Int. J. Metrol. Qual. Eng. **13**, 13 (2022)

https://doi.org/10.1051/ijmqe/2022011

International Journal of Metrology and Quality Engineering

Available online at:
www.metrology-journal.org

RESEARCH ARTICLE OPEN ACCESS

Research on the enhancement of machine fault evaluation model based on data-driven

Peng Cui[1], Xuan Luo[2,*], Xiaobang Li[2], and Xinyu Luo[2]

[1] School of Information Science and Engineer, Yanshan University, Qinhuangdao 066004, China
[2] Industrial Technology Center, Hebei Petroleum University of Technology, Chengde 067000, China

Received: 24 December 2021 / Accepted: 14 September 2022

Abstract. Recently fault data diagnosis-based deep learning methods have achieved promising results. However, most of these methods' performances are difficult to improve once they have achieved accuracy. This paper mainly uses fusion theory based on data-driven to solve this problem. Firstly, the diagnostic models are divided into feature extraction and neural network. Then, four feature extraction methods are fused by pre-allocation. The neural network part consists of three single models, and the weight of the three output results is determined by regression analysis. Experiments show that the accuracy of diagnostic models is improved. Finally, we combine the two studies and propose a Fusion-Ensemble superposition (FES) model. The AUC value of the model is higher than 98% in most tasks of the DCASE2020 machine failure dataset.

Keywords: Feature extraction / convolutional neural network (CNN) / out-of-distribution (OOD) / multi-model fusion / model ensemble

1 Introduction

Abnormal sound can be used as an essential standard to identify whether the machine is abnormal. Normal sounds of a working machine are often smooth and regular but accompanied by obviously anomalous sounds when the machine is out of order. Anomalous sounds [1,2] indicate that a machine may have malfunctioned, including the rupture of mechanical components, stuck, or the failure of completing a specific function [3]. Timely discovery of faults can avoid heavy losses and reduce production costs. Most machine failures occur slowly, and uncertainty makes it difficult to predict, so data collection is extremely difficult. Out-of-distribution (OOD) [4] detection has methods suitable for supervised data and semi-supervised data. Therefore, the OOD detection method based on deep learning is often used for anomalous sound recognition. Now many researchers pay more attention to model innovation, but we find that feature extraction also impacts the overall recognition effect. This paper will show the impact on machine fault recognition from two aspects: feature extraction and network structure.

Aiming at handling abnormal sound detection problems in the early stages [5], Koizumi et al. [6] proposed using the Gaussian mixture model to calculate anomaly scores [7], and Foggia et al. [5] used audio streams to perform sound detection to determine dangerous situations. However, traditional algorithms cannot handle high-dimensional data, and feature extraction capabilities are weak. Deep anomaly detection (DAD) advocates for solving this problem, and auto-encoder (AE) is one of the commonly used DAD algorithms. Long- and short-term memory network adversarial networks (GANs) [8] and OC-NN [9] have also been widely used in various sound detection scenarios. Suedusa et al. [10] used an interpolation-based deep learning network for abnormal sound detection. The spectrogram of the removed center frame is used as the input of the model, and the interpolation prediction result of the removed frame is used as the output of the model. Komatsu et al. [11] proposed to use WaveNet and I-Vector to detect abnormal acoustic events based on time, location, and changes in the surrounding environment.

The main contributions of the paper are:

- In the task of audio recognition, the results of different feature extraction methods are also different. The four feature extraction methods are fused to improve the accuracy of machine fault diagnosis.
- The maximum limit of single model accuracy is broken through the method of a model ensemble.
- A method for machine fault diagnosis based on multi-feature fusion and model ensembles is proposed.

The rest of this paper is organized as follows: Section 2 introduces the data set and evaluation method. Section 3 presents the model structure and method. Section 4 shows the model accuracy and comparative test results.

* Corresponding author: cptt0000@163.com

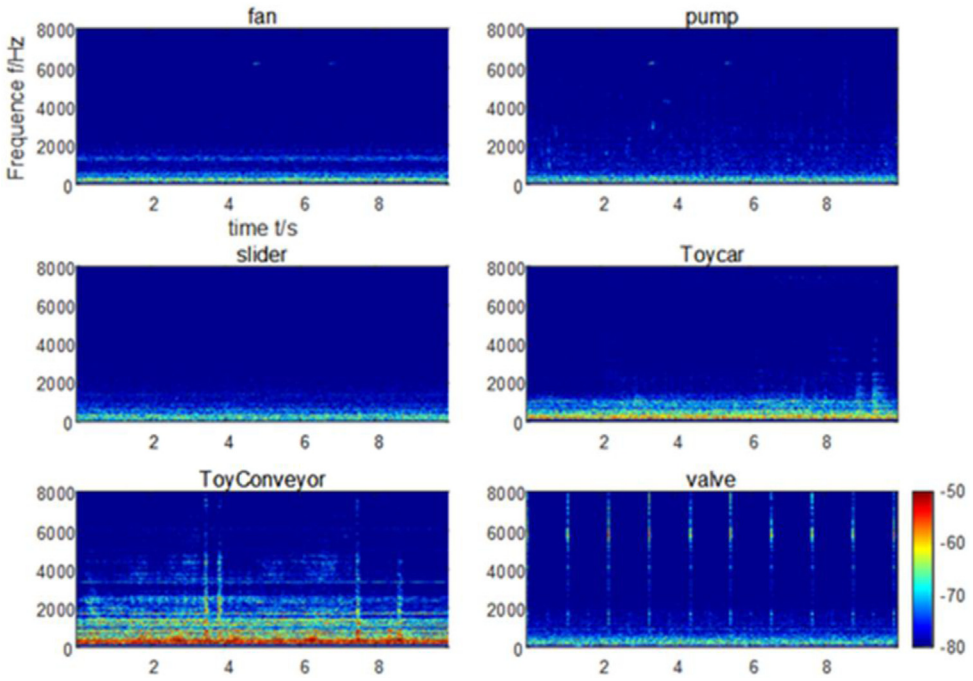

Fig. 1. Various data Log-Mel spectrogram. The horizontal axis represents time, and the vertical axis represents frequency.

2 Dataset and evaluation metrics

DCASE2020 TASK2 data set was used to verify the performance of the proposed model. The main challenge of this task is to detect unknown anomalous sounds under the condition that only normal sound samples have been provided as training data. In real-world factories, actual anomalous sounds rarely occur and are highly diverse. Therefore, exhaustive patterns of anomalous sounds are impossible to deliberately make and/or collect. This means we have to detect unknown anomalous sounds that were not observed in the given training data. The data set was composed of ToyADMOS and MIMII, which were single-channel recordings. The down-sampling rate of all audio clips was 16 kHz, and the length was about 10 s. The normal sound sample data used in the TASK2 are divided into six categories: toy-car, toy, valve, pump, fan, and slider. The first two are from toy machines, whereas the rest are from real machines [12]. Figure 1 shows the spectrograms of the six groups of samples.

There are several fixed detection indicators for OOD detection: true positive rate (TPR) is calculated in equation (1), where TP and FN represent true positive and false negative, respectively

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{1}$$

False positive rate (FPR) is calculated in equation (2), where FP and TN indicate false positive and true negative, respectively

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{2}$$

OOD detection usually uses the area under the curve (AUC) and partial-AUC (pAUC) to evaluate the quality of the model. AUC is the area under the receiver operating characteristic (ROC) curve. The abscissa of the ROC curve is FPR, and the ordinate is TPR. AUC is the indicator for judging the pros and cons of a two-class prediction model. AUC is more commonly used than accuracy and recall rate [13]. AUC can demonstrate the overall performance of the model. A high AUC value indicates that the model's performance is excellent, and the error probability of the positive prediction example is low. The pAUC is the AUC within a specific false positive rate range.

In the feature extraction, the sampling rate is 16 kHz, the window length is 1024 samples, the skip length is 512 samples (64 ms). 1024 FFT (fast Fourier transform) points are used, and 128 mel filters are used. The sequence length of a training sample is $n_{\mathrm{frames}} = 640$ audio frames, and every five frames are connected.

3 Materials and methods

3.1 Multi-feature fusion

Four feature extraction methods are applied in the feature extraction part: log-linear [14], Log-Mel [15], HPSS_h [16], HPSS_p [17]. The principle of Log-Mel is to extract sound features by simulating the human ear structure. However, when the sound signal has high and low tones, the high tones will be covered by the low tones. HPSS (harmonic/percussive source separation) technology was first applied in the music field. Music signals are distributed in two forms, continuously and smoothly, along with time and frequency. These two distributed music sources are called harmonic sources (HPSS_h) and shock sources (HPSS_p) [16].

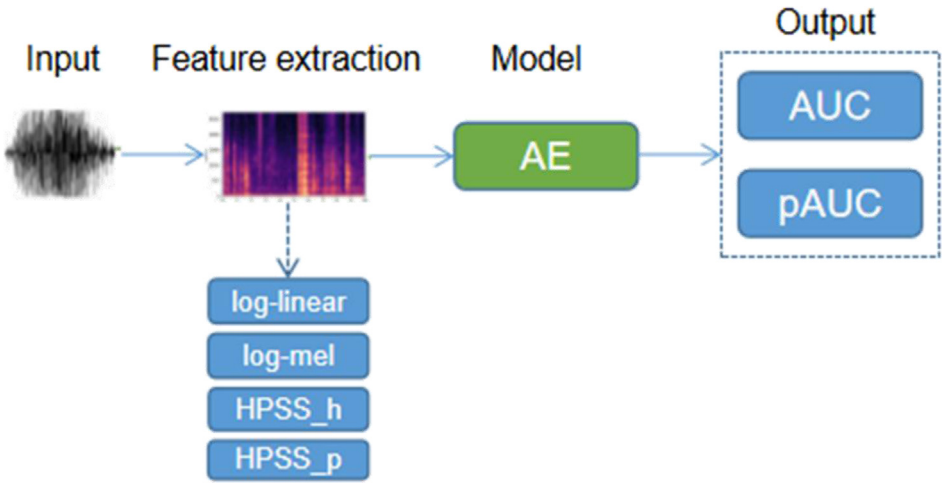

Fig. 2. The input data selects different feature extraction methods according to the corresponding relationship in Table 1, and the features are trained through AE network. During the test, the root mean square error of the reconstruction error is calculated, and the current state is obtained by comparing the threshold value.

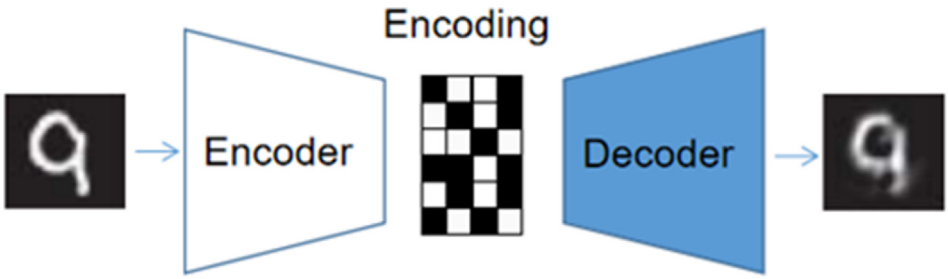

Fig. 3. AE is a special neural network architecture, and the input and output are the same architecture. It is trained in an unsupervised method to obtain the lower dimensional expression of the input data. These low-latitude information expressions are reconstructed back to high-dimensional data expressions.

3.1.1 Log-Mel

Log-Mel adopts the signal energy as the basic feature, and its signal processing can be employed as the output feature. This feature is not affected by the nature of the signal and has no restriction on the input signal, which has a better recognition effect when the signal-to-noise ratio is low.

3.1.2 HPSS_h

The harmonic source contains a fixed tone, which can form a series of smooth instantaneous envelopes on the frequency. It is smooth and continuous on the time axis and discontinuous on the frequency axis.

3.1.3 HPSS_p

The shock source is concentrated quickly, forming a series of vertical broadband spectral envelopes on the frequency spectrum, so it is discontinuous on the time axis and smoothly continuous on the frequency axis.

Figure 2 is the structure diagram in which a machine fault recognition model is established by an AE [17] (the structure is shown in Fig. 3). Different feature extraction methods have different application ranges. Log-Mel is the most widely used, so it is the basic feature extraction method. Log-linear is suitable for data with strong correlation, HPSS_p has a good extraction effect for sound with a complete period, HPSS_h is more inclined to extract the features of discontinuous sound and burst. We present a fusion strategy to allocate the best feature extraction method according to the audio characteristics of the tasks (the scheme is shown in Tab. 1).

Table 1. Tasks and feature extraction methods pre-allocation table.

Task	Feature
ToyCar ToyConveyor Fan	Log-Mel
Pump	HPSS-p
Slider Valve	HPSS-h

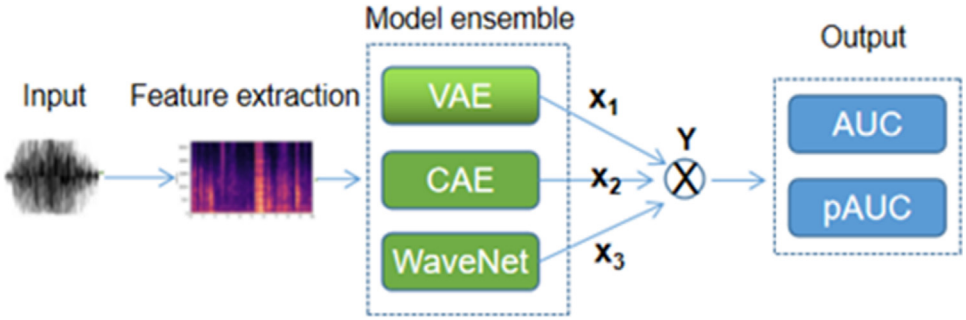

Fig. 4. Log-Mel is used for data feature extraction to train VAE, CAE and WaveNet network respectively. The training set is used to collect the output results, and the voting weights of the three networks are determined by regression analysis.

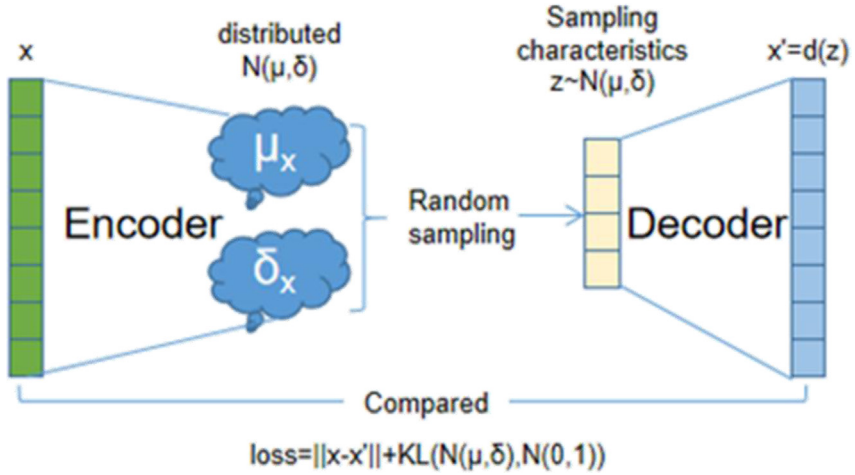

Fig. 5. Variational autoencoder structure diagram. Where x is the input data, x' is the reconstructed data, and μ_x and δ_x are the mean and standard deviation of normal distribution respectively.

3.2 Model ensemble

The purpose of a multi-model ensemble is to integrate the advantages of each model through scientific methods and obtain a stronger ability to solve unknown problems [18]. So, multi-model ensemble has attracted more attention in practical application.

The model ensemble structure is proposed as shown in Figure 4, which is composed of three models: variational autoencoder (VAE) [19], contractive autoencoder (CAE) [20], and WaveNet [21]. Log-Mel is used for feature extraction.

AE could compress input data into a lower dimension manner and decode the data into the original input data unsupervised. The encoded data is reconstructed by decoding, and the difference between the reconstructed data and the original input data is the reconstruction error. If the reconstruction error is large, it is guaranteed to be a poor auto-encoder. The unsupervised training algorithm layer by layer is used to complete the pre-training of the hidden layer between the encoders and decoders. Then the backpropagation algorithm is used to optimize and adjust the system parameters of the whole neural network, which improves the learning ability and is beneficial to the pre-training. VAE and CAE are both variants of AE.

The structure of VAE is shown in Figure 5. It contains two encoders used to calculate the mean and variance, respectively. Gaussian noise is added to the encoder network for calculating the mean value so that the decoder can be robust to noise. KL loss is applied to make the mean value 0 and the variance 1 and append a regularizer to the encoder so that the encoder data has zero mean value. The function of the network for calculating variance is to adjust the intensity of noise dynamically.

CAE replaces the Hessian matrix of AE with the Jacobian matrix [22], and other parts are almost the same.

WaveNet model is a sequence generation model that can directly learn the mapping of sampling value sequence, so it has a good synthesis effect. At present, WaveNet is applied in speech synthesis, acoustic model modeling, and vocoder and has great potential in speech synthesis. The structure is shown in Figure 6.

X_1, X_2, X_3 represents the output of the three network models, and the final result of the model ensemble is marked Y. The relationship between them is calculated according to equation (3):

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3. \tag{3}$$

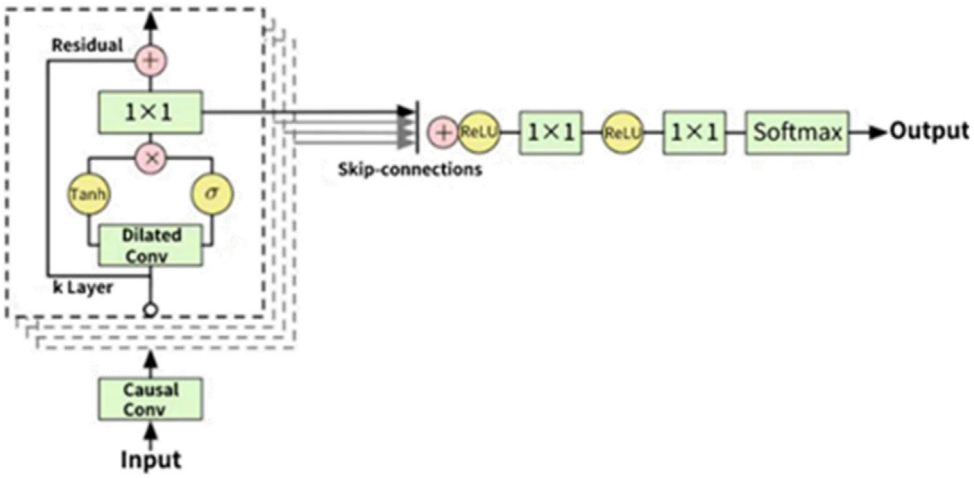

Fig. 6. Overview of the WaveNet entire architecture. A residual module in the model is shown in the dotted line. Multiple such modules will be stacked together in the network. K is the layer index. The nodes of each layer in the hidden layer will add the original value and the value of the activation function and pass it to the next layer. The 1×1 convolution kernel is used to reduce the number of channels. Then the results of the over activated function of each hidden layer are added to do a series of operations and transmitted to the output layer. The output layer uses softmax to calculate the probability of each sampling point.

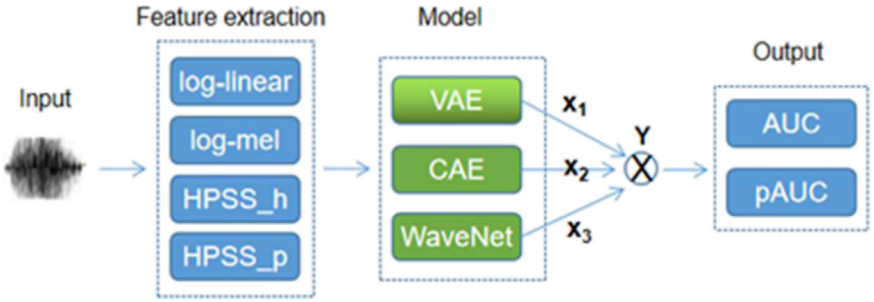

Fig. 7. FES model structure diagram.

The linear regression analysis obtains the weights a_1, a_2, a_3 of the three network models. Finally, by the weight of the training results, the network model parameters are optimized, which could get the optimal network architecture of the machine fault diagnosis.

3.3 Fusion-ensemble superposition model

We propose a fusion-ensemble superposition (FES) model based on multi-feature fusion and model ensemble (the model structure is shown in Fig. 7). It is divided into a feature extraction module and a neural network module. The feature extraction module adopts the multi-feature fusion method in Section 3.1, and the neural network module uses the model ensemble method in Section 3.2.

In order to verify the effect of FES, three models are selected for comparison: ResNet [23], MobileFaceNet [24], and PLG. PLG is a model ensemble composed of principal component analysis (PCA) [25], local outlier factor (LOF) [26], and Gaussian mixture model (GMM) [27].

ResNet solves the degradation problem through the residual learning depth network, which can train a deeper network (the structure is shown in Tab. 2). The convergence speed of ResNet is faster, so it is much easier to directly learn the residual than to learn the mapping between input and output directly, and the classification accuracy can be improved by adding layers.

MobileFaceNet has made five improvements based on MobileNetV2: separable convolution instead of average pool layer, Insightface loss function for training, reduces channel expansion multiple, PReLU instead of ReLU, and employs batch normalization. The structure is shown in Table 3. Both PReLU and ReLU are activation functions. PReLU can retain some information less than zero, and achieve the purpose of activating functions at the same time. See their expressions for specific differences:

$$\mathrm{ReLU}(x) = \begin{cases} x, if\, x > 0 \\ 0, if\, x \leq 0 \end{cases},$$

Table 2. ResNet18 network structure diagram.

Layer name	Output size	18-layer
Conv 1	112 × 112	7 × 7, 64, stride 2 3 × 3 max pool, stride 2
Conv 2_x	56 × 56	$\begin{bmatrix} 3 \times 3 & 64 \\ 3 \times 3 & 64 \end{bmatrix} \times 2$
Conv 3_x	28 × 28	$\begin{bmatrix} 3 \times 3 & 128 \\ 3 \times 3 & 128 \end{bmatrix} \times 2$
Conv 4_x	14 × 14	$\begin{bmatrix} 3 \times 3 & 256 \\ 3 \times 3 & 256 \end{bmatrix} \times 2$
Conv 5_x	7 × 7	$\begin{bmatrix} 3 \times 3 & 512 \\ 3 \times 3 & 512 \end{bmatrix} \times 2$
	1 × 1	Average pool, 1000-d fc, softmax
FLOPs		1.8×10^9

Table 3. MobileFaceNet network structure diagram.

Operator	t	c	n	s
Conv 2D	–	16	1	2
Bottleneck	1	8	1	1
Bottleneck	6	16	2	2
Bottleneck	6	16	3	2
Bottleneck	6	32	4	2
Bottleneck	6	48	3	1
Bottleneck	6	80	3	2
Bottleneck	6	160	1	1
Conv 2D	–	1280	1	1
Avg. pool	–	1280	1	–
Dense	–	Num classes	1	–

Table 4. Using the same convolutional neural network and different feature extraction methods, AUC/pAUC value summary.

Algorithm	AE	AE	AE	AE	AE
Feature	Log-Mel	Log-linear	HPSS-p	HPSS-h	Fusion
ToyCar	0.768/0.662	0.717/0.659	0.658/0.531	0.606/0.553	0.768/0.662
ToyConveyor	0.733/0.609	0.679/0.565	0.728/0.609	0.597/0.601	0.733/0.609
Fan	0.644/0.536	0.593/0.551	0.635/0.537	0.523/0.531	0.644/0.536
Pump	0.732/0.612	0.684/0.605	0.809/0.641	0.726/0.598	0.809/0.641
Slider	0.851/0.664	0.911/0.756	0.828/0.617	0.928/0.791	0.928/0.791
Valve	0.660/0.510	0.737/0.544	0.562/0.509	0.843/0.665	0.843/0.665

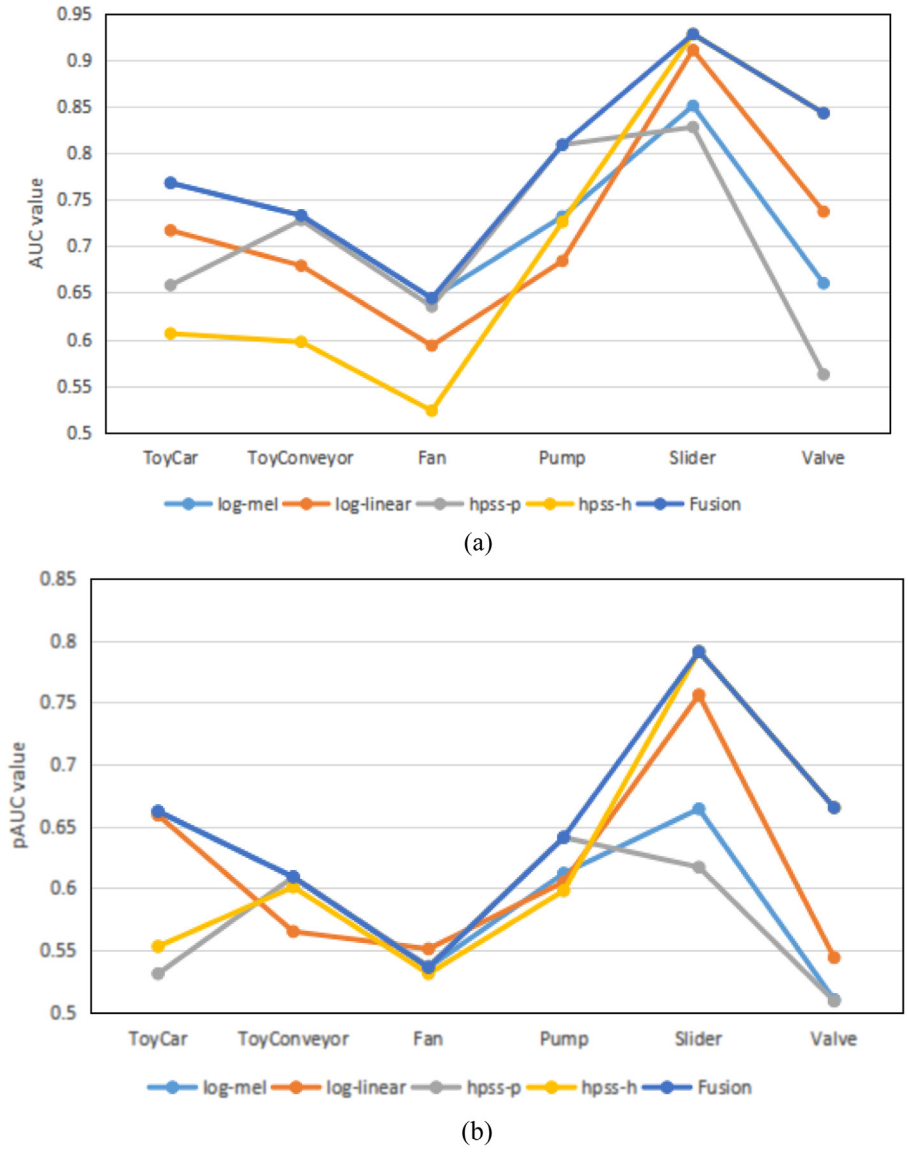

Fig. 8. Comparison line chart of each model: (a) line chart of AUC value; (b) line chart of pAUC value.

$$\mathrm{PReLU}(x) = \begin{cases} x_i, if\, x_i > 0 \\ a_i x_i, if\, x_i \leq 0 \end{cases},$$

where a_i is automatically calculated by the network feedback, and i represents different channels.

PCA is a statistical method that converts a set of potentially correlated variables into a set of linearly unrelated variables through an orthogonal transformation. It is widely used in many fields such as satisfaction measurement, pattern recognition, image compression, etc. LOF mainly determines whether the point is an outlier by comparing the densities of each point and its neighbors. Points with low densities are identified as outliers. GMM uses the Gaussian probability density function to quantify the data accurately. It is a model that decomposes the data into multiple normal distribution curves. The ensemble strategy of PLG is to convert the outlier scores into a standardized scale and then calculate the average standardized values for the three models.

3.4 Experiment and discussion

3.4.1 Effect comparison of multi-feature fusion

Table 4 lists the recognition effects of four different feature extraction methods using the same neural network, and the results of the multi-feature fusion model are also listed. The experimental results agree with the pre-allocation in Table 1, proving that each task's applicable characteristic hypothesis is valid.

In the feature extraction based on HPSS_p, the identification accuracy of the Pump is 80.9%/64.1%, which is better than other feature extraction schemes. In slider and value, the accuracy of HPSS_h is 7.7%/12.7%

Table 5. Using the same feature extraction methods and different convolutional neural networks, AUC/pAUC value summary. All values are in %.

Feature	Log-Mel	Log-Mel	Log-Mel	Log-Mel
Algorithm	VAE	CAE	WaveNet	Ensemble
ToyCar	93.39/85.46	91.25/87.36	96.73/89.30	98.30/93.55
ToyConveyor	83.46/68.98	72.23/60.23	87.22/72.59	89.02/73.89
Fan	80.94/66.55	81.82/76.98	93.51/85.10	94.12/88.23
Pump	85.26/74.35	88.17/80.36	95.87/89.53	97.31/92.56
Slider	95.64/90.74	86.49/74.65	97.36/94.60	97.85/94.54
Valve	91.54/77.10	84.59/62.41	97.94/91.58	98.35/92.11

and 18.3%/15.5% higher than Log-Mel, respectively. In the Pump task, the accuracy of HPSS_p is 6.7%/2.9% higher than Log-Mel. The experimental results show that Log-Mel performs well in signal processing: larger background sound, less obvious sound characteristics, or signal processing with lower SNR. Other feature extraction methods perform better than Log-Mel in some tasks. HPSS-h can extract more features for periodic smooth sound. Correspondingly, HPSS-p has a good performance in extracting the features of irregular sounds. None of log-linear is the best, but it is the most comprehensive and can achieve satisfactory results in different situations. In Figure 8, it is obvious that our model performance is significantly improved over the single feature model. Therefore, it is proved that the multi-feature fusion method is better than the single feature method in machine fault diagnosis.

3.4.2 Effect comparison of model ensemble

Table 5 lists the recognition effects of a single network and a Multi-model ensemble network.

WaveNet can generate the deep neural network of the original audio waveform, which is specially designed for audio. The experimental results show that among the three single model networks used in the model ensemble network, WaveNet has achieved the best results in all projects. As described in Section 3.2, although CAE simply replaces the Hessian matrix in AE, it performs slightly better in periodic stable sound. In Fan and Pump projects, CAE is 0.88% and 2.91% higher than the AUC of VAE, and pAUC is 10.43% and 6.01% higher than VAE. A large number of improvements made by VAE relative to AE. The experimental results also confirm that the VAE improvement is successful. The AUC and pAUC of VAE are 11.23% and 16.09% higher than AUC at most other projects.

It can be seen from Figure 9 that the ensemble network composed of three single models is better than any of them. WaveNet is the best fault diagnosis among the three single models. However, the average AUC of the ensemble network is 1.05% higher than WaveNet, and the average pAUC is 2.03% higher than WaveNet. Therefore, it is proved that the model ensemble method can improve the effectiveness of the single model method in mechanical fault diagnosis.

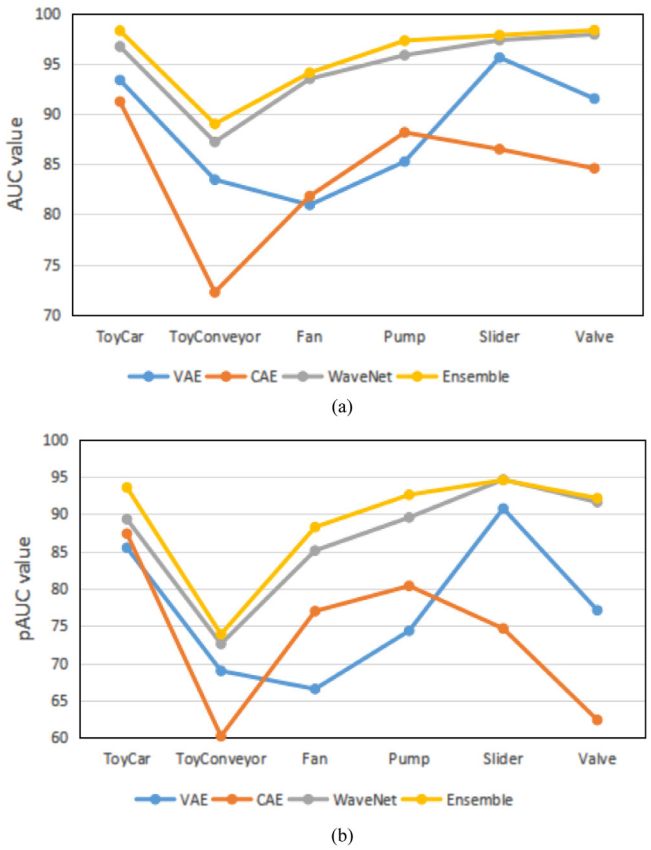

Fig. 9. Comparison line chart of each model: (a) line chart of AUC value; (b) line chart of pAUC value.

3.4.3 Effect comparison of Fusion-Ensemble superposition model

Table 6 lists the recognition effects of FES network and three comparison networks. ResNet, MobileFaceNet and PLG have different effects on different projects. From the overall effect, the fusion model is better than the single model. Although PLG only achieved the best results in the valve project, the average AUC was 1.46% and 2.62% higher than ResNet and MobileFaceNet, respectively. The average effect of the two single models has little difference,

Table 6. The recognition effect is significantly different using the same convolutional neural network and different feature extraction methods. All values are in %.

Algorithm	ResNet	PLG	Mobile FaceNet	FES
ToyCar	92.55/81.47	90.12/76.19	85.92/80.59	98.30/93.55
ToyConveyor	88.68/71.20	85.62/65.02	92.17/80.88	89.02/73.89
Fan	92.28/86.74	92.00/74.62	94.58/92.84	94.12/88.23
Pump	91.15/86.65	89.24/76.74	90.05/76.55	98.11/93.76
Slider	85.06/70.21	93.42/86.61	93.56/88.51	98.24/94.97
Valve	87.28/79.35	95.34/75.72	73.77/66.26	98.95/92.43
Average	89.50/79.27	90.96/75.82	88.34/80.94	96.12/89.47

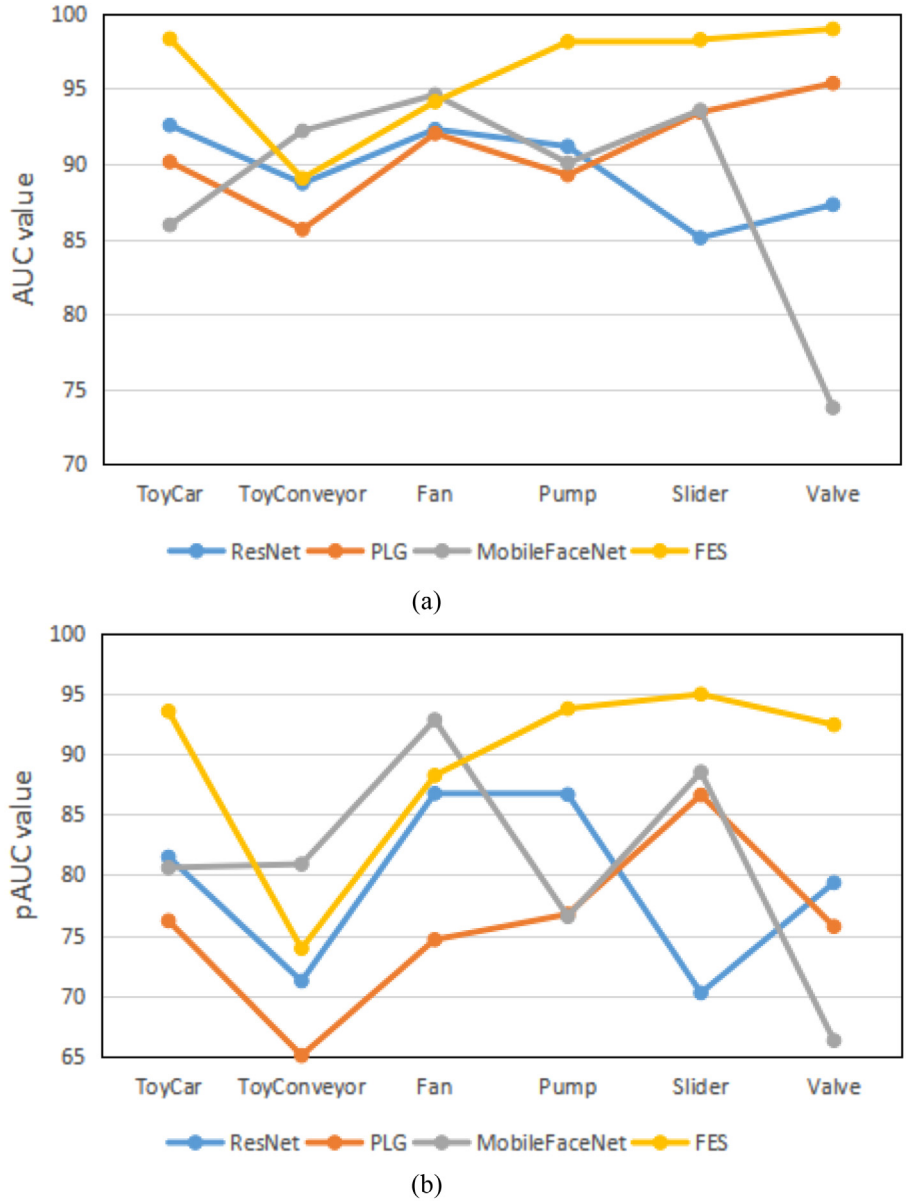

Fig. 10. Comparison line chart of each model: (a) line chart of AUC value; (b) line chart of pAUC value.

Table 7. Method generality experiment results. Average precision and accuracy are reported with 68% confidence intervals.

Algorithm	Average precision	Accuracy
DenseNet-201	87.1%	80.3%
FES	88.3%	81.1%

and each has its own advantages. ResNet works best in ToyCar and Pump projects, MobileFaceNet achieved the best results in the ToyConveyor, Fan and Slider projects, especially in the ToyConveyor project.

The average AUC of FES was 6.62%, 5.16%, and 7.78% higher than ResNet, MobileFaceNet, and PLG, respectively. Moreover, the increase of the minimum value is more than 10% for most projects. It can be proved that our FES model can complete the task of machine fault diagnosis. The specific effect is shown in Figure 10.

3.5 Method generality experiment

In order to test the effect of our proposed method on other data sets, we use FSDnoisy18k dataset [28], the comparison model DenseNet-201 is a Densenet [29] with 201 layers, and use ImageNet [30] for pre-training. Although ImageNet is an image data set, we found that it is very effective for pre-training. The results are shown in Table 7.

FSDnoisy18k dataset is a multi-classification task and contains 18532 audio clips across 20 classes, totaling 42.5 h of audio. The clip durations range from 300 ms to 30 s. DenseNet-201 has used ImageNet for pre training, and then used the target data set for training. It is not only a deep learning method, but also represents a data fusion method. The results show that the proposed method is also suitable for multi classification problems, and is 1.2% and 0.8% higher than the data fusion method in average precision and accuracy, respectively.

4 Conclusion

In the field of mechanical fault diagnosis, vibration signals are often used as data sources, but there are three problems: (1) in order to ensure the diagnosis effect, vibration sensors need to be deployed in each monitoring position, and the equipment cost will be very high; (2) some compact devices do not have space to install vibration sensors; (3) the stability of heavy equipment is good, and the vibration effect is not obvious, resulting in inaccurate diagnosis. We propose a fault diagnosis method based on sound characteristics, which can solve these problems. Compared with the traditional method, our method has lower cost and wider applicability.

In this paper, we discuss the influence of multi-feature fusion and model ensemble on the effect of machine fault diagnosis. The experiment shows that each feature extraction method is suitable for different machine fault types. Selecting an appropriate feature extraction method plays an important role in improving the accuracy of machine fault diagnosis. For tasks with multiple data types, we propose a feature fusion method to configure appropriate feature extraction methods for each type. It is helpful to improve the overall performance of the task. The multi-model ensemble can fuse excellent models through some scientific methods to break through the bottleneck of the generalization ability about a single model to unknown problems and integrate the advantages of each model to obtain the optimal solution to the same problem. An ensemble model is proposed with the average AUC value reaching 95.83%, and it is higher than a single model network. The superiority of the ensemble network is further proved. A new mechanical fault diagnosis model FES is proposed by combining the two experiments. The results show that the AUC value of the model in most projects is more than 98% and has good fault identification accuracy.

References

1. T. Hayashi. T. Komatsu. R Kondo et al., Anomalous sound event detection based on WaveNet, in: European Signal Processing Conference (EUSIPCO), 2018
2. D. Chakrabarty, M. Elhilali, Abnormal sound event detection using temporal trajectories mixtures, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016
3. Y. Li, X. Li, Y. Zhang et al., Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads. IEEE Access 1–1 (2018)
4. D. Hendrycks, K.A. Gimple, A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv:1610.02136 (2017)
5. P. Foggia, N Petkov, A Saggese et al., Audio surveillance of roads: a system for detecting anomalous sounds, IEEE Trans. Intell. Transport. Syst. **17**, 279–288 (2015)
6. Y. Koizumi, S. Saito, H. Uematsu et al., Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson Lemma, in: European Signal Processing Conference (EUSIPCO), 2017, pp. 698–702
7. D. Putri, D.O. Siahaan, Software feature extraction using infrequent feature extraction, in: 6th International Annual Engineering Seminar (InAES), 2016
8. V.K. Mittal, B. Yegnanarayana, Production features for detection of shouted speech, in: Consumer Communications and Networking Conference (CCNC), 2013, pp. 106–111
9. S. Advanne, P. Pertila, T. Virtanen et al., Sound event detection using spatial features and convolutional recurrent neural network, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017
10. K. Suefusa, T. Nishida, H. Purohit et al., Anomalous sound detection based on interpolation deep neural network, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 271–275

11. T Komatsu, T. Hayashiy, R. Kondo et al., Scene-dependent anomalous acoustic-event detection based on conditional Wavenet and I-vector, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 870–874
12. K. Yuma, S. Shoichiro, U. Hisashi, H. Noboru, I . Keisuke et al., ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection, in: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019, pp. 308–312
13. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 1145–1159 (1997)
14. P. Vellaisamy, V. Vijay, Log-linear modeling using conditional log-linear structures, Ann. Inst. Statist. Math. **61**, 309–329 (2009)
15. Y.-K. Lee, O.-W. Kwon, A phase-dependent a priori SNR estimator in the Log-Mel spectral domain for speech enhancement, IEEE Int. Conf. Consumer Electron. **1**, 413–414 (2011)
16. Y. Masuyama, K. Yatabe, Y. Oikawa et al., Phase-aware Harmonic/percussive source separation via convex optimization, in: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 985–989
17. G. Jia, G. Liu, Z. Yuan, J. Wu et al., An anomaly detection framework based on autoencoder and nearest neighbor, in: Proceedings of the 2018 15th International Conference on Service Systems and Service Management (ICSSSM), 2018, pp. 1–6
18. C. Tebaldi, R. Knutti, The use of the multi-model ensemble in probabilistic climate projections, Philos. Trans. R. Soc. **365**, 2053–2075 (2007)
19. D.P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv:1312.6114 (2018)
20. S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio et al., Contractive auto-encoders: explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 2011, pp. 833–840
21. A. Oord, S. Dieleman, H. Zen et al., WaveNet: a generative model for raw audio. arXiv:1609.03499 (2016)
22. R. Bivand, J. Hauke, T. Kossowski et al., Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods, Geogr. Anal. **45**, 150-179 (2013)
23. K. He, X. Zhang, S. Ren, J. Sun et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778
24. C. Sheng, L. Yang, G. Xiang et al., MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices, arXiv:1804.07573 (2018)
25. I.-T. Jolliffe, Principal component analysis, J. Market. Res. **87**, 513 (2002)
26. J. Lee, B. Kang, S.H. Kang et al., Integrating independent component analysis and local outlier factor for plant-wide process monitoring, J. Process Control **21**, 1011–1021 (2011)
27. F. Najar, S. Bourouis, N. Bouguila, S. Belghith et al., A comparison between different gaussian-based mixture models, in: IEEE/ACS 14th International Conference on Computer Systems and Applications, 2017, pp. 704–708
28. E. Fonseca, M. Plakal, D.P.W. Ellis, F. Font, X. Favory, X. Serra, Learning sound event classifiers from web audio with noisy labels, in: International Conference on Acoustics, Speech and Signal F Processing, S Brighton, F UK, 2019, pp. 21–25
29. G. Huang, Z. Liu, L.,v. der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 2017, pp. 2261–2269
30. J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: Conference on Computer Vision and Pattern Recognition, S Miami, F FL, 2009, pp. 248–255

Cite this article as: Peng Cui, Xuan Luo, Xiaobang Li, Xinyu Luo, Research on the enhancement of machine fault evaluation model based on data-driven, Int. J. Metrol. Qual. Eng. **13**, 13 (2022)