

Model	MW	SDUSTBearing	MCC5-THU	HUSTBearing
Model_M	Mexhat	99.61 ± 0.52	89.63 ± 1.39	98.14 ± 2.03
Model_L	Laplace	98.41 ± 0.43	39.65 ± 2.85	93.98 ± 1.73
Model_D	DoG	99.81 ± 0.11	88.42 ± 3.91	98.86 ± 0.98
WKAConvLSTM	Morlet	99.95 ± 0.06	91.74 ± 1.11	98.94 ± 0.23

selected from bearing dataset, and used as inputs to the model. For the same input, feature maps are extracted from the WKAConv layer, the conventional convolutional layer, and the KANConv layer, respectively. The differences in their feature extraction capabilities are then comparatively analyzed, as illustrated in Fig. 13.

The analysis of these results indicates that the WKAConv layer demonstrates significant advantages in extracting fault-related features. Specifically, its activation responses are highly aligned with the impact components in the original vibration signals, and the resulting feature maps display sharp boundaries and strong contrast in high-response regions. In comparison, traditional CNN and KANConv layers tend to produce numerous irrelevant or spurious activations when processing time-series data, with the KANConv layer being particularly. Such redundant activations blur the distinction between fault features and background information, which to a certain extent interferes with the accuracy of subsequent fault pattern recognition and classification results.

7. Conclusions

In this article, an interpretable wavelet Kolmogorov–Arnold convolutional LSTM (WKAConvLSTM) is proposed for spatial-temporal feature extraction, where the WKAK with learnable scale factor and translation factor is embedded into traditional convolutional layer to extract interpretable spatial features, and the multi-head attention enhanced LSTM is designed for capturing crucial temporal features of vibration signal. Experimental findings indicate that the WKAConvLSTM outperforms the compared baseline models under both noisy, nonstationary and data imbalance conditions. Moreover, its effectiveness is further verified through interpretability analysis.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 52505107), and Yunnan Fundamental Research Projects (NO. 202501CF070179).

Declare of Interests

The authors declare no conflicts of interest .

References

[1] C. Ding, W. Huang, C. Shen, X. Jiang, J. Wang, and Z. Zhu, "Synchroextracting frequency synchronous chirplet transform for fault diagnosis of rotating machinery under varying speed conditions," *Structural Health Monitoring,* vol. 23, no. 3, pp. 1403-1422, 2024.

[2] C. Guo, Z. Zhao, J. Ren, S. Wang, Y. Liu, and X. Chen, "Causal explaining guided domain generalization for rotating machinery intelligent fault diagnosis," *Expert Systems with Applications,* vol. 243, p. 122806, 2024.

[3] T. Li, C. Sun, S. Li, Z. Wang, X. Chen, and R. Yan, "Explainable graph wavelet denoising network for intelligent fault diagnosis," *IEEE Transactions on Neural Networks and Learning Systems,* 2022.

[4] T. Liu, S. Wang, X. Li, Y. Li, and K. Noman, "Time-frequency mode adaptive decomposition based on the maximum kurtosis for extracting fault component of bearings," *IEEE Transactions on Instrumentation and Measurement,* 2025.

[5] S. Wei, B. Hou, D. Wang, S. Liu, and Z. Peng, "Generalized difference mode decomposition for adaptively

extracting fault components of rotating machinery under non-stationary conditions," *Journal of Sound and Vibration,* vol. 609, p. 119089, 2025.

[6] D. Peng, M. Yazdanianasr, A. Mauricio, T. Verwimp, W. Desmet, and K. Gryllias, "Physics-driven cross domain digital twin framework for bearing fault diagnosis in non-stationary conditions," *Mechanical Systems and Signal Processing,* vol. 228, p. 112266, 2025.

[7] H. Wang, X. Wang, Y. Yang, K. Gryllias, and Z. Liu, "A few-shot machinery fault diagnosis framework based on self-supervised signal representation learning," *IEEE Transactions on Instrumentation and Measurement,* vol. 73, pp. 1-14, 2024.

[8] R. Kumar and R. Anand, "Bearing fault diagnosis using multiple feature selection algorithms with SVM," *Progress in Artificial Intelligence,* vol. 13, no. 2, pp. 119-133, 2024.

[9] B. Zhao, Q. Wu, K. Zhao, Z. Mo, Z. Zhang, and X. Zhang, "MNHP-GAE: A novel manipulator intelligent health state diagnosis method in highly imbalanced scenarios," *IEEE Internet of Things Journal,* 2024.

[10] X. Zhao *et al.*, "Model-assisted multi-source fusion hypergraph convolutional neural networks for intelligent few-shot fault diagnosis to electro-hydrostatic actuator," *Information Fusion,* vol. 104, p. 102186, 2024.

[11] B. Wang, W. Qiu, X. Hu, and W. Wang, "A rolling bearing fault diagnosis technique based on recurrence quantification analysis and Bayesian optimization SVM," *Applied Soft Computing,* vol. 156, p. 111506, 2024.

[12] Y. Ming, H. Shao, B. Cai, and B. Liu, "rgfc-Forest: An enhanced deep forest method towards small-sample fault diagnosis of electromechanical system," *Expert Systems with Applications,* vol. 238, p. 122178, 2024.

[13] T. Li, C. Sun, R. Yan, and X. Chen, "A novel unsupervised graph wavelet autoencoder for mechanical system fault detection," *Journal of Intelligent Manufacturing,* pp. 1-18, 2024.

[14] Z. Lei *et al.*, "Unsupervised graph transfer network with hybrid attention mechanism for fault diagnosis under variable operating conditions," *Reliability Engineering & System Safety,* vol. 255, p. 110684, 2025.

[15] Y. Li, X. Wang, Y. He, Y. Wang, Y. Wang, and S. Wang, "Deep spatial-temporal feature extraction and lightweight feature fusion for tool condition monitoring," *IEEE Transactions on Industrial Electronics,* vol. 69, no. 7, pp. 7349-7359, 2021.

[16] Y. Wang, D. Bao, and S. Li, "Dynamic graph embedding PCA to extract spatio–temporal information for fault detection," *IEEE Transactions on Industrial Informatics,* 2024.

[17] J. Li, D. Zhao, L. Xie, Z. Zhou, L. Zhang, and Q. Chen, "Spatial–temporal synchronous fault feature extraction and diagnosis for proton exchange membrane fuel cell systems," *Energy Conversion and Management,* vol. 315, p. 118771, 2024.

[18] R. Yan *et al.*, "Knowledge Driven Machine Learning Towards Interpretable Intelligent Prognostics and Health Management: Review and Case Study," *Chinese Journal of Mechanical Engineering,* vol. 38, no. 1, p. 5, 2025.

[19] T. Li *et al.*, "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE*

Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 4, pp. 2302-2312, 2021.

[20] H. Wang, Y.-F. Li, T. Men, and L. Li, "Physically interpretable wavelet-guided networks with dynamic frequency decomposition for machine intelligence fault prediction," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 2024.

[21] T. Huang, Q. Zhang, X. Tang, S. Zhao, and X. Lu, "A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems," *Artificial Intelligence Review,* vol. 55, no. 2, pp. 1289-1315, 2022.

[22] S. Zhao, Y. Duan, N. Roy, and B. Zhang, "A deep learning methodology based on adaptive multiscale CNN and enhanced highway LSTM for industrial process fault diagnosis," *Reliability engineering & system safety,* vol. 249, p. 110208, 2024.

[23] M. T. Singh, R. K. Prasad, G. R. Michael, N. H. Singh, and N. Kaphungkui, "Spatial-Temporal Bearing Fault Detection Using Graph Attention Networks and LSTM," *arXiv preprint arXiv:2410.11923,* 2024.

[24] S. Li, T. Li, C. Sun, X. Chen, and R. Yan, "WPConvNet: An interpretable wavelet packet kernel-constrained convolutional network for noise-robust fault diagnosis," *IEEE Transactions on Neural Networks and Learning Systems,* 2023.

[25] Q. Li, H. Li, W. Hu, S. Sun, Z. Qin, and F. Chu, "Transparent operator network: A fully interpretable network incorporating learnable wavelet operator for intelligent fault diagnosis," *IEEE Transactions on Industrial Informatics,* 2024.

[26] Z. Zhao *et al.*, "Model-driven deep unrolling: Towards interpretable deep learning against noise attacks for intelligent fault diagnosis," *ISA transactions,* vol. 129, pp. 644-662, 2022.

[27] S. Li, T. Li, C. Sun, R. Yan, and X. Chen, "Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis," *Journal of manufacturing systems,* vol. 69, pp. 20-30, 2023.

[28] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate Shapley value feature attributions," *Nature Machine Intelligence,* vol. 5, no. 6, pp. 590-601, 2023.

[29] Z. Liu *et al.*, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756,* 2024.

[30] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation,* vol. 31, no. 7, pp. 1235-1270, 2019.

[31] J. Wang *et al.*, "Attention guided multi-wavelet adversarial network for cross domain fault diagnosis," *Knowledge-based systems,* vol. 284, p. 111285, 2024.

[32] D. Ruan, J. Wang, J. Yan, and C. Gühmann, "CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis," *Advanced Engineering Informatics,* vol. 55, p. 101877, 2023.

[33] A. D. Bodner, A. S. Tepsich, J. N. Spolski, and S. Pourteau, "Convolutional kolmogorov-arnold networks," *arXiv preprint arXiv:2406.13155,* 2024.

[34] M. Beck *et al.*, "xlstm: Extended long short-term memory," *arXiv preprint arXiv:2405.04517,* 2024.

[35] H. Wang *et al.*, "Convformer: Revisiting transformer for sequential

user modeling," *arXiv preprint arXiv:2308.02925,* 2023.

[36] S. Chen, Z. Liu, X. He, D. Zou, and D. Zhou, "Multi-mode fault diagnosis datasets of gearbox under variable working conditions," *Data in brief,* vol. 54, p. 110453, 2024.

[37] C. Zhao, E. Zio, and W. Shen, "Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study," *Reliability Engineering & System Safety,* vol. 245, p. 109964, 2024.