Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 15, No. 3, 2025

A Lightweight Method Integrating Dynamic Frequency-Aware Convolution and SoftPool for Abnormal Sound Detection in Wind Turbines

Qingzheng Li

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

Abstract: Aiming at the problems of insufficient feature expression capability and high model computation complexity in traditional wind turbine group abnormal sound detection methods, this paper proposes a lightweight detection method based on improved MobileNetV3 network. First, the SincNet bandpass filter and Mel spectrum are integrated to construct multi-dimensional acoustic features, taking into account the original signal time-domain features and frequency-domain features. Second, a dynamic frequency-aware convolution (DFC) module is introduced into the MobileNetV3 network architecture to adaptively adjust the parameters of the convolution kernel through the frequency-domain attention mechanism to strengthen the frequency feature capture of abnormal sounds; the feature downsampling process is optimized by combining with SoftPool to reduce the loss of high-frequency information. On the dataset of Danish University of Science and Technology, the AUC reaches 94.71%, and the number of parameters is only 2.38M, which is 62.3% lower than the mainstream model ResNet-18, providing a high-precision edge-end solution for wind turbine status monitoring.

Keywords: Wind Turbines, Abnormal Sound Detection, MobileNetV3.

1. Introduction

Wind turbines, as core equipment in clean energy generation, have operational stability that directly impacts power generation efficiency and maintenance costs. Traditional vibration monitoring methods are susceptible to environmental interference and incur high installation and maintenance expenses. Industrial noise exhibits complex characteristics such as wind noise and electromagnetic interference, rendering conventional detection methods insufficiently sensitive to high-frequency anomalies and unknown fault types. In contrast,acoustic analysis technology enables non-contact detection by capturing operational sounds, effectively identifying mechanical damage.

Abnormal Sound Detection (ASD)systems, based on acoustic signal analysis and pattern recognition technologies, have gained widespread attention in the field of industrial intelligent monitoring. This recognition stems from their non-intrusive monitoring nature, real-time responsiveness, and adaptability to complex acoustic environments. Currently, ASD has been successfully applied in diverse scenarios including livestock health monitoring, industrial equipment condition assessment, and medical diagnostic assistance.

In wind turbine abnormal sound detection, supervised methods rely on labeled fault samples and achieve high accuracy for known anomalies, but they face two major limitations, The acquisition of abnormal samples through destructive experiments is costly and impractical for industrial applications,Model performance significantly degrades after cross-condition transfer due to domain shifts in operational environments. In contrast, unsupervised methods only require normal sound data for modeling, effectively overcoming data scarcity constraints. For instance, autoencoders can detect blade cracks through reconstruction error analysis while integrating noise suppression techniques to mitigate wind noise-induced misclassification. This study adopts an unsupervised approach because industrial scenarios inherently provide abundant normal operation data, and such methods enable identification of unknown anomalies without dependency on predefined fault categories, addressing the critical challenges of supervised models.The unsupervised anomaly sound system monitoring framework is illustrated in Figure 1.

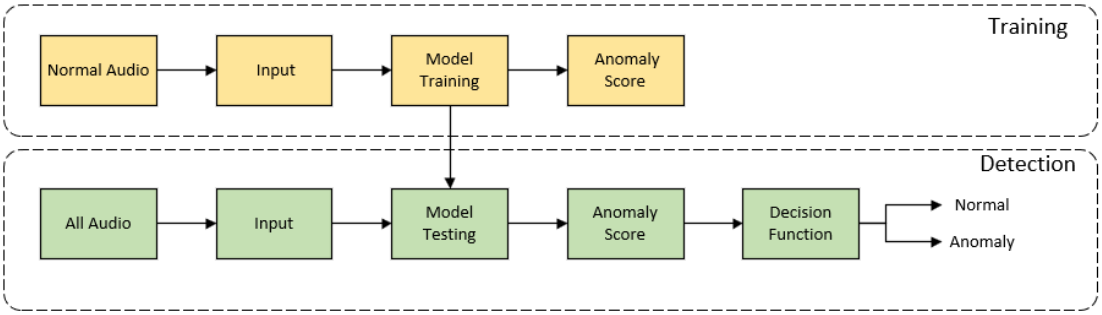

Figure 1. Unsupervised Anomalous Sound System Monitoring Framework

Acoustic feature extraction techniques have undergone continuous evolution from traditional methods to deep learning approaches. In anomaly detection tasks, feature extraction plays a critical role. Early widely adopted Mel-

frequency cepstral coefficients (MFCC), though computationally efficient, demonstrated limitations in high-frequency feature preservation due to their anthropomorphic auditory modeling bias toward low-frequency characteristics. This deficiency becomes particularly pronounced in noisy industrial environments, significantly constraining the robustness of conventional approaches.

To address high-frequency information loss, researchers have introduced deep learning-based solutions like SincNet, which employs learnable bandpass filters to substantially enhance the capture of critical high-frequency features. This study adopts a hybrid strategy integrating Log-Mel spectrograms with SincNet spectrograms. This synergistic combination leverages Log-Mel's superior low-frequency sound capture capability in industrial production environments and SincNet's high-frequency resolution advantages, effectively enhancing feature representation in complex acoustic scenarios [1].

With the development of machine learning technology, many studies on abnormal sound detection in wind turbines have adopted deep learning methods. For example, Zhang et al [2] proposed a gearbox fault detection method based on time-frequency analysis, which effectively dealt with the problem of weak fault feature extraction in high-noise environments.Wang et al [3] realized anomaly detection of wind turbine gearboxes by reconstructing the error using a deep selfencoder, but it was still deficient in high-frequency feature capture.Deng et al [4] designed a lightweight convolutional neural network for real-time industrial anomaly detection, but the robustness in complex noise environments needs to be improved.MobileNetV3, as a lightweight convolutional neural network, achieves efficient feature extraction while maintaining computational efficiency through techniques such as depth-separable convolution and dynamic frequency-aware convolution. Dynamic frequency-aware convolution can dynamically adjust the shape and size of the convolution kernel according to the frequency distribution of the input signal to better capture the time-frequency features.The network structure of MobileNetV3 fused with dynamic frequency-aware convolution has stronger robustness and feature extraction ability in complex noise environments.Therefore, in this paper, MobileNetV3 is chosen as the base network and combined with the dynamic frequency-aware convolution to improve the performance of abnormal sound detection of wind turbines.

SoftPool performs weighted aggregation of the activation values within the pooling window through the Softmax function, which retains more high-frequency detailed features compared to traditional pooling methods [5]. This method shows the advantage of sensitivity to transient components in acoustic signal processing, and is especially suitable for feature extraction in non-smooth scenes. Aiming at the problem that high-frequency transient impulses are easily lost in the abnormal sound of wind turbines, this paper adopts SoftPool instead of the global pooling layer in the MobileNetV3 network to strengthen the model's parsing ability for key acoustic features.

In summary, this paper chooses to use a new combination of acoustic features in sound extraction by fusing Mel spectrograms and SincNet spectrograms into MS spectrograms, which is a method that can be better adapted to the noisy environment in the working environment of wind turbines. In terms of neural network, this paper proposes an improved machine anomalous sound detection network DS-MobileNetV3, which enhances the feature processing capability and robustness of the network by adding dynamic frequency-aware convolution and combining with SoftPool pooling method. The technology roadmap is shown in Figure. 2.

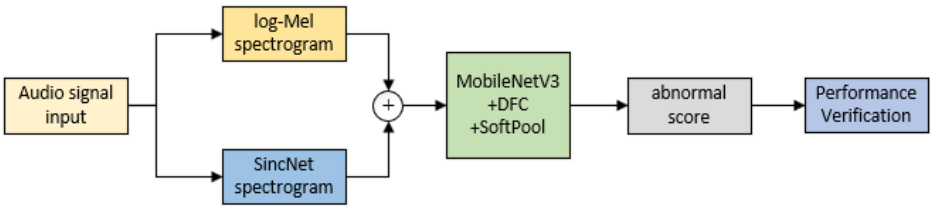

Figure 2. Technology roadmap

2. Methods

2.1. Feature extraction

The Log-Mel spectrogram highlights low-frequency harmonic features based on the human ear's auditory characteristics, but its fixed filter has insufficient resolution in the high-frequency band, making it difficult to capture transient impact signals, while the SincNet spectrogram enhances high-frequency feature resolution under noisy environments by directly modeling band parameters through learnable band-pass filters. The MS spectrogram is formed by combining the advantages of the two: Log-Mel retains the low-frequency energy structure, and SincNet strengthens the high-frequency transient response, forming a feature expression that is complementary to the wide-frequency coverage and local sensitivity, and enhancing the robustness of the detection of the wind turbine's complex working conditions.The extraction process of the MS spectrogram is shown in Figure. 3.

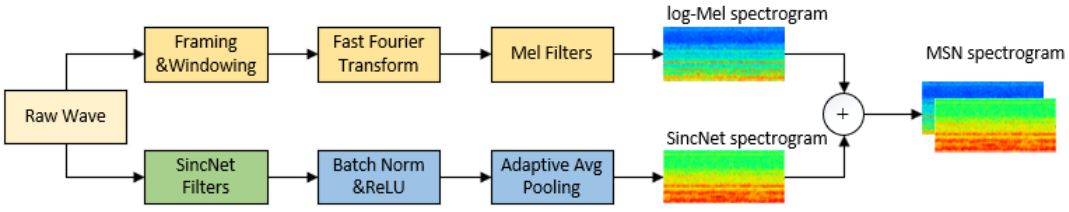

Figure 3. MS spectrogram extraction process

2.1.1. Log-Mel extraction

Log-Mel spectrograms are suitable for sound detection in industrial production environments, and in wind turbine abnormal sound detection, Log-Mel spectrograms are used as audio feature representations, which are required to be able to capture the key information of the sound signal. Its extraction process firstly, the original audio signal is divided into frames, and the long signal is divided into short-time frames, and the length of each frame is N sampling points.Next, a Hamming window $w(n)$ is applied to each frame to minimize the spectral leakage, denoted by the formula (1):

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (1)$$

Then, the fast Fourier transform is applied to the windowed signal to obtain the spectrum of each frame. $X(k)$ In order to simulate the human auditory system's perception of frequency, the linear spectrum is converted to a Mel spectrum. The conversion relationship between Mel frequency and actual frequency is shown in Equation (2):

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

Where f is the actual frequency and m is the Mel frequency. The spectrum is then filtered by a bank of Mel filters, each covering a certain range of Mel frequencies. After filtering, the logarithmic energy of the output of each filter is calculated to obtain the Log-Mel spectrum. Finally, the multi-frame Log-Mel spectrum is stitched into a two-dimensional matrix as the Log-Mel spectrogram of the audio signal [6].

2.1.2. SincNet extraction

SincNet is a neural network architecture specialized in audio processing that provides a way to capture the time-domain features of audio signals by using sinc function-based filters in the first convolutional layer. Unlike standard CNNs, which learn all elements of each filter, SincNet only learns the low cutoff frequency and the high cutoff frequency of the bandpass filter. Learning the high cutoff frequency compensates for Log-Mel's insensitivity to high frequencies.

The SincNet workflow mainly consists of preprocessing, designing a set of Sinc filters, and using the Sinc filter bank to convolve the audio signal to extract features in each frequency band.The SincNet filter feature extraction operation is shown in equation (3):

$$x_k[n] = x[n] \cdot g_k^w(n), k = 1,2,\ldots,K \quad (3)$$

where $x[n]$ denotes the input raw sound signal, K denotes the number of SincNet filters, and $g_k^w(n)$ denotes that it is the kth SincNet filter. Subsequently, one-dimensional batch normalization and ReLU nonlinear activation functions are applied to obtain the filter outputs. In order to make the SincNet spectrograms have the same dimensional size as the log-Meier spectrograms, Adaptive Average Pooling is used on the output of each filter $x_k[n]$, which automatically adjusts the size of the pooling region and the step size according to the target output size. The operation procedure is as in equation (4):

$$m_k[n] = AdaptiveAugPool(x_k[n]) \quad (4)$$

The SincNet spectrogram optimizes band segmentation through end-to-end learning, autonomously captures subtle frequency-domain patterns of devices, and preserves phase information in time-domain convolution to improve sensitivity to shock-type anomalies. However, its band resolution is limited by the number of filters, which makes it difficult to cover broadband noise interference. Therefore, this paper fuses SincNet and Log-Mel spectra: Log-Mel provides global energy distribution, and SincNet focuses on local high-frequency details, forming complementary acoustic representations to enhance the discriminative power.

2.1.3. Spectrogram Fusion

In the study of acoustic feature fusion, the multimodal spectrogram integration strategy can improve the accuracy of anomaly detection in complex scenes by integrating features from different frequency domains. In this paper, Log-Mel spectrograms and SincNet spectrograms are spliced with cross-domain features to form a fusion feature matrix with complementary characteristics. As in equation (5):

$$F_{fusion}(c,t)\begin{cases}\dfrac{F_{Mel}(m,t)-\mu_{Mel}}{\sigma_{Mel}}, c \in [1, C_{Mel}] \\ \dfrac{F_{Sinc}(s,t)-\mu_{Sinc}}{\sigma_{Sinc}}, c \in [C_{Mel}+1, C_{Mel}+C_{Sinc}]\end{cases} \quad (5)$$

where μ_{Mel}, σ_{Mel} and $\mu_{Sinc}, \sigma_{Sinc}$ represent the channel statistics, 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 localized band transient response. Using this fusion method preserves the advantages of Log-Mel spectrograms and SincNet spectrograms, allowing the two spectrograms to complement each other in the frequency domain coverage [7].

2.2. Anomaly detection network optimization based on MobileNetV3

2.2.1. Introduction of the DFC module

In order to enhance the model's ability to extract frequency domain features of sound signals, this paper introduces the Dynamic Frequency Convolution (DFC) module into the MobileNetV3 network structure. This module performs multi-scale frequency analysis of the input signal by means of a learnable frequency domain filter bank, which, combined with lightweight parameter design, enhances the network's ability to capture high-frequency details without significantly increasing the computational volume [8].The DFC module adopts a dynamic weight allocation mechanism, which adaptively adjusts the filter parameters according to the

characteristics of the input spectrum, effectively extracting the non-smooth high-frequency transient components of the anomalous sound. The DFC module is connected in parallel with the depth-separable convolutional layer of MobileNetV3 to form a complementary feature fusion structure, which takes into account the efficiency of the model operation and the ability to fine-tune the expression of acoustic features.The structure diagram of the DFC module is shown in Figure. 4.

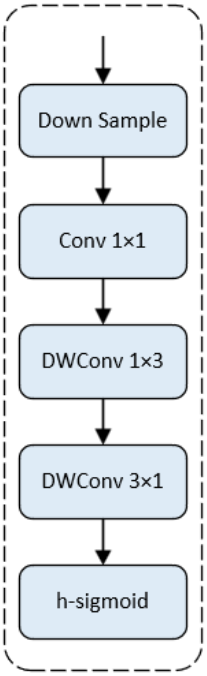

Figure 4. DFC module diagram

2.2.2. Integration of the SoftPool

In order to improve the model's ability to extract acoustic features in a fine-grained way, this study introduces SoftPool to replace the traditional pooling layer in MobileNetV3. This pooling method adopts an exponentially weighted average strategy, which reduces the spatial resolution through weighted aggregation on the basis of retaining the consistency of the input and output dimensions, and reduces the loss of high-frequency details compared with maximum pooling [9]. Aiming at the significant characteristics of high-frequency transient features in the wind turbine abnormal sound detection task, SoftPool can effectively alleviate the information degradation problem of traditional pooling and enhance the robustness. Figure 5 shows the schematic diagram of the SoftPool weighted pooling process. Its computational process can be described as Equation (6):

$$\omega_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}} \tag{6}$$

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.

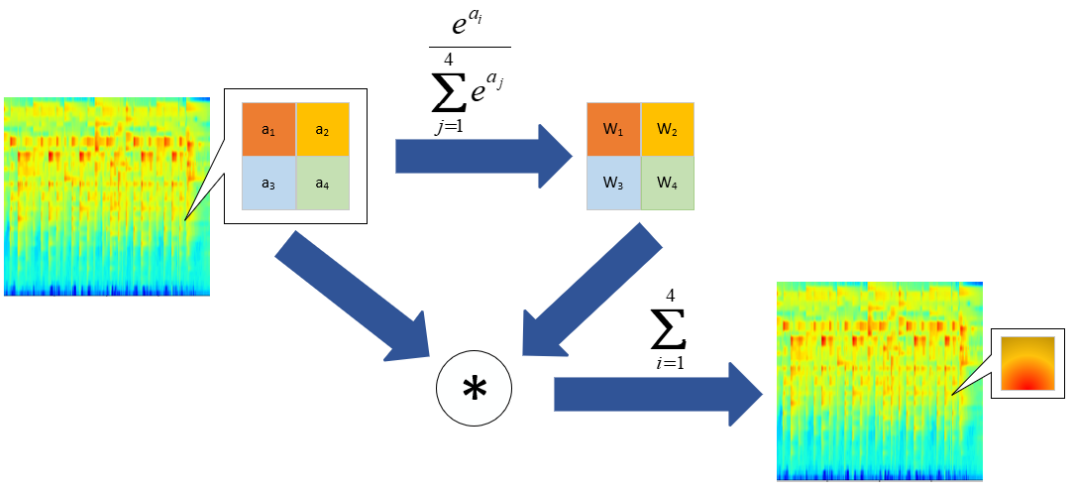

Figure 5. SoftPool calculation process

2.2.3. Network architecture

In this paper, we integrate the dynamic frequency convolution module and optimize the pooling layer design on the basis of MobileNetV3 network to construct a lightweight acoustic feature extraction network, and the overall architecture is shown in Figure 6.

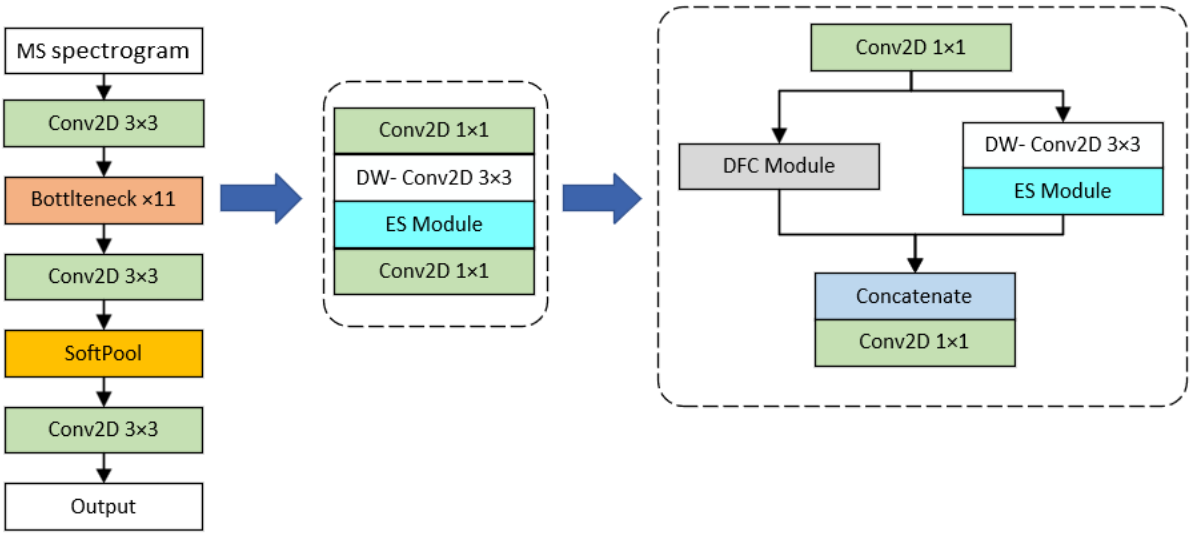

Figure 6. DS-MobileNetV3 network architecture

The input of the network is the pre-processed audio spectrogram, and after the shallow features are extracted by the initial convolutional layer, the multi-level feature abstraction is realized by the improved bottleneck modules at multiple levels. Each bottleneck module adopts a two-branch parallel structure: the main branch retains MobileNetV3's original depth-separable convolution and SE attention mechanism for spatial feature extraction and channel dimensionality recalibration; the newly added DFC module performs multiscale frequency analysis of the input signal through a dynamic frequency-domain filter bank, focusing on capturing the high-frequency transient components of the anomalous sound. The two-branch output is feature spliced and dimension matching is realized by 1x1 compression convolution.

At the end of the network, SoftPool is used to replace the conventional global average pooling layer to avoid the feature smoothing problem caused by conventional pooling operations. This design significantly improves the resolution of high-frequency acoustic features while maintaining the efficient computational characteristics of MobileNetV3, which is suitable for turbine anomaly detection scenarios under complex working conditions.

3. Experimental Data and Analysis

3.1. Data set

The experiment uses the wind turbine acoustic data set [10] from the Danish University of Science and Technology, which contains the sound signals of the three core components, namely, gearbox, bearings, and blades, under abnormal and normal conditions. The total number of sound signal samples is 8760, of which the number of samples containing only normal working condition sound data is 6160 as the training set. The number of samples containing abnormal sound data is 2600, which is used as the test set. The duration of a single sample is 5 seconds, the sampling rate is 44.1 kHz, and the background wind noise environment is covered by 40-70 dB, which highly reproduces the complex working conditions of wind farms.

3.2. Experimental environment

The input features are MS spectrograms with a dimension of 128×313×2, and the joint characterization of high-frequency transient and low-frequency harmonic features is enhanced by frequency domain complementarity. Training is performed using the 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. The data enhancement includes time domain stretching, Gaussian noise injection and frequency domain masking to improve the model robustness. The development platform used for the experiments is Pycharm, conducted using PyTorch 1.13.0 and Python 3.7.13, with training done on NVIDIA RTX 3090 GPUs, and edge deployment testing based on a Jetson Nano with 4GB of RAM.

3.3. Evaluation indicators

Enhanced Accuracy Gain (AUC) and Partial Enhanced Accuracy Gain (pAUC) are used as the core classification metrics to evaluate the model performance. aUC is defined as the area under the ROC curve, which evaluates the overall anomaly detection capability. pAUC is defined as the AUC value of the restricted high-frequency anomaly samples, which is calculated in the interval [0, p] where the false-positive rate (FPR) is low. aUC reflects the the magnitude of the model's classification accuracy improvement on the noise-enhanced test set, and calculates the accuracy difference between the improved model and the baseline model. pAUC further refines the assessment of the proportion of gain in high-frequency anomalous samples, and measures the model's sensitivity to critical fault features. The experiments also count the number of parameters, inference elapsed time, and edge inference, and assess metrics such as computational power and reliability to ensure that the improved solution meets the lightweight deployment requirements.

3.4. Ablation experiments

To validate the contribution of DFC and SoftPool, the following comparison experiments are designed:

Table 1. Ablation comparison experiments

Model Configuration	Params (M)	FLOPs (G)	AUC(%)	pAUC(%)
MobileNetV2	3.50	0.535	83.61	-
MobileNetV3	2.10	0.423	89.14	-
MobileNetV3+DFC	2.28	0.439	92.37	87.92
MobileNetV3+SoftPool	2.15	0.428	91.62	88.45
DS-MobileNetV3	2.38	0.420	94.71	92.26

The effects of the DFC and SoftPool modules are verified by a module-by-module comparison on a machine anomaly sound dataset from the Technical University of Denmark. As shown in Table 1, the DFC module alone improves the AUC by 3.23%, SoftPool replaces the original pooling operation to improve the AUC by 2.48%, and the two modules synergistically produce a gain-stacking effect of 1.35%, and the number of parameters increases by only 4.8%, which experimentally proves the feasibility of the method of this paper.

3.5. Comparative tests

In order to verify the superiority of the proposed method, a comparative analysis of the machine anomaly detection performance is carried out based on the wind turbine acoustic dataset from the Danish University of Science and Technology, comparing the method in this paper with other mainstream methods. The comparison results are shown in Table 2.

Table 2. Model comparison experiments

Mould	AUC	pAUC	Params (M)	FLOPs (G)	Training time (min)	Marginal Reasoning (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	91.16	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	-	-	-	-
DS-MobileNetV3	94.71	92.26	2.38	0.210	168	382

As shown in Table 2, this paper's method outperforms the comparison model in both core indexes of AUC and pAUC, and the number of parameters is 62.3% less than that of ResNet-18, and the computation amount is reduced by 0.03G, which verifies the superiority of the model. The enhancement of high-frequency transient features by the DFC module and SoftPool is demonstrated.

4. Conclusions

In this paper, we propose a lightweight network for abnormal sound detection in wind turbines by incorporating the dynamic frequency convolution module and optimizing the pooling layer with SoftPool in MobileNetV3. The DFC module innovatively adopts the frequency band adaptive mechanism, enhances the ability of high-frequency transient feature extraction with a learnable band-pass filter, and SoftPool significantly reduces the loss of high-frequency details. The experiments are validated based on the acoustic dataset of Danish University of Science and Technology, and the improved model improves 5.57% in average AUC over the original MobileNetV3, with only 4.8% increase in the number of parameters. 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] Wang M ,Mei Q ,Song X , et al.A Machine Anomalous Sound Detection Method Using the lMS Spectrogram and ES-MobileNetV3 Network[J].Applied Sciences,2023,13(23):

[2] Shipeng H ,Yihang C ,Zhifang W , et al.Deep learning bird song recognition based on MFF-ScSEnet [J]. Ecological Indicators, 2023,154

[3] Yixing F ,Chunjiang Y ,Yan Z , et al.Classification of birdsong spectrograms based on DR-ACGAN and dynamic convolution [J]. Ecological Informatics,2023,77

[4] Liu H ,Wang H .Real-Time Anomaly Detection of Network Traffic Based on CNN[J].Symmetry,2023,15(6):

[5] 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.

[6] 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.

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

[8] 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.

[9] Meng H ,Yan T ,Yuan F , et al.Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network.[J].IEEE Access,2019,7125868-125881.

[10] Emre B ,Jun W Z ,Zhong W S , et al.Consistent modelling of wind turbine noise propagation from source to receiver.[J].The Journal of the Acoustical Society of America, 2017, 142(5): 3297.