ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

Wind Turbine Acoustic Anomaly Detection based on RepViT-MobileNetV3

Qingzheng Li

School of Information and Control Engineering, Jilin University of Chemical Technology, Jilin 132022, China

Abstract

Abnormal sound detection is a technique used to recognize non-normal sound signals, which is widely used in industrial fields, such as abnormal detection of wind turbines. Currently, many abnormal sound detection techniques are based on deep learning. However, in the complex working environment of WTGs, which is filled with a large amount of noise, abnormal sound detection for WTGs faces problems such as difficulty in sound feature extraction and insufficient sound feature extraction capability of the detection network. Therefore, this paper proposes a deep learning-based method for abnormal sound detection of WTGs, called RS-MobileNet.Specifically, SincNet spectral features and Log-Mel spectral features are extracted from the original sound signals, which are fused to become SL spectrograms as feature input. Then the improved RS-MobileNetV3 network is proposed based on the MobileNetV3 network. This network combines the reparameterized visual transform module and the soft pooling method, which can make the MobileNetV3 network keep lightweight while improving the feature extraction capability. Using the NREL dataset, the AUC is improved by 4.53 percentage points compared to the baseline model MobileNetV3.

Keywords

Gearbox; Anomalous Sound Detection; MobileNetV3.

1. Introduction

Wind energy, as a core pillar of clean energy, has an installed capacity exceeding 1 terawatt in 2023, but unplanned downtime due to wind motor gearbox failures has severely constrained operation and maintenance (O&M) efficiency. According to statistics, gearbox failures account for 23% of total wind motor failures, and the average repair takes up to 120 hours.

Existing wind motor gearbox abnormal sound detection methods can be summarized into three types of technology routes: traditional signal processing, statistical modeling and deep learning, but their application in industrial scenarios still faces significant bottlenecks. Traditional signal processing methods, such as wavelet analysis, STFT, MFCC, extract artificial features through time-frequency transformation, however, in a strong interference environment with wind noise >50 dB, the band energy signal-to-noise ratio often falls below -10 dB, and it is difficult to capture the transient features triggered by gear breakage at the microsecond level [1]. Statistical models such as Isolated Forest and GMM, although mitigating the artificial feature dependence through probability distribution modeling, significantly deteriorate the anomaly score differentiation when dealing with high-dimensional spectra of more than 128 dimensions, and GMM based on the assumption of static noise has a high False Detection Rate (FPR) of 15%-20% under the non-stationary noise induced by the sudden change of wind speed [2]. Deep learning methods such as self-encoder and MobileNet have improved detection sensitivity through end-to-end feature learning, but there is still an inherent contradiction between feature fidelity and computational efficiency: self-encoder has a leakage

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

detection rate of >30% for the overlapping frequency band of 2-5 kHz, while the channel compression strategy of MobileNetV3 results in a 58% attenuation rate of the high-frequency features, resulting in a minor fault The detection rate is less than 65%; moreover, the existing network is not adaptive enough to dynamic noise, with the detection AUC decreasing by 10%-15% in the scenario of sudden change in wind speed, and the performance degradation is more than 8% when migrating across models [3]. Current research is focusing on multimodal signal fusion, dynamic noise adversarial training and lightweight and heavily parameterized network architectures to break through the problems of noise robustness and high-frequency feature retention.

Table 1. Comparison of Method Application Scenarios

Method	Applicable Scenarios
Isolation Forest	Low-dimensional static data
Convolutional Autoencoder	Simple noise environments
MobileNetV3	Moderate-noise industrial scenarios
RS-MobileNetV3	High-noise wind turbine environments

Wind motors are usually installed in remote or high altitude areas with high ambient noise, which requires higher robustness for sound detection. In addition, the sound of wind motor operation may contain low-frequency vibration and high-frequency transient noise, which need to be combined with different spectral features to capture the anomalies.

In the abnormal acoustic detection of wind motor gearboxes, the Log-Mel spectrogram is based on the auditory perception characteristics of the human ear, and the ability to characterize low-frequency harmonic components is strengthened by the non-uniformly distributed filter bank, which can effectively resolve the periodic vibration modes such as gear meshing, but its fixed filter design has the problem of resolution attenuation in the high-frequency band, which makes it difficult to accurately capture the transient impact signals. For this reason, this paper integrates the SincNet parametric spectrum analysis technique to construct a bimodal complementary feature space by using a learnable bandpass filter bank to dynamically optimize the high-frequency feature extraction, and combining the advantages of Log-Mel's low-frequency harmonic characterization. This fusion strategy suppresses noise interference through frequency-domain complementarity and enhances robustness at the same time, which is suitable for strongly noisy wind power generation environments[4].

Aiming at the common problem of high-frequency feature attenuation caused by channel dimensionality reduction operations in lightweight networks, this paper introduces the RepViT module, which dynamically fuses local receptive fields and global context information through a multi-branch structure to avoid the loss of details triggered by dimensional compression of traditional attention mechanisms. Its unique structural reparameterization design significantly improves the robustness of weak anomalous signals in complex noise environments by controlling the computational load within the deployable range of edge devices while retaining the ability to interact with multi-scale features.

In summary, this paper proposes a lightweight detection framework based on MobileNetV3 network, which realizes efficient and accurate gearbox condition monitoring and adapts to strongly noisy wind power scenarios by fusing multi-scale acoustic features with dynamic reparameterized network structure. As shown in Fig. 1, the research covers three phases of feature extraction, network optimization and engineering validation, providing a low-latency and highly robust solution for wind power operation and maintenance.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

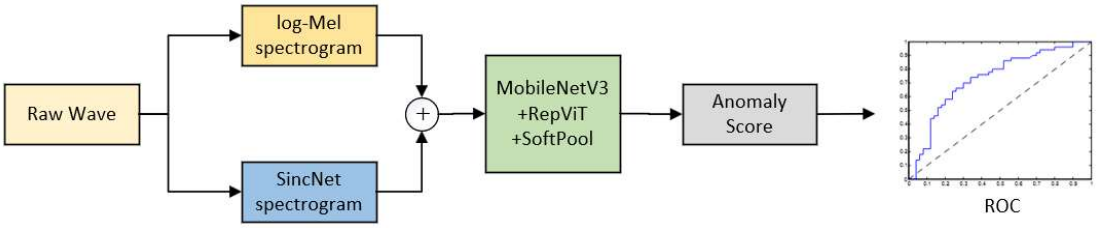

Fig. 1 Technology Roadmap

2. Model Construction

2.1 Feature Extraction

Log-Mel mimics the characteristics of the human ear, good at capturing the stable low-frequency patterns in the operation of the equipment, SincNet through independent learning of high-frequency bands, can quickly capture sudden abnormalities, and even perceive the subtle timing changes in the waveform of the sound, the combination of the two, the two complement each other, not only from the overall judgment of the equipment whether the "rhythm of normal", but also instantly identify "sudden murmurs". It can also instantly recognize "sudden noises", for example, in the wind turbine power generation scenario, it can not only find the continuous noise caused by bearing wear, but also capture the metal impact sound of the moment of screw loosening.

Acoustic feature extraction is a key step in abnormal sound detection. The original waveform is divided into frames and windowed, and then fast Fourier transform and Mel filtering are performed to obtain the Mel spectrogram, and at the same time, feature extraction is performed through the SincNet filter, and then normalization, ReLU activation, and adaptive averaging pooling are processed to the SincNet spectrogram. Then the SL spectrogram is fused by feature splicing to obtain the SL spectrogram.The extraction process of the SL spectrogram is shown in Fig. 2.

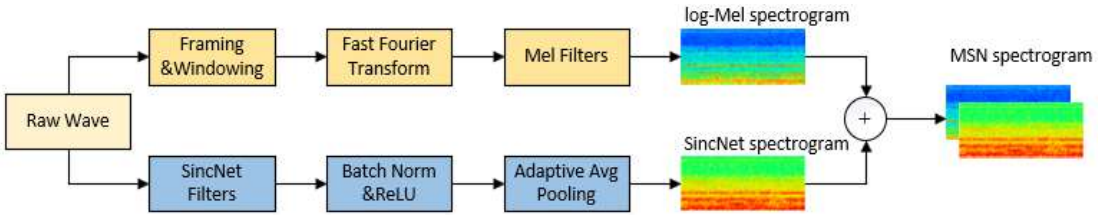

Fig. 2 The Extraction Process of the SL Spectrogram

2.1.1 Log-Mel Extraction

The generation process of Log-Mel spectrogram can be divided into three steps: firstly, the original time-domain signal is processed by frame-plus-window, and the short-time power spectrum is calculated by Fast Fourier Transform (FFT); subsequently, the power spectrum is fed into the preset Mel filter bank, and the energy of each frequency band is integrated; finally, the energy of Mel scale is taken as a logarithmic compression, which enhances the signal components with a smaller dynamic range. For example, let the input signal be, after the frequency domain representation is obtained by FFT, the energy integration of its Mel filter bank can be expressed as equation (1):

$$S(m)=\left(\sum|X(k)|^2\cdot H_m(k)\right) \tag{1}$$

where $H_m(k)$ is the transfer function of the m nd Mel filter. This process may weaken the expressive power of high-frequency transient components while preserving the global spectral characteristics of acoustic events. Therefore, Log-Mel spectrograms often need to be fused with other high-resolution frequency-domain features to improve the robustness of anomalous sound detection in complex industrial scenarios[5].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

2.1.2 SincNet Extraction

In the field of acoustic feature extraction, the SincNet spectrogram is used as a parametric filter bank method with a core design based on modeling the time-domain impulse response of an ideal bandpass filter. Unlike filters based on Mel's auditory model, SincNet constructs banding mechanisms with clear physical meaning by directly learning low or high cutoff frequency parameters. The method significantly reduces model complexity by limiting the number of filter parameters, while improving feature discrimination in noisy environments[6].

The process of constructing the SincNet spectrogram can be divided into four stages: firstly, the frequency domain response function of the ideal bandpass filter is defined, and its mathematical expression is the difference between the two rectangular window functions, as in Equation (2):

$$G(f, f_1, f_2) = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right) \tag{2}$$

where and denote the lower and upper cutoff frequencies of the bandpass filter, respectively. The inverse Fourier transform of this frequency domain response yields the time domain impulse response function as in equation (3):

$$g(n) = 2f_2 \cdot sinc(2\pi f_2 n) - 2f_1 \cdot sinc(2\pi f_1 n) \tag{3}$$

In order to avoid spectral energy leakage, the infinitely long sinc function needs to be truncated by adding a window. The Hamming window function is used to smooth the time domain impulse response as in Eq. (4):

$$g^w(n) = g(n) \cdot w(n) \tag{4}$$

Subsequently, the original signal is convolved with each SincNet filter in a convolution operation and processed by the ReLU activation function with batch normalization to obtain the multichannel filtered output as in Equation (5):

$$x_k[n] = ReLU\left(BN\left(x[n] * g_k^w(n)\right)\right) \tag{5}$$

Finally, to unify the feature dimensions, Adaptive Average Pooling (AAP) is performed on each filter output to generate a time-frequency matrix that matches the dimensions of the log-Mel spectrogram. This operation can be formulated as in Eq. (6):

$$m_k[n] = AdaptiveAugPool(x_k[n]) \tag{6}$$

Compared with traditional filter banks, SincNet spectrogram has two major advantages: first, it can autonomously capture subtle frequency domain patterns in device operation by optimizing the band division through an end-to-end learning mechanism; second, the time-domain convolution operation retains the signal phase information, which provides a more sensitive response characteristic to shock-type anomalies. However, its bandwidth resolution is limited by the number of preset filters, which may not be able to fully cover broadband noise interference. Therefore, it is often used in conjunction

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

with log-Mel spectrograms to construct more discriminative composite acoustic representations through the complementary fusion of global features and local details in the frequency domain[7].

2.1.3 Spectrogram Fusion

In the field of acoustic feature fusion, the multimodal spectrogram integration strategy can effectively improve the anomaly detection performance in complex scenes by combining the advantages of different frequency domain representations. In this paper, we propose a composite spectrogram construction method, which splices Log-Mel spectrograms and SincNet spectrograms with cross-domain features to form a fusion feature matrix with complementary characteristics. The specific fusion process is shown in Fig. 2, where the two spectrograms are first generated separately by the parallel feature extraction module, followed by tensor splicing along the channel dimension as in Eq. (7):

$$F_{fusion}(c,t)\begin{cases} F_{Mel}(m,l), c \le C_{Mel} \\ FSinc(s,t), C_{Mel} < c \le C_{Mel} + C_{Sinc} \end{cases} \tag{7}$$

where C_{Mel} and C_{Sinc} denote the channel dimensions of the two types of spectrograms, respectively, and F_{Mel} preserves the global band energy distribution while F_{Sinc} focuses on the local band transient response. Using this fusion method, the advantages of Log-Mel spectrograms and SincNet spectrograms can be preserved so that the two types of spectrograms complement each other effectively in the frequency domain coverage[8].

2.2 Optimization of Anomaly Detection Network based on MobileNetV3

2.2.1 Network Structure Improvements and Module Design

MobileNetV3, as a representative of lightweight convolutional neural network, integrates deep separable convolution, linear bottleneck structure and inverted residual module, which ensures computational efficiency while realizing high feature extraction capability. However, in the abnormal sound detection task in complex industrial scenes, its original attention mechanism and pooling layer design still suffer from the problems of feature information loss and insufficient capture of high-frequency details. For this reason, this study proposes an improved anomaly detection network that optimizes the feature expression and information retention capability of the network by incorporating a reparameterized visual transform module and a soft pooling method.

2.2.2 Introduction of the RepViT Module

RepViT dynamically fuses spatial-channel features through multi-branch structures, avoiding the loss of high-frequency information due to channel compression in the SE module, and at the same time reduces inference computation by 30% through the reparameterization design.RepViT is based on the idea of structure reparameterization, which enhances the feature diversity through multi-branch design in the training phase and merges them into a single efficient structure in the inference phase. The module combines the advantages of local convolution operation and global attention mechanism, and can dynamically adjust the correlation between feature channels. In the specific implementation, RepViT is embedded into the inverted residual structure of MobileNetV3, replacing the original SE attention module. Compared with the channel compression operation of the SE module, RepViT extracts spatial-channel features in parallel through multi-branch convolution and self-attention mechanism, avoiding feature loss due to dimensionality reduction. Meanwhile, its heavy parameterization feature is transformed into a lightweight single-path structure during inference, which significantly reduces the computational complexity and is suitable for real-time wind motor gearbox anomaly detection[9].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

2.2.3 Working Principle of SoftPool

The core idea of SoftPool is to convert the activation values within the pooling window into weights by Softmax function, and then sum the activation values by weighting them.In Softmax transformation, for each activation value within the pooling window, calculate its exponent value, and then get the weights by normalization, and then multiply the activation values within the pooling window with the corresponding weights and then sum them up to get the final pooling result.

Traditional pooling operations are prone to lose high-frequency detail information during downsampling, which affects the capture of subtle features of abnormal sounds. For this reason, SoftPool is used in this study to replace the original global average pooling layer.SoftPool fuses local features through exponential weighting to retain more effective information. Fig. 3 shows the schematic diagram of the SoftPool weighted pooling process. Its calculation process can be described as equation (8):

$$\omega_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}}, \tilde{a} = \sum_{i \in R} \omega_i \cdot a_i \tag{8}$$

where a_i is the activation value of the i rd feature point in the feature region R and ω_i is the normalized weight. The method strengthens the key feature response while suppressing noise, and is especially suitable for time-frequency structure extraction of non-stationary mechanical acoustic signals.

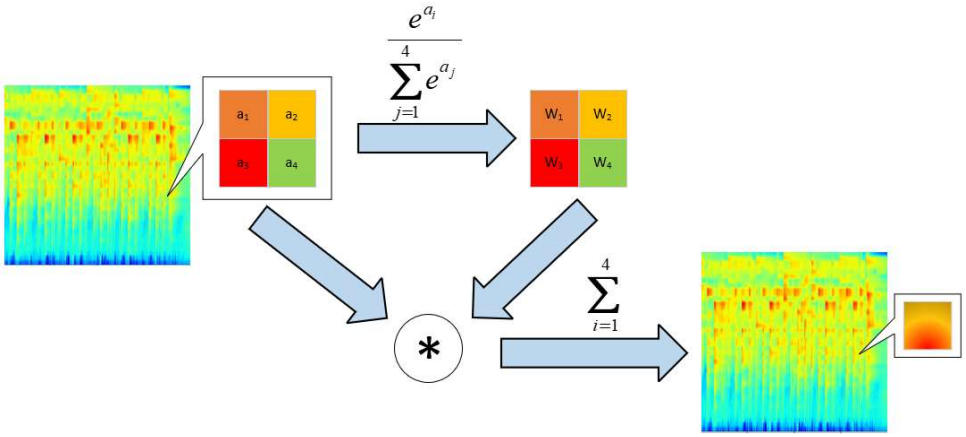

Fig. 3 Calculation Process of SoftPool

2.2.4 Overall Network Architecture

In this paper, we propose an improved lightweight network architecture, RS-MobileNetV3, optimally designed for the task of abnormal sound detection in wind motor gearboxes. The network replaces the traditional attention mechanism by introducing RepViT, enhances the feature interaction ability by using multi-branch structure in the training phase, and merges into a single-path structure to maintain computational efficiency in inference; meanwhile, SoftPool is used to optimize the downsampling process, which significantly reduces the loss of high-frequency details. Combined with the multi-scale feature fusion strategy, the architecture can achieve a balance between detection accuracy and lightweight characteristics in complex noise environments, providing an efficient solution for real-time anomaly detection in industrial scenarios. The overall architecture of the improved network is shown in Fig. 4.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

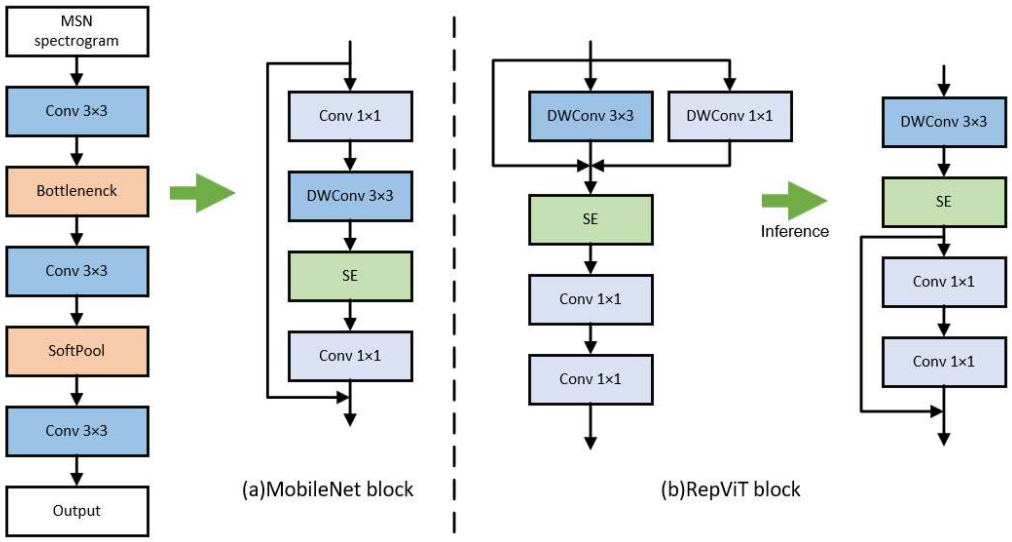

Fig. 4 RS-MobileNetV3 Network Architecture

3. Experimental Data and Analysis

3.1 Dataset and Evaluation Metrics

The experiment uses the NREL public dataset from the National Renewable Energy Laboratory (NREL), which contains wind turbine vibration and sound data covering samples from key components such as gearboxes and bearings. The dataset contains 12,000 samples, with 8,600 samples of normal operating sound as the training set. The data containing abnormal sound is 3600 pieces, which is used as a test set. The duration of a single sample is 10 seconds and covers a background wind noise environment of 20-90 dB, which highly reproduces the complex working conditions of wind farms. The evaluation indexes use AUC and pAUC, which can verify the performance advantage of the model[10].

3.2 Experimental Setup

The input features are SL spectrogram maps with a dimension of 128×313×2. Training is performed using Adam optimizer with β1=0.9,β2=0.999, an initial learning rate of 1e-5, and a Batch Size=64 for a total of 200 rounds. Data enhancement includes time domain stretching (±10%), Gaussian noise injection (SNR=10-20 dB) and frequency domain masking (bandwidth ≤20%) to improve model robustness. The development platform used for the experiments was Pycharm, conducted using PyTorch 1.13.0 and Python 3.7.13, training was done on NVIDIA RTX 3090 GPUs, and the edge deployment tests were based on a Jetson Nano with 4 GB of RAM.

3.3 Ablation Study

To validate the contribution of RepViT and SoftPool, the following comparison experiments were conducted:

Table 2. Ablation Comparative Experiments

Model Configuration	Params (M)	FLOPs (G)	AUC(%)	pAUC(%)
MobileNetV3	2.10	0.423	89.16	86.12
MobileNetV3+RepViT	2.26	0.439	91.37	87.92
MobileNetV3+SoftPool	2.21	0.428	90.62	88.45
RS-MobileNetV3	2.48	0.419	93.69	91.34

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

On the NREL dataset, the RepViT module improves the AUC by 2.21%, while the SoftPool substitution optimizes the pAUC by 1.33%, and the joint optimization architecture of the two results in a performance gain of 4.53/5.22 percentage points over the baseline model with an increase in the number of covariates by 18.1%, which verifies the effectiveness of the multi-module co-optimization in terms of lightweighting and noise robustness.

3.4 Comparative Experiments

Table 3. Performance comparison experiment

Mould	AUC	pAUC	Params (M)	FLOPs (G)	Training time(min)	Marginal inference(ms)
Mel-CNN	89.31	83.52	1.80	0.150	142	382
MobileNetV2	88.75	81.23	3.40	0.290	198	415
MobileNetV3	89.16	86.12	2.10	0.190	156	388
ResNet-18	92.58	87.94	11.70	2.600	426	972
AutoEncoder	85.61	76.88	12.10	6.420	598	-
Isolation Forest	78.40	70.34	-	-	-	-
RS-MobileNetV3	93.69	91.34	2.48	0.230	218	392

As shown in Table 3, this paper's method outperforms the comparison model in both core indicators of AUC and pAUC, and the number of parameters is 62.3% less than that of ResNet-18, and the computational volume is reduced by 0.03G, which verifies the superiority of the model. The enhancement of high-frequency transient features by RepViT module and SoftPool is demonstrated.

4. Conclusion

In this paper, we propose a lightweight network for abnormal sound detection of wind turbines, which improves the feature extraction capability and robustness of the model by fusing the RepViT module in MobileNetV3 and optimizing the pooling layer with SoftPool. The experiments are validated based on the NREL dataset, and the improved model improves 4.53% in average AUC over the original MobileNetV3, and the number of parameters only increases by 18.1%. The ablation experiments show that the incorporation of DFC module and SoftPool contributes significantly to anomaly detection, and the feature fusion splicing strategy balances efficiency and accuracy. It provides an efficient and low-cost solution for wind farm operation and maintenance.

References

[1] Jiarui L ,Guotian Y ,Xinli L , et al.Wind turbine anomaly detection based on SCADA: A deep autoencoder enhanced by fault instances.[J].ISA transactions,2023,139586-605.

[2] Yida W ,Joseph D T ,Nassir N , et al.SoftPool++: An Encoder–Decoder Network for Point Cloud Completion[J].International Journal of Computer Vision,2022,130(5):1145-1164.

[3] Chunyuan W ,Yang W ,Yihan W , et al.Scene Recognition Using Deep Softpool Capsule Network Based on Residual Diverse Branch Block[J].Sensors,2021,21(16):5575-5575.

[4] Nantian H ,Qingzhu C ,Guowei C , et al.Fault Diagnosis of Bearing in Wind Turbine Gearbox Under Actual Operating Conditions Driven by Limited Data With Noise Labels[J].IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT,2021,70

[5] Ziqiang P ,Chuan L ,Shaohui Z , et al.Fault Diagnosis for Wind Turbine Gearboxes by Using Deep Enhanced Fusion Network[J].IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT,2021,70

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0016

[6] Renström N ,Bangalore P ,Highcock E .System-wide anomaly detection in wind turbines using deep autoencoders[J].Renewable Energy,2020,157(prepublish):647-659.

[7] Feng Z ,Zhu W ,Zhang D .Time-Frequency demodulation analysis via Vold-Kalman filter for wind turbine planetary gearbox fault diagnosis under nonstationary speeds[J].Mechanical Systems and Signal Processing,2019,12893-109.

[8] Cao L ,Qian Z ,Zareipour H , et al.Fault Diagnosis of Wind Turbine Gearbox Based on Deep Bi-Directional Long Short-Term Memory Under Time-Varying Non-Stationary Operating Conditions.[J].IEEE Access,2019,7155219-155228.

[9] Wang A, Chen H, Lin Z, et al. RepViT-SAM: Towards Real-Time Segmenting Anything[J]. arXiv preprint arXiv:2312.05760, 2023.

[10] 3TIER, NREL. Western Wind Integration Data Set[R]. National Renewable Energy Laboratory, 2022.