Journal of Electronics and Information Science (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2025.100210
ISSN 2371-9524 Vol. 10 Num. 2

A rolling bearing fault diagnosis method based on the improved sparrow search algorithm optimized VMD and multi-scale convolutional neural networks

Gaolei Mao[1,a,*], Yali Sun[1,b]

[1]*School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou, China*
[a]*870768944@qq.com*, [b]*2636491670@qq.com*
[*]*Corresponding author*

Keywords: Fault diagnosis, Rolling bearings, Variational modal decomposition algorithm, Improved sparrow search algorithm, Multi-scale convolutional neural networks

Abstract: To address the issues of low diagnostic accuracy in traditional rolling bearing fault diagnosis models and the ineffective extraction of spatial and temporal features from vibration signals, this paper proposes a rolling bearing fault diagnosis method based on the improved sparrow search algorithm optimized VMD and multi-scale convolutional neural networks. First, the improved sparrow search algorithm is employed to adaptively optimize the penalty parameter and mode count in variational modal decomposition (VMD). This achieves finer frequency band segmentation and effectively suppresses energy leakage, thereby yielding high quality frequency domain representations. Second, a multi-scale convolutional neural networks (MSCNN) is constructed, with feature level fusion implemented. Subsequently, a bidirectional long short-term memory networks (BiLSTM) is introduced to model the temporal dependencies of the fused features, enabling fault mode learning. A softmax layer is employed to achieve multi-class classification. Finally experimental results and comparisons based on the CWRU bearing dataset demonstrate the effectiveness of the proposed method in the rolling bearing fault classification task, providing significant application value for achieving efficient and reliable bearing fault detection.

1. Introduction

Currently, rotating machinery has been widely adopted across multiple industries including energy, transportation, and aerospace [1]. As critical components in rotating machinery, rolling bearings primarily support rotating parts and transmit power. Their operational health directly determines whether equipment can run stably [2-4]. When subjected to prolonged operation under complex conditions, factors such as high loads, severe impacts, and friction wear can cause bearings to experience premature degradation and failure, potentially leading to shutdowns and economic losses [5]. Therefore, conducting fault diagnosis for rolling bearings is of great significance.

The health monitoring and fault diagnosis of rolling bearings have become increasingly complex.

Vibration signals are prone to noise interference during the feature extraction process, which has prompted researchers to propose various signal decomposition methods. Huang et al [6] proposed empirical mode decomposition (EMD), which can adaptively decompose signals under noisy conditions, but suffers from mode aliasing and endpoint effects. Dragomiretskiy et al [7] proposed the VMD effectively mitigates end-point effects and suppresses modal aliasing. Chen et al [8] applied VMD to bearing signal decomposition, but its penalty factor and mode number are difficult to determine. Improper parameter settings can easily generate spurious components, thereby affecting subsequent classification accuracy. In their related research, Wang et al [9] employed a combined approach of minimum redundancy and maximum correlation to determine the number of modal components. Liu et al [10] selected K based on the center frequencies of each component derived from VMD and energy spectrum analysis, but neither study simultaneously considered the coupling effects between α and K. However, in the aforementioned studies, only one of the two parameters was optimized, without considering their mutual influence.

With the advancement of machine learning, several intelligent algorithms have been applied to bearing fault diagnosis. Commonly used algorithms include convolutional neural networks (CNN) [11], long short-term memory networks (LSTM) [12], and gated recurrent unit networks (GRU) [13]. Xu et al [14] applied LSTM to fault diagnosis in ship power plants and AC motor systems, achieving satisfactory diagnostic results. Dong et al [15] proposed a bearing fault diagnosis model combining CNN and BiLSTM. Although it considers the model's feature representation capabilities in both spatial and temporal domains, it exhibits certain limitations due to its single-channel architecture and the potential degradation of temporal information following convolution.

To address the aforementioned issues, a rolling bearing fault diagnosis method based on the improved sparrow search algorithm optimized VMD and multi-scale convolutional neural networks is proposed. First, the bearing vibration signals are preprocessed, and features are subsequently extracted using VMD. To enhance decomposition quality, an improved sparrow search algorithm is designed to adaptively optimize the penalty factor and number of modes in VMD, thereby obtaining optimal modal components. Second, we construct a multi-scale convolutional neural networks (MSCNN) to enhance the network's perceptual capabilities regarding input data. Finally, a BiLSTM network is introduced to model the temporal dependencies of the fused features for fault diagnosis, thereby enhancing the accuracy of bearing fault detection.

2. Proposed method

2.1 Improved sparrow search algorithm

The sparrow search algorithm (SSA) [16] is a swarm intelligence optimization method inspired by the foraging behavior of sparrows. It divides individuals into discoverers and followers to collaboratively perform global and local searches, thereby solving combinatorial optimization problems. Its advantages lie in its simplicity of implementation, minimal parameter requirements, and inherent global optimization capabilities. However, it is prone to local optima, exhibits slow convergence, and random initialization may result in uneven population distribution, thereby reducing subsequent search efficiency.

Before implementing VMD decomposition, the modal number k and penalty parameter α must be preset. Parameter values that are too large or too small may lead to overfitting and information redundancy. Therefore, it is necessary to jointly determine the optimal parameter pair (k,α). To this end, intelligent optimization strategies are commonly employed to search for VMD parameters.

To overcome the aforementioned shortcomings, this paper proposes an improved sparrow search algorithm, which integrates the osprey algorithm, Cauchy mutation strategy, and sparrow search

algorithm. This method enable joint global optimization of (k,α) to obtain the optimal combination, thereby enhancing decomposition quality and subsequent fault diagnosis performance. The specific improvements are as follows.

(1) Employing the Logistic chaotic map to initialize the diversity of the population.

(2) Based on Tizhoosh's inverse learning, the refraction principle is introduced to generate "refraction inverse solutions" , which both expand the search scope and mitigate premature convergence issues in later stages.

In the global exploration phase of stage 1, the osprey optimization algorithm replaces the locator position update in SSA: it randomly selects prey points based on the detect-dive mechanism and executes attack updates, reducing dependence on the previous generation's positions. Simultaneously, the explorer displacement rules are rewritten using the osprey's fish-chasing movement model, enhancing global search capabilities while mitigating premature convergence risks. The formula for the global exploration strategy of the osprey optimization algorithm in the first phase is shown in formula (1):

$$x_{i,j}^{P1} = x_{i,j} + r_{i,j} \cdot \left(SF_{i,j} - I_{i,j} \cdot x_{i,j} \right) \tag{1}$$

In the formula: $x_{i,j}^{P1}$ denotes the new position in the j-th dimension of the i-th fish eagle during the first phase; $x_{i,j}$ represents the individual; $SF_{i,j}$ is a random number between [0,1]; $I_{i,j}$ is a random number from the set {1,2}.

During the local exploitation phase, replace the displacement update of SSA followers with Cauchy variation. Since the Cauchy distribution exhibits heavier tails and slower decay compared to the normal distribution, it generates more substantial random perturbations. Applying these perturbations to individual positions expands the search radius and enhances population diversity, thereby improving the ability to escape local optima and achieve global optimization. During the foraging process, employing the Cauchy mutation strategy enhances the algorithm's capabilities. The updated position of the new follower is as follows:

$$x_{i,j}^{i+1} = x_{best}(t) + cauchy(0,1) \oplus x_{best}(t) \tag{2}$$

In formula (2), $cauchy(0,1)$ denotes the Cauchy distribution function; $\oplus$ denotes the multiplication operation between the two.

To ensure more precise position updates, this paper constructs a greedy principle defined by comparing the fitness values of two new positions. The greedy rule is expressed as formula (3), where $f(x)$ represents the fitness value of position x.

$$\begin{cases} x_{best} = x_{i,j}^{t+1}, f\left(x_{i,j}^{t+1}\right) < f\left(x_{best}\right) \\ x_{best} = x_{best}, f\left(x_{i,j}^{t+1}\right) \geq f\left(x_{best}\right) \end{cases} \tag{3}$$

2.2 Multi-scale feature extraction module

To maximize the extraction of feature information from the input signal, this paper designs a multi-scale feature extraction module, as shown in Figure 1. Compared to traditional CNN with single-path processing and early pooling, this module employs parallel convolutions to obtain multi-scale representations and lightweight concatenation to enhance nonlinear expressions, as detailed below. The first layer configures three parallel convolutions: 1×1, 1×3 and 1×5, with channel counts of 16, 8 and 8 respectively. The second layer further stacks 1×5 and 1×3

convolutions on corresponding branches, each with 16 channels. The third layer employs two 1×1 convolutions for channel compression and fusion, with channel counts of 32 and 16 respectively. All convolutions are followed by batch normalization and activation functions, with outputs from each branch concatenated via a concat layer.

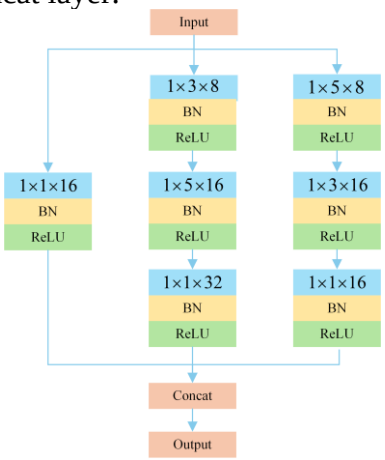


Figure 1: Multi-scale feature extraction module structure.

2.3 Bidirectional long short-term memory network

BiLSTM [17] introduces two independent branches-forward and backward-based on LSTM, as shown in Figure 2. The same sequence is input into two LSTM layers in both forward and reverse order to extract contextual features, and their output vectors are concatenated and fused to form a bidirectional temporal representation. This architecture simultaneously captures past and future dependencies, making it suitable for variable length sequences. Compared to unidirectional LSTM, it demonstrates superior feature representation capabilities in temporal modeling and recognition tasks.

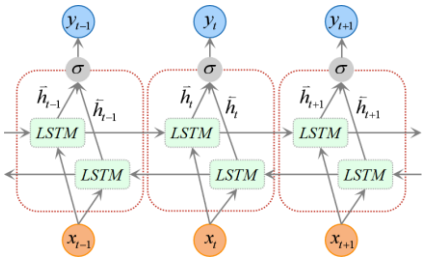
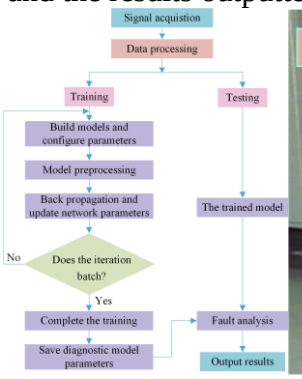


Figure 2: BiLSTM network archiecture diagram.

2.4 Diagnostic Process for the proposed fault diagnosis method

The diagnostic flow chart for the rolling bearing fault diagnosis method based on the improved sparrow search algorithm optimized VMD and multi-scale convolutional neural networks is shown in Figure 3. As shown in Figure 3, the process is primarily divided into three parts: data preprocessing, model training, and fault diagnosis. First, sensor data is collected to capture vibration signals from various types of bearing faults, and the bearing vibration signals undergo preprocessing. The data is divided into a training set and a test set, with the former used for model training and the latter for independent evaluation. Subsequently, a fault diagnosis model is constructed and its parameters are configured. The model parameters are initialized and iteratively updated via backpropagation. The model parameters are initialized and iteratively updated through backpropagation. Training continues until the convergence criterion is met, after which the model is

saved. Finally, the trained model is evaluated on the test set to perform feature extraction and fault identification, and the results outputted.



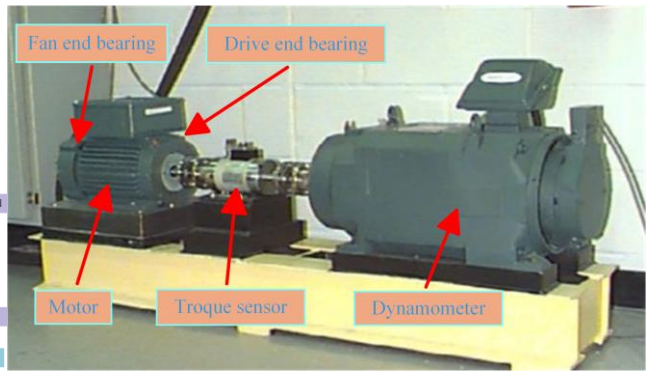


Figure 3: Diagnostic flow chart Fig 4: CWRU test bench

3. Experimental analysis

To validate the effectiveness and generalization capability of the proposed method, this paper employs the bearing dataset from Case Western Reserve University (CWRU) for verification. The dataset is divided into training and testing sets. The experimental setup utilizes an Intel® Core™-i7-12900H@2.50GHz.

3.1 Dataset description

To evaluate the diagnostic performance of the proposed model, the CWRU[18] bearing fault dataset is employed for validation. The test bench, as shown in Figure 4, consists of a 1.5 kw motor, a torque sensor, a dynamometer, and an electronic controller. The bearing under test is an SKF 6205, with a sampling frequency of 12 kHz. Four load conditions are configured: 0, 1, 2, and 3 hp, with corresponding rotational speeds of 1797, 1772, 1750, and 1730 r/min, respectively.Vibration signals are captured by acceleration sensors and categorized into datasets A, B, C, and D based on operating conditions. Specific details are provided in Table 1. To better validate the diagnostic performance of the proposed method, experiments are conducted under varying operating conditions, noise levels, and loads. The proposed method is compared with five other methods. The comparison methods include CNN-LSTM, CNN-BiLSTM, VMD-CNN-BiLSTM (VCBiLSTM), and GWO-VMD-CNN-BiLSTM (GVCBiLSTM).

Table 1: Description of the dataset

Dataset	Rotating speed (r/min)	Load/ (hp)
A	1797	0
B	1772	1
C	1750	2
D	1730	3

3.2 Variable noise fault diagnosis experimental results and analysis

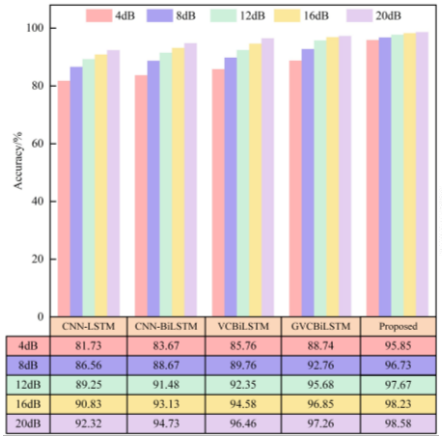
Bearing operation is often accompanied by significant noise interference. Diagnostic models must possess robust noise resistance, only models with superior anti-interference capabilities can achieve precise fault identification. To validate the noise immunity advantages of the proposed method, a noise immunity performance test is designed. Gaussian white noise with varying signal to noise ratios is added to the original vibration signal to simulate real-world noise environments.

Experiments are conducted using data from a 2hp load, with Gaussian white noise added at SNR of 4dB, 8dB, 12dB, 16dB, and 20dB. Figure 5 shows the recognition accuracy rates for various methods.

As shown in Figure 5, the proposed method achieves higher accuracy than the other four methods. The average accuracy of the CNN-LSTM method is lower than that of the other methods. At an SNR of 12 dB, the accuracy of the proposed method reached 97.67%, outperforming the CNN-LSTM, CNN-BiLSTM, VCBiLSTM, and GVCBiLSTM methods by 1.99 percentage points, 5.32 percentage points, 6.19 percentage points, and 8.42 percentage points, respectively. When the SNR is 20 dB, the diagnostic accuracy of the method proposed in this paper reaches 98.58%, demonstrating high accuracy. This demonstrates that the designed method exhibits superior noise immunity and stability compared to the reference method.

3.3 Variable load fault diagnosis results and analysis

In actual industrial settings, bearing loads fluctuate significantly with operating conditions. Fault diagnosis models must possess strong cross-load generalization capabilities to maintain efficient and accurate identification performance across multiple operating conditions. To evaluate the cross-load generalization capability of the proposed method, experiments are conducted using load data from four load conditions: 0, 1, 2, and 3 hp. Each experiment uses one load dataset as the training set, with the remaining three load datasets serving as independent test sets. For simplicity, denote 0-1, 0-2, and 0-3 as training under 0 hp conditions and testing under 1, 2, and 3 hp conditions, respectively. Other combinations follow this pattern. The proposed method is compared with other approaches, with the results shown in Figure 6.



	CNN-LSTM	CNN-BiLSTM	VCBiLSTM	GVCBiLSTM	Proposed
4dB	81.73	83.67	85.76	88.74	95.85
8dB	86.56	88.67	89.76	92.76	96.73
12dB	89.25	91.48	92.35	95.68	97.67
16dB	90.83	93.13	94.58	96.85	98.23
20dB	92.32	94.73	96.46	97.26	98.58

Figure 5: Results for dfferent SNR

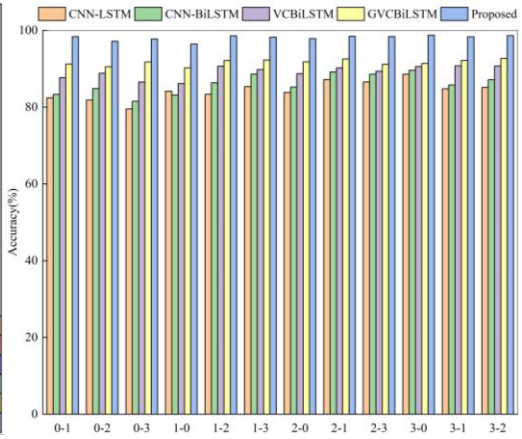
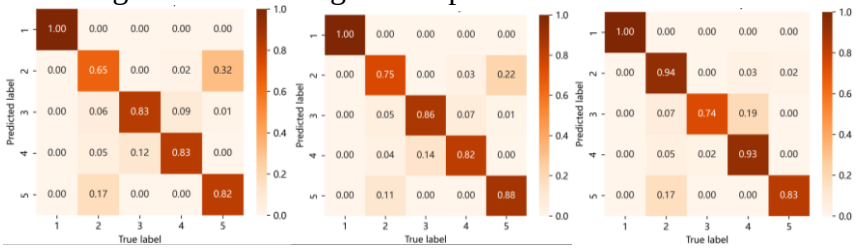


Fig 6: Variable load results

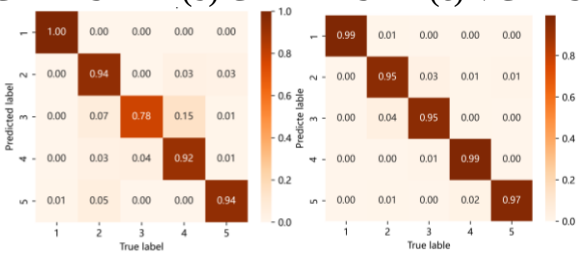
As shown in Figure 6, the results obtained using the proposed method in this paper consistently outperform those of the reference method in variable load experiments. The average accuracy rates of CNN-LSTM, CNN-BiLSTM, VCBiLSTM, and GVCBiLSTM under variable operating conditions are 84.39%, 86.10%, 89.18%, and 91.66%, respectively. The average accuracy rate of the proposed method reaches 98.08%. Compared with the aforementioned four methods, the proposed method achieves improvements of 13.69, 11.98, 8.90, and 6.42 percentage points, respectively. The results demonstrate that the proposed method exhibits significant advantages and excellent generalization capabilities under variable load conditions. The reason lies in the fact that different loads cause changes in the frequency and amplitude distribution of bearing vibration signals. The feature extraction module of this method can adaptively handle these variations, enabling stable extraction of discriminative features and thereby enhancing the model's generalization performance.

3.4 Confusion matrix results and analysis

To further validate the fault classification performance of the proposed method, we conduct a confusion matrix analysis on the test results. Under operating conditions of a 2 hp load and an SNR of 8 dB, the results of the confusion matrix are shown in Figure 7. The value along the main diagonal represents the proportion of correct classifications within that category, while values in other positions indicate the proportion misclassified into the corresponding category. As shown in Figure 7, the other four comparison methods exhibit significant misclassification and lower diagnostic accuracy. The proposed method achieves accuracy rates exceeding 95% across all five categories, with a significant reduction in off-diagonal elements. These results demonstrate that the proposed method maintains high classification accuracy and stability under low SNR and different load conditions, exhibiting robust fault diagnosis capabilities.



(a) CNN-LSTM (b) CNN-BiLSTM (c) VCBiLSTM



(d) GVCBiLSTM (e) Proposed

Fig 7: Confusion matrix for different methods

4. Conclusion

To ensure the safe operation of rolling bearings in mechanical equipment and enhance the accuracy and generalization capability of fault diagnosis, this paper proposes a rolling bearing fault diagnosis method based on the improved sparrow search algorithm optimized VMD and multi-scale convolutional neural networks. First, an improved sparrow search algorithm is employed to adaptively adjust the penalty factor and number of modes in VMD, enabling more precise frequency band segmentation and effectively suppressing energy leakage to yield high quality frequency domain features. Second, a multi-scale convolutional neural networks is constructed and feature fusion is performed at the feature level to further enhance feature expression capabilities. Then, a bidirectional long short-term memory networks is introduced to model the temporal dependencies of the fused features, and a softmax layer is employed to perform multi-class fault classification. Finally, experiments and comparative analyses based on the CWRU bearing dataset demonstrate that the proposed method exhibits significant advantages in rolling bearing fault classification tasks, providing substantial application value for achieving efficient and reliable bearing fault detection.

References

[1] Lu K., Jin Y.L., Chen Y.S., et al. (2019) Review for Order Reduction Based on Proper Orthogonal Decomposition and Outlooks of Applications in Mechanical Systems [J]. Mechanical Systems and Signal Processing, 123(3): 264-297.

[2] Yan G., Chen J., Bai Y., et al. (2022) A Survey on Fault Diagnosis Approaches for Rolling Bearings of Railway Vehicles[J]. Processes, 10(4):724.

[3] Liu Z.P., Zhang L. (2020) A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings[J]. Measurement, 149: 107002.

[4] Yan R., Shang Z., Xu H., et al. (2023) Wavelet transform for rotary machine fault diagnosis: 10 years revisited[J]. Mechanical systems and signal processing, 200: 110545.

[5] Liu R., Yang B., Zio E., et al. (2018) Artificial intelligence for fault diagnosis of rotating machinery: A review[J]. Mechanical Systems and Signal Processing, 108: 33-47.

[6] Huang N.E., Shen Z., Long S.R., et al. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings Mathematical Physical & Engineering Sciences, 454: 903-995.

[7] Dragomiretskiy K., Zosso D. (2014) Variational Mode Decomposition[J]. IEEE Transactions on Signal Processing, 62(3): 531-544.

[8] Chen G., Lu X., He L., et al. (2023) Subway train rolling bearing fault diagnosis method based on SSA-VMD. Equipment Manufacturing Technology, (7), 42–46.

[9] Wang Y., Cheng Y. (2021) Vibration signal analysis of cylinder head during engine combustion process[J]. Journal of Vibration and Shock, 40(13): 210-215, 254.

[10] Liu Y.S., Wei Z.G., Shu H.X., et al. (2023) Weak fault feature extraction of rolling bearings based on parameter adaptive VMD and MCKD[J]. Noise and Vibration Control, 43(3): 102-109.

[11] Hu P., Zhao C., Huang J., et al. (2023) Intelligent and small samples gear fault detection based on wavelet analysis and improved CNN[J]. Processes, 11(10):2969.

[12] Malhotra P., Ramakrishnan A., Anand G., et al. (2016) LSTM-based encoder-decoder for multi-sensor anomaly detection[J]. arxiv preprint arxiv:1607.00148.

[13] Chung J., Gulcehre C., Cho K.H., et al. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arxiv preprint arxiv:1412.3555.

[14] Xu Z.H., Pan T.L. (2021) Noise reduction method of fan gearbox vibration signal based on variational mode decomposition[J]. Journal of Mechanical & Electrical Engineering, 38(1): 129-132.

[15] Dong S.J., Li Y., Liang T., et al. (2022) Fault diagnosis method of rolling bearing based on CNN-BiLSTM under variable working conditions[J]. Journal of Vibration, Measurement & Diagnosis, 42(5): 1009-1016, 1040.

[16] Sun X.Y.,Xiang D.,Ding W.,et al. (2023) Research on individual pitch control of wind turbine based on sparrow search algorithm[J]. ACTA ENERGIAE SOLARIS SINICA,44(10):266-274.

[17] Lample G., Ballesteros M., Subramanian S., et al. (2016) Neural architectures for named entity recognition[J]. arxiv preprint arxiv:1603.01360 .

[18] L Smith W.A., Randall R.B. (2015) Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study[J]. Mechanical systems and signal processing, 64: 100-131.