Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

EfficientNetV2 Pump Anomaly Detection Method Based on Improved CBAM Attention Mechanism

Fuyue Sun SUNFUYUE2022@163.COM
International School, Beijing University of Posts and Telecommunications, Beijing, 100876, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

In the process of industrial production, pump as the core equipment, its stable operation directly affects the safety and production efficiency of the factory. However, in practical applications, pumps often face various abnormal conditions, which may lead to equipment failure and even safety accidents in serious cases. To solve this problem, we design a new anomaly detection method, which improves the EfficientNetV2 network: The original Attention Module is replaced with the optimized CBAM module, its channel attention is retained to capture the cross-channel dependencies, and the Simple Attention Module (SimAM) is introduced into the spatial attention part to effectively reduce the computational complexity and enhance the sensitivity of the model to local details and global context information. In order to better deal with the problem of data imbalance, we use MixUp data augmentation and Label Smoothing regularization strategy in the training process, and choose BCEWithLogitsLoss as the loss function. In the pre-training phase, the pump-related audio modules in the MIMII dataset are used for weight initialization, which is subsequently fine-tuned on the pump anomaly binary classification task. Experimental results show that the proposed model improves the classification accuracy by 0.76% compared with the original EfficientNetV2-Small network under the same data set and evaluation metrics, which verifies the effectiveness and superiority of the architecture optimization in pump anomaly detection.

Keywords: SimAM; CBAM; EfficientNetV2; BCEWithLogitsLos; Spectrogram

1. Introduction

In modern industrial production, pump as a key fluid delivery device, widely used in petroleum, chemical, electric power and other fields. With the extension of operation time, the pump often appears vibration, noise, abnormal temperature rise and other faults, these problems not only affect the stability of the equipment operation, but also may cause safety accidents and economic losses. Therefore, how to realize the timely and accurate identification of pump abnormal state become an important problem to be solved.

Previously, data-driven methods have shown significant advantages in the field of equipment fault diagnosis. Among them, Convolutional Neural Network (CNN),especially EfficientNet, has become a mainstream tool for industrial anomaly detection due to its excellent performance in image processing and feature modeling, with many achievement in this fleid. In China, Peng et al. (2025) proposed an intelligent pneumothorax diagnosis method based on improved EfficientNet. Zhao et al. (2025) optimized EfficientNet through transfer learning to realize skin disease classification. The Lite-EfficientNet proposed by Zhang and Luo (2024) significantly improves the efficiency of the model by combining spatial information optimization and lightweight design. In terms of international research, Ye et al. (2022) proposed an improved EfficientNetV2 model combined with visual attention mechanism, which has excellent performance in cassava disease recognition. Fused

EfficientNetV2 and Swin Transformer to realize high-precision recognition of tomato leaf diseases. Zhao et al. (2024) verified the effectiveness of EfficientNetV2 in the corrosion image classification task, demonstrating its application potential in complex image recognition tasks. What's more, Yang et al. (2025), Hu et al. (2024), Sun et al. (2024), and Dai et al. (2024) have respectively applied modified EfficientNet models to address image recognition challenges in the fields of mining, agriculture and manufacturing. These works provide a solid theoretical basis and reference for the design and implementation of pump anomaly detection system.

However, traditional CNN often has problems such as insufficient feature extraction ability and low information utilization when dealing with high-dimensional and complex industrial signal data. To alleviate this challenge, attention mechanism has been gradually introduced into the model architecture design. The Attention mechanism represented by Convolutional Block Attention Module (CBAM) can weight the input feature map in the channel and spatial dimensions, significantly improve the model's ability to focus on key areas, thereby improving the detection accuracy and model discrimination performance.

Therefore, based on the above research background, this paper proposes a EfficientNetV2-based lightweight pump anomaly detection model incorporating an improved convolutional block attention module (CBAM) mechanism.

2. Experimental Methods and Materials

2.1. Dataset

In this study, the Malfunctioning Industrial Machine Investigation and Inspection (MIMII) dataset (Purohit et al., 2019) was used as the basis for model pre-training. The dataset covers sound signals of four types of common industrial equipment (pumps, fans, valves, and slides) under normal operation and multiple abnormal states (such as rotation imbalance, leakage, contamination, etc.). The data of each type of device is divided into two categories: normal and abnormal, so as to facilitate the training of binary classification tasks. During the audio acquisition process, the research team placed a circular microphone array consisting of eight independent microphones at a distance of about 500 mm from the device to ensure high-quality capture of the sound emitted by the target device. At the same time, the background noise in the real industrial environment is introduced into the data set and fused with the audio signal of the device itself to simulate the complex and variable practical application scenarios.

In order to enhance the robustness of the model under noise interference conditions, this study selects the -6 dB background noise version with low signal-to-noise ratio as the pre-training corpus to improve the adaptability of the model in the real industrial environment. The specific data are shown in Table 1.

Table 1: Data Sources.

Name	Audio duration(s)	Class	Audio quantit
MIMII Dataset	$\leq$10	0_dB_pump	4205
MIMII Dataset	$\leq$10	-6_dB_pump	4205
MIMII Dataset	$\leq$10	6_dB_pump	4205

2.2. Model Architecture

EfficientNetV2-Small is a lightweight variant introduced by Google in their family of Efficient Convolutional neural networks, which aims to achieve better recognition performance while significantly reducing computational resource consumption. The network is automatically generated based on Neural Architecture Search (NAS) technology. In the model design process, multiple factors such as training speed, inference efficiency and model accuracy are fully weighed, which is especially suitable for image recognition tasks in resource-sensitive edge computing devices. The network structure is shown in Table 2, which is mainly composed of multi-layer Fuded-MBConv and MBConv modules. Each module is responsible for shallow and deep feature extraction tasks at different stages, and Sigmoid weighted Linear Unit (SiLU) is used as the activation function. To enhance the nonlinear modeling ability and improve the training stability, which further improves the overall performance.

Table 2: EfficientNetV2-Small architecture

Stage	Operator	Repeats	Output Channels	Expansion Ratio	Kernel Size	Stride	SE	Activation
0	Conv3×3	1	24	-	3×3	2	-	SiLU
1	Fused-MBConv	2	24	1	3×3	1	-	SiLU
2	Fused-MBConv	4	48	4	3×3	2	-	SiLU
3	Fused-MBConv	4	64	4	3×3	2	-	SiLU
4	MBConv	6	128	4	3×3	2	✓	SiLU
5	MBConv	9	160	6	3×3	1	✓	SiLU
6	MBConv	15	256	6	3×3	2	✓	SiLU
7	Conv1x1+Pool+FC	1+1+1	1280	-	-	-	-	SiLU

The network structure of EfficientNetV2-Small is based on Fuded-MBConv and MBConv modules. The former improves the efficiency of shallow feature extraction by fusing 1×1 convolutions and 3×3 convolutions. The latter retains the efficient feature extraction capabilities based on depthwise separable convolution and Squeeze-and-Excitation (SE) mechanism in the original EfficientNet series. The initial stage of the network is a 3×3 convolutional layer for basic feature extraction, followed by multiple stacked Fuded-MBConv layers, which omit the expansion and compression process and are suitable for processing low-level features. The middle and later sections use the MBConv module, which firstly performs channel expansion, then extracts spatial features through deep convolution, and finally uses 1×1 convolution for channel compression, and combines the SE attention mechanism to enhance the feature expression ability. By gradually deepening the network structure and increasing the number of channels, the model extracts more complex semantic information layer by layer. In the tail of the network, EfficientNetV2-Small uses a 1×1 convolutional layer to integrate global features, followed by a global average pooling and a fully connected layer for classification output. This structure design effectively improves the expression ability and generalization performance of the model while maintaining the compactness of the model, and achieves a high degree of unity of inference speed and accuracy.

2.2.1. THE LIGHTWEIGHT NEURAL NETWORK PROPOSED IN THIS PAPER

The lightweight neural network proposed in this study is based on EfficientNetV2-Small with several key structural improvements, especially in attention mechanism and module design. The original Attention Module is replaced with an optimized Convolutional Block Attention Module

(CBAM) module, and the Simple Attention Module (SimAM) is introduced into the spatial attention part. SimAM is a parameterless attention mechanism, which finely models spatial information by measuring the importance of each position in the feature map based on the principle of neuronal energy minimization. In the network, both Fuded-MBConv and MBConv modules adopt improved CBAM attention mechanism. The Fuded-MBConv module improves the efficiency of shallow feature extraction by combining 1×1 convolution and 3×3 convolution. The MBConv module further optimizes the feature extraction process by depthwise separable convolution and improved CBAM attention mechanism. The reasonable collocation of these modules not only enhances the feature extraction ability, but also effectively reduces the computational complexity, so that the network has strong computational efficiency while maintaining high accuracy. These innovative designs enable the network to show significant performance advantages in pump anomaly detection tasks, which can provide more accurate results while ensuring efficient computation. The specific architecture and parameters of the improved model are shown in Figure 1 and Table 3 respectively.

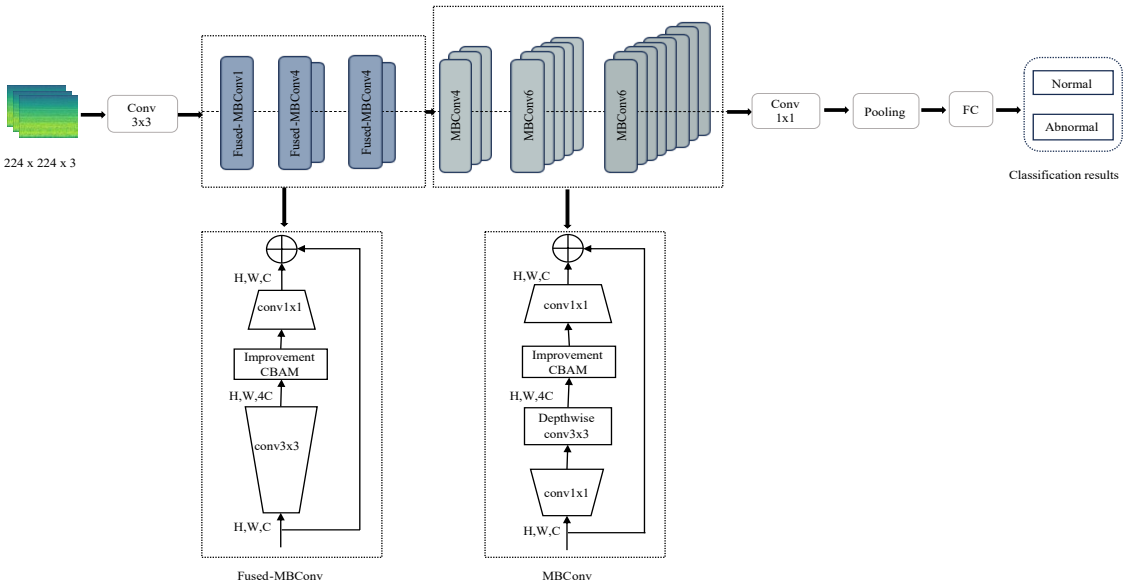

Figure 1: Improved EfficientNetV2-Small.

Table 3: Improve the EfficientNetV2-Small network structure.

Stage	Operator	Repeats	Output Channels	Expansion Ratio	Kernel Size	Stride	Improvement CBAM	Activation
0	Conv3×3	1	24	-	3×3	2	-	SiLU
1	Fused-MBConv	2	24	1	3×3	1	-	SiLU
2	Fused-MBConv	4	48	4	3×3	2	-	SiLU
3	Fused-MBConv	4	64	4	3×3	2	-	SiLU
4	MBConv	6	128	4	3×3	2	✓	SiLU
5	MBConv	9	160	6	3×3	1	✓	SiLU
6	MBConv	15	256	6	3×3	2	✓	SiLU
7	Conv1x1+Pool+FC	1+1+1	1280	-	-	-	-	SiLU

2.2.2. IMPROVED CBAM

CBAM consists of a channel attention module and a spatial attention module with its complete structure shown in Figure 2, which helps to enhance the feature extraction ability of the model. The channel attention module optimizes the feature representation by capturing the dependency between different feature channels. Firstly, the global description information of each channel is extracted by global average pooling and global Max pooling, and then the weight of each channel is calculated by using Multi-layer Perceptron (MLP) to adjust the importance of feature channels. In this way, the model can pay more attention to those channels with higher importance in the feature extraction process to improve the overall performance. The spatial attention module focuses on capturing the spatial dependencies of features. In the traditional spatial attention module, the spatial attention map is usually generated by pooling and convolution operations, but some detailed information may be lost in this process. In order to overcome this problem, Instead, we introduce a parameter-free Attention mechanism Simple Attention Module (SimAM) based on the principle of neuronal energy minimization. SimAM can effectively model the importance of spatial location by constructing neuronal activity functions without introducing additional learnable parameters and computational burden, and overcome the problems of feature suppression and insufficient receptive field caused by traditional spatial attention dependent convolutional structures. This improvement not only has better spatial modeling ability in theory, but also shows stronger feature discrimination and lower computational complexity in experiments, which provides an effective path for lightweight modeling.

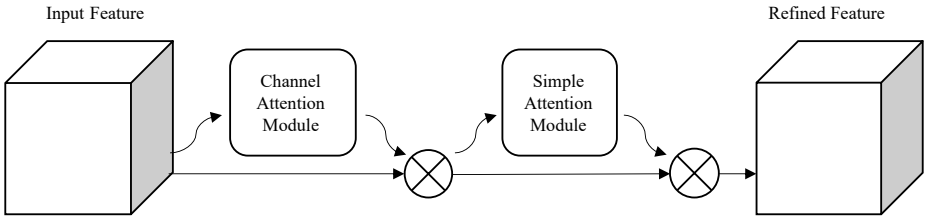

Figure 2: Diagram of the improved CBAM structure.

2.2.3. SIMAM

SimAM (Yang et al., 2021) is a lightweight, parametrically free attention mechanism. Its design is inspired by the modeling method of neuron activity in neuroscience. Different from traditional attention mechanisms that generate explicit attention maps by introducing convolution, MLP or gated units, SimAM starts from the perspective of "energy minimization" and constructs an energy function based on neuron discriminability to measure the importance of a neuron, so as to realize attention weighting.

If a neuron x_i can be isolated without affecting the overall neural activity, it is discriminative and should be given a higher attention weight. To this end, SimAM constructs an energy function as follows.

$$E\left(x_i\right)=\left(x_i-\mu_{-i}\right)^2+\lambda\cdot\sigma_{-i}^2 \tag{1}$$

Among them:

x_i denotes the value of the current neuron;

$\mu_{-\mathrm{i}}$ and $\sigma^2_{-\mathrm{i}}$ are the mean and variance of neurons except those in the current channel, respectively;

λ hyperparameter (usually set to a small constant, such as 1)

The smaller the $\mathrm{E}(\mathrm{x_i})$ energy function is, the easier the neuron is to be separated and the stronger the discrimination ability is.

To map the above energy into attention weights, SimAM uses an attention function of the following form:

$$\alpha_{\mathrm{i}} = \frac{1}{1 + \mathrm{E}\left(\mathrm{x_i}\right)} \tag{2}$$

This value is applied as an attention weight on the original feature map to achieve saliency enhancement in the spatial dimension. The design of SimAM not only avoids the problem of local information loss caused by the limitation of convolution kernel in the traditional spatial attention mechanism, but also significantly improves the feature identifiability while maintaining computational efficiency. In addition, SimAM has good model generalization and deployment flexibility because it does not introduce any additional parameters, which is especially suitable for practical application scenarios that are sensitive to resources. In conclusion, SimAM is an effective and efficient parameterless spatial attention method, which can be used as an alternative to the CBAM spatial attention module to further enhance the network's ability to model spatial information.

2.2.4. COMBINING MIXUP AND BCEWITHLOGITSLOSS

MixUp (Zhang et al., 2017) is a data augmentation technique that improves the generalization ability and robustness of the model by weighted combination of input images and labels. Its core idea is to generate new training samples by linearly interpolating the images and labels of two samples and its specific process is shown in Figure 3. The operations are as follows:

Image interpolation: For two samples $\mathrm{x_i}$ and $\mathrm{x_j}$, MixUp performs a linear interpolation on the image:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{3}$$

Where λ is a coefficient sampled from the Beta distribution, typically λ is between 0 and cc1. It controls the proportion in which the two samples are mixed.

Label interpolation: The labels are also interpolated with weights:

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{4}$$

Here, $\mathrm{y_i}$ and $\mathrm{y_j}$ are the labels of samples $\mathrm{x_i}$ and $\mathrm{x_j}$. For classification tasks, labels are usually one-hot-encoded vectors.

In Binary classification tasks, binary Cross-Entropy Loss (BCE) is one of the commonly used loss functions to measure the difference between the probability distribution output by the model and the true label. However, in many deep learning frameworks, linear activation functions (i.e., raw logits) are often used in the output layer, so an additional Sigmoid activation operation is applied to the logits values to convert them into probability values. To solve this problem, the BCEWithLogitsLoss loss function combines Sigmoid activation and binary cross-entropy loss into a single overall operation, which avoids the step of additional activation of logits, thus improving computational efficiency and enhancing numerical stability. For a single sample, the BinaryCross-EntropyLoss(BCE) is calculated as follows:

$$\mathrm{BCE}(p, y) = -[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)] \tag{5}$$

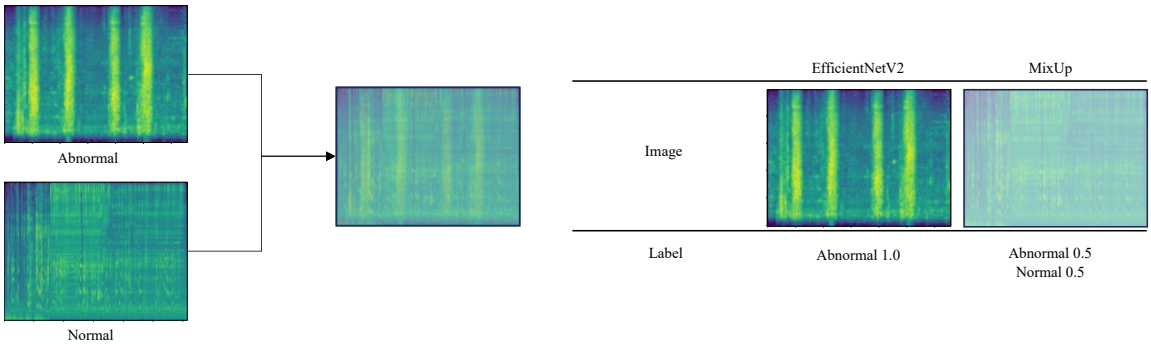

Figure 3: MixUp Implementation.

Among them:

p is the probability that the model predicts the positive class (class 1), which is usually obtained from the output of the model with a Sigmoid activation function in the range $0 \leq \mathrm{p} \leq 1$.

y is the true label of the sample and takes the value 0 or 1, where $\mathrm{y} = 1$ indicates the positive class and $\mathrm{y} = 0$ indicates the negative class.

The combination of MixUp and BCEWithLogitsLoss significantly improves the performance of the model in several ways. By linearly interpolating samples from different categories to generate virtual samples and assigning smooth labels, MixUp helps to alleviate overfitting and improve the generalization ability of the model. At the same time, BCEWithLogitsLoss can directly process the output of logits without Sigmoid activation, avoiding the problem of numerical instability and effectively optimizing the model performance in binary classification tasks. The combination of the two not only improves the diversity of data and training stability, but also improves the performance of the model when dealing with imbalanced data sets and difficult to classify samples.

3. Experiment

3.1. Experiment Environment

The experiment is conducted using the Python programming language, with an NVIDIA GeForce GTX 3080Ti GPU processor, 16GB of memory, and the Ubuntu 20.04 operating system. The deep learning framework used is PyTorch.

3.2. Parameter Settings

After several experiments, we determined the best training parameter configuration for this model. The image normalized size (Norm_size) was set to 224, the batch size (Batch_size) was 32, and the whole training process was carried out for a total of 200 epochs. When the validation set loss (Best_loss) was lower than 2, the model weights were saved to further reduce resource consumption. The cosine annealing learning rate strategy is used, the learning rate is set to 0.01 in the initial stage, and the parameter lrf is set to 0.1. During the training process, the learning rate will eventually decay to 10% of the initial value (that is, about 0.001).

3.3. Experiment Results and Analysis

In this experiment, in order to ensure the independence of the data and avoid the occurrence of a certain image in the training set and test set at the same time after expansion, thus generating false accuracy, each spectrogram was randomly divided into training set with 1890 images, validation set with 540 images, and test set with 270 images according to 7:2:1. On the test set, the accuracy is about 96% for both normal and abnormal pump sound spectrogram images. Compared with the unimproved EfficientNetV2 model and other deep learning network models such as MobileVit, CNN, ResNet50, Conformer, etc., the comparison data is shown in Table 4.

Table 4: Performance metrics of different models on the MIMII dataset.

Model	Input Features	Accuracy
Improved EfficientNetV2-Small	Spectrogram	0.9603±0.25
EfficientNetV2-Small	Spectrogram	0.9527±0.25
EfficientNetV0	Spectrogram	0.8613
EfficientNetV1	Spectrogram	0.9454
EfficientNetV2	Spectrogram	0.9286
EfficientNetV3	Spectrogram	0.9496
EfficientNetV4	Spectrogram	0.9328
EfficientNetV5	Spectrogram	0.9286
EfficientNetV6	Spectrogram	0.9328
EfficientNetV7	Spectrogram	0.9486
MobileVit	Spectrogram	0.9316

During the model training process, it can be observed that the improved CBAM attention mechanism in this model achieves higher accuracy compared to other attention mechanisms, such as SOCA, SimAm, ECA, and NAM. The comparison data is shown in Table 5.

Table 5: Performance metrics of different attention mechanisms on the dataset.

Attention	Input Features	Accuracy	Loss
Improved CBAM	Spectrogram	0.9603±0.25	0.0624
ECA	Spectrogram	0.9501	0.3228
SCOA	Spectrogram	0.9542	0.0617
SimAm	Spectrogram	0.9482	0.0323
NAM	Spectrogram	0.9402	0.2236

This study compares four audio feature extraction methods: Spectrogram, Melspectrogram, Log-Frequency-Power-Spectrogram, and Linear-Frequency-Power-Spectrogram. As shown in Figure 4, after experimentation, it was found that the Spectrogram produced higher accuracy for the model, as presented in Table 6.

4. Conclusion

This study proposes a pump anomaly detection method that integrates the improved CBAM attention mechanism and EfficientNetV2, and combines the MixUp data enhancement strategy and the

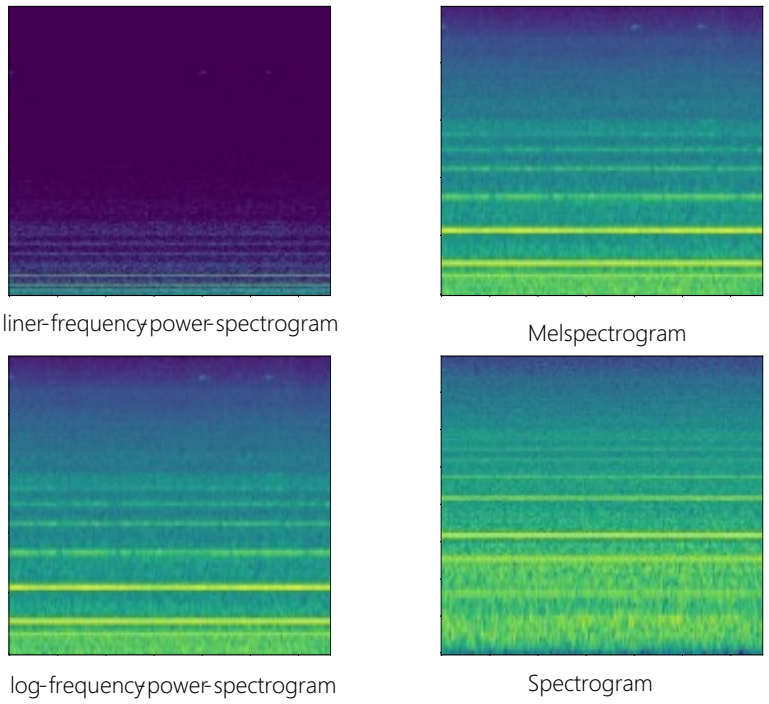

Figure 4: Spectrogram.

Table 6: Feature Comparison

Model	Input Features	Accuracy	Loss
Improved EfficientNetV2-Small	Spectrogram	0.9603±0.25	0.0624
Improved EfficientNetV2-Small	Melspectrogram	0.9520±0.25	0.1152
Improved EfficientNetV2-Small	Log-Frequency-Power-spectrogram	0.9378±0.15	0.0595
Improved EfficientNetV2-Small	Liner-Frequency-Power-spectrogram	0.9248±0.15	0.1631

BCEWithLogitsLoss loss function to improve the detection accuracy and robustness of the model under complex operating conditions. In terms of network structure, we introduce an improved CBAM module based on the EfficientNetV2-Small backbone network, and retain its original Channel Attention structure to model the dependency between feature channels. At the same time, the traditional spatial Attention mechanism is replaced by Simple Attention Module (SimAM) based on neuron separation. By constructing a parameterless energy function, the module effectively measures the importance of neurons at each position in the feature map, so as to realize efficient and non-redundant spatial attention modeling, which significantly improves the expression ability of features in the spatial dimension. The experimental results show that the proposed method achieves excellent performance in the pump anomaly detection task on the MIMII dataset. It is not only significantly better than the traditional method in accuracy, but also shows stronger stability and generalization ability in a variety of noise environments, which provides an efficient and practical intelligent diagnosis scheme for industrial equipment condition monitoring.

References

Yingyu Dai, Jingchao Li, Ying Zhao, Yanli Liu, Shenhua Wang, and Bin Zhang. A lightweight rolling bearing fault diagnosis method based on improved efficientnet model. *Manufacturing Technology & Machine Tool*, (9):9–15, 2024. doi: 10.19287/j.mtmt.1005-2402.2024.09.001.

Shiwei Hu, Jianxin Deng, Haoyu Wang, and Lin Qiu. Grape leaf disease identification method based on improved efficientnetb0 model. *Modern Electronics Technique*, 47:73–80, 2024. doi: 10.16652/j.issn.1004-373x.2024.15.012.

Sheng Peng, Zhigao Zeng, Shengqiu Yi, and Xinpan Yuan. A method of pneumothorax diagnosis based on improved efficientnet network. *China Science and Technology Information*, pages 88–90, 2025. doi: 10.3969/j.issn.1001-8972.2025.02.026.

Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *CoRR*, abs/1909.09347, 2019.

Yubing Sun, Lixin Ning, Bin Zhao, and Jun Yan. Tomato leaf disease classification by combining efficientnetv2 and a swin transformer. *Applied Sciences (Switzerland)*, 14, 2024. doi: 10.3390/app14177472.

Hailong Yang, Yiping Yuan, Panpan Fan, and et al. Foreign object recognition for mine conveyor belt iron separators based on transfer learning with efficientnet. *Coal Science and Technology*, 2025. doi: 10.12438/cst.2024-0772.

Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11863–11874. PMLR, 2021.

Yuanbo Ye, Houkui Zhou, Huimin Yu, Haoji Hu, Guangqun Zhang, Junguo Hu, and Tao He. An improved efficientnetv2 model based on visual attention mechanism: Application to identification of cassava disease. *Computational Intelligence and Neuroscience*, 2022. doi: 10.1155/2022/1569911.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017.

Shufang Zhang and Xizhe Luo. Lite-efficientnet: spatial information optimization and lightweight network integration. *Journal of Optoelectronics·Laser*, 2024. doi: 10.16136/j.joel.2024.11.002.

Haiyan Zhao, YouTeng Wu, and Menghan Ren. Research on skin lesion classification using the improved efficientnet network based on transfer learning. *Journal of Inner Mongolia University for Nationalities (Natural Science Edition)*, 40:22–27, 2025. doi: 10.14045/j.cnki.15-1220.2025.01.004.

Ziheng Zhao, Elmi Bin Abu Bakar, Norizham Bin Abdul Razak, and Mohammad Nishat Akhtar. Corrosion image classification method based on efficientnetv2. *SSRN*, 2024. doi: 10.1016/j.heliyon.2024.e36754.