Self-supervised Anomalous Sound Detection for Machine Condition Monitoring

Ying Zeng(✉), Hongqing Liu, Yu Zhao, and Yi Zhou

School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
s200101158@stu.cqupt.edu.cn

Abstract. Automatic detection of anomalous sounds is very important for industrial equipment maintenances. However, anomalous sounds are difficult to collect in practice, and self-supervised methods have received extensive attentions. It is well-known that the self-supervised methods show poor performances on certain machine types. To improve the detection performance, in this work, we introduce other types of data as targets to train a general classifier. After that, the model has certain prior knowledge, and then we fine tune the parameters of the model for a specific machine type. We also studied the impact of input features on performance, and it is shown that for machine types, filtering out low-frequency noise interference can significantly improve model performance. Experiments conducted using the DCASE 2021 Challenge Task2 dataset showed that the proposed method improves the detection performance on each machine type and outperforms the DCASE 2021 Challenge first-placed ensemble model by 8.73% on average according to the official scoring method.

Keywords: Machine condition monitoring · Anomalous sound detection · Self-supervised learning

1 Introduction

Anomalous sound detection (ASD) is receiving increasing attentions, especially in the industrial fields, where mechanical failures cause companies a great financial risk. Due to the shortage of maintenance workers in companies around the world, there is a growing need for automatic diagnostic technology using machine sounds [1,2].

The purpose of ASD is to identify whether the sound emitted from a machine is normal or abnormal in determining the machine operation status. If the anomaly score of the data exceeds a threshold, the said sound is identified as

This work is supported by the Natural Science Foundation of Chongqing, China (No. cstc2021jcyj-bshX0206).

F. Gao et al. (Eds.): ChinaCom 2022, LNICST 500, pp. 220–230, 2023.
https://doi.org/10.1007/978-3-031-34790-0_17

the outlier. Because of the variability of the anomaly data and the high cost of damaging machines, it is difficult to collect them, and the unsupervised anomaly detection methods are usually preferred.

The Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) has played a great role in advancing this technology. In DCASE 2020 Challenge Task 2, teams from all over the world participated and a good detection performance was achieved [3], but this was under ideal conditions. In order to simulate more realistic scenarios, the DCASE 2021 Challenge added a domain shift condition [4]. The task is performed under the condition that the acoustic properties of training data and testing data are different, i.e., domain shift, and the differences include speed, machine load, ambient noise, etc. Compared to 2020, this task only provides very little normal audio data for the target domain, and this extra setting simulates a more realistic situation.

The traditional approaches are to train an autoencoder that maps normal data to a low-dimensional space through the encoder, and then reconstruct the data with the decoder, and the reconstruction error is used as the loss function [5]. The parameters in the network are continuously updated to reduce the reconstruction error. Since the network is trained with normal data, the basic assumption is that normal data can be well reconstructed after training, but the abnormal data cannot. During the test, the reconstruction error is used as the anomaly score. However, under the influence of complex environments, the model trained by this method often cannot distinguish the normal and abnormal sounds.

The normalized flow (NF) is another commonly used method [6], which is a series of reversible transformations between the input data distribution $p(x)$ and the known distribution $p(z)$ to perform accurate likelihood estimation. The negative log-likelihood function is used as optimization objective and anomaly score. It can fit the distribution of normal samples well, but for other normal samples whose domain is shifted, it is easy to be judged as abnormal.

The latest research is often based on self-supervised learning [7–9], and the assumption is that there are sound data from multiple machines of the same machine type. This is a realistic assumption since multiple machines of the same type are often installed in factories. For a specific machine type, by training a dedicated classifier to distinguish between different machine ID, the model can learn the inherent attributes of the machine ID and anomalies are determined based on the output probability of the corresponding IDs. In this case, the normal sound often outputs a higher probability, while the output probability of abnormal data corresponding to ID is often lower than normal output.

The key to self-supervised anomalous sound detection is to use normal sound samples to learn the inherent properties of normal sound samples, so as to distinguish abnormalities. However, due to the powerful fitting ability of the neural network, it is very easy to distinguish different machine IDs of the same machine type, which means that it is very easy to overfit. Although these methods improved detection performance on average, research showed that the significant low scores were obtained on some machine types, see DCASE 2021. To overcome the above problems, we propose a more robust self-supervised anomaly detection model. First, we use all machine types to train a classifier that can simultaneously distinguish different machine types with different IDs. Second, we fine-tune the model parameters to train a dedicated classifier for each machine type. Since the features of some machine types are mainly concentrated in high frequencies, we design a high-pass filter to filter out low-frequency noise signal to improve the detection performance.

In addition to classifier confidence, Mahalanobis distance is served as an additional option for computing the anomaly scores. The experimental results show that the proposed method is superior to the first-place model of DCASE 2021 challenge task2, especially on the target domain.

The main contributions of this paper can be summarized as follows.

- We found that by selecting an appropriate anomaly score, the detection performance of the same model will be greatly improved.
- We preprocess the input features to reduce the interference caused by noise. Specifically, we adopt a high pass filter for some machine types to filter out the low frequency interference before Mel filtering.
- We train a general anomalous sound detection model, and finally fine tune the parameters of the model to obtain a special anomalous sound detection model for a specific machine type. After this, the model can learn more discriminative and robust latent acoustic representations.

The rest of the paper is organized as follows. Section 2 gives an overview of the proposed method in details. Section 3 describes the experiment settings and the experimental results. Finally, Sect. 4 concludes the paper.

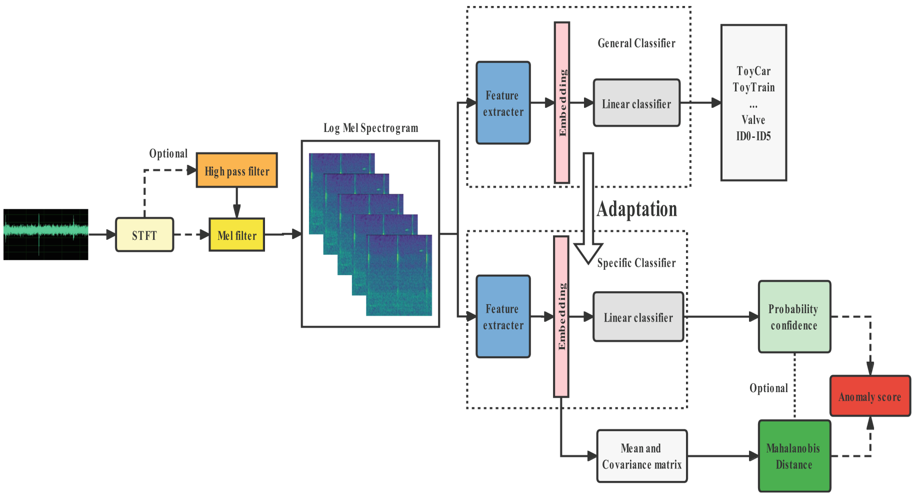
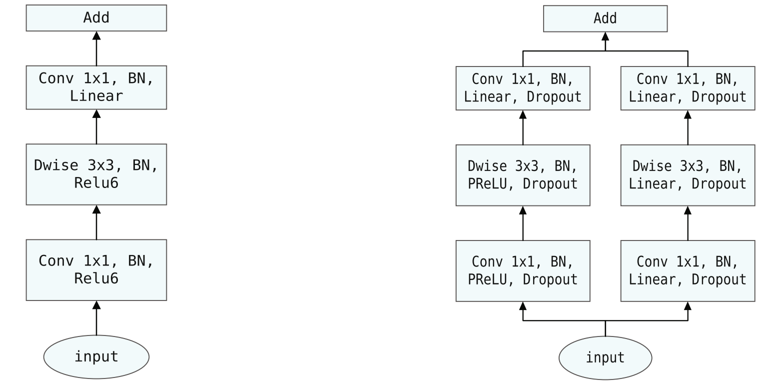


Fig. 1. A overview of our anomalous sound detection model.

2 Proposed Method

The overall structure of our model is shown in Fig. 1. First, we use all machine types to train a classifier that can simultaneously distinguish different machine types with different IDs. Finally, we fine tune the model parameters to train a dedicated classifier for each machine type. In the fine-tuning phase, we use the enhanced input features, the time spectrum after short-time Fourier transform is passed through a high-pass filter with a selectable cut-off frequency. The model will continuously optimize the parameters according to the loss function. We use the training dataset to calculate the average vector and covariance matrix of each class separately to measure the anomaly score.



(a) Original block (b) Proposed dual block

Fig. 2. Comparison of Inverted residuals block when stride = 2.

2.1 Model

We use the inverted residuals proposed by Mobilenet v2 [10] to build our model. Similar to mobilefacenet [11], PRelu is selected as the activation function, and a linear branch without using any activation function is added to reduce the loss of information caused by PRelu. Figure 2 shows the case of stride = 2. When stride = 1, the input is added to the output through skip connection. By utilizing this structure, our final model is depicted in Table 1, where the global depthwise convolution (GDConv) [11] means that the kernel size is equal to the input dimension size, and the bottleneck consists of n numbers of the proposed inverted residuals. This model first uses two layers of 2D Convs to extract features and three layers of the proposed bottleneck, and followed by convolutional layers and finally outputs the probability of section IDs.

Table 1. Model architecture, where k is the number of section IDs, t indicates the expansion factor, c is the output channels, and s is the stride. The first layer of each sequence has a stride s and others use stride 1, and linear means that the activation function is not employed.

Operator	t	c	n	s
Conv2d 3x3	-	64	-	2
Conv2d 3x3	-	64	-	1
Dual block	2	128	2	2
Dual block	4	128	2	2
Dual block	4	128	2	2
Conv2d 1x1	-	512	-	1
Linear GDConv2d	-	512	-	1
Linear Conv2d 1x1	-	128	-	1
Dropout	-	-	-	-
Conv2d 1x1	-	k	-	-

2.2 Loss Function

For self-supervised anomaly sound detection, a classifier is trained as an auxiliary task, and usually the softmax loss function is used as the loss function. We also select the center loss [12] as the loss function. Compared with using only the softmax loss function, it effectively increases the inter-class distance and reduces the intra-class distance. Ideally, since the same kind is more concentrated, for normal test samples, the anomaly score will be smaller than using only softmax loss. Since abnormal sounds are not involved in training, its anomaly score has little change compared with only softmax loss, and then AUC and pAUC can be improved. The joint loss function now is

$$\mathcal{L}_S = -\sum_{i=1}^{m}\sum_{j=1}^{k} y_{ij} \log \hat{y}_{ij}, \tag{1}$$

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2, \tag{2}$$

$$\mathcal{L} = \lambda \mathcal{L}_S + \mathcal{L}_C, \tag{3}$$

where (1) and (2) respectively represent the softmax loss and center loss, λ is a tunable parameter, and m is the size of mini-batch. In (1), $\hat{y}_i \in \mathbb{R}^k$ denotes that the i-th sample passes through the output of the network and then passes through the output value of the softmax function, y_i is the one-hot encoding of the label, $\hat{y}_{ij}$ and y_{ij} are the j-th values of $\hat{y}_i$ and y_i, respectively. In (2), $c_{y_i} \in \mathbb{R}^{128}$ denotes the real class center of i-th sample and x_i is the output of Linear Conv2d 1x1 layer. In actual training, we first initialize a class center for each class, and then continuously update the class center.

2.3 Calculation of Anomaly Score

The calculation of anomaly scores is very important. We found that the results obtained by using different calculation anomaly scores for the same model are often very different. This indicates that the model might be good enough, but we only need to select appropriate anomaly scores for different machine types.

Classifier Confidence. The models is trained to identify from which section the observed signal was generated and outputs softmax value for each machine section ID. After feature extraction, the features of an audio segment are segmented into B segments. The anomaly score is calculated as

$$Score = \frac{1}{B}\sum_{i=1}^{B} \log \frac{1 - P_i}{P_i}, \tag{4}$$

where P_i is the prediction probability of machine ID corresponding to the i-th feature segment.

Mahalanobis Distance. Mahalanobis distance [13,14] is commonly used as an indicator of similarity measure. During the experiments, it is found that using Mahalanobis distance as anomaly score effectively classify anomalies, and it also shows a high performance under the condition that the target domain has very few normal data. The formula for calculating Mahalanobis distance is

$$Score_{ma} = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}, \tag{5}$$

where $\mathbf{x}$ is the mean vector, μ is the mean vector representation corresponding to the ID and $\boldsymbol{\Sigma}$ is the covariance matrix corresponding to the ID. For the source domain and the target domain, we use the data of different domains to calculate the average vector, and use all the training data of the machine ID to calculate the covariance matrix. Because the amount of data corresponding to the target domain is too small, the covariance matrix will be inaccurate if calculation only uses the data from the target domain.

2.4 Data Augmentation

Data augmentation is an effective way to improve neural network generalization and prevents overfitting. In our system, we use the Mixup method [15] in the training phase to randomly introduce other samples as noise to the original data. The mixing operation on the training samples is

$$\tilde{x} = \eta x_i + (1 - \eta) x_j, \tag{6}$$

where x_i is considered as the original sample and x_j is considered as the random sample of this mini-batch.

Since the sound characteristics of some machine types are mainly concentrated in high frequencies, we pass a high-pass filter with a manually tuneable cutoff frequency to produce better embedding.

3 Experiments

3.1 Datasets

We conduct experiments using the DCASE 2021 Challenge Task2 dataset, which includes recordings of 10 s in length and a sampling rate of 16 kHz, and in this dataset, ToyCar and ToyTrain belong to ToyADMOS2 [16], and Fan, Gearbox, Pump, Slider, Valve belong to MIMIIDUE [17]. The dataset for each machine type has 6 different machine numbers, and for each number, about 1000 normal samples belong to the source domain and 3 normal samples are in the target domain. We use sections 3, 4, and 5 for each machine type as the test set. For both domains, the same number of test samples is provided, with approximately 100 normal samples and 100 abnormal samples.

3.2 Experimental Setup

We load the audio data using the default sample rate and apply a short time Fourier transform (STFT) with a window size of 1024 and a hop length of 512 samples. We pass the data through a high pass filter, and then convert the STFT spectrogram into a Mel spectrogram with a 128-band Mel filter. After this, we generate a 128 $\times$ 313 logarithmic Mel spectrogram. We used Adam as the optimizer and the learning rate of the model was set to 0.001. Training runs with a batch size of 64 for 50 epochs.

Table 2. The effect of anomaly scores on results, where m represents Mahalanobis distance and p represents probability confidence.

	m		p	
	AUC	pAUC	AUC	pAUC
ToyCar	**62.92%**	**54.42%**	54.58%	54.17%
ToyTrain	**68.21%**	**56.04%**	49.04%	49.50%
fan	72.64%	54.59%	**86.15%**	**71.45%**
gearbox	**66.18%**	**57.68%**	52.02%	51.95%
pump	63.31%	54.94%	**78.81%**	**67.57%**
slider	74.17%	61.30%	**89.15%**	**73.82%**
valve	57.51%	52.46%	**59.23%**	**59.67%**

3.3 Results and Discussions

To show the performance, we evaluate the detection performance of the area under the receiver operating characteristic curve (AUC) and the partial AUC (pAUC) with $p = 0.1$.

Table 3. Anomaly detection results for different machine types, where General represents pre-training with all data, and Specific represents the final result after fine-tuning.

	Machine Type	Baseline	Top1	general	specfic
AUC	ToyCar	65.93%	75.27%	65.09%	**78.40%**
	ToyTrain	68.51%	69.15%	71.24%	**73.43%**
	fan	60.68%	61.01%	86.15%	**91.44%**
	gearbox	65.49%	63.07%	71.09%	**73.40%**
	pump	58.30%	86.76%	78.81%	**89.06%**
	slider	57.22%	83.18%	**89.15%**	89.11%
	valve	51.87%	65.36%	74.78%	**80.03%**
pAUC	ToyCar	52.32%	59.71%	58.24%	**68.73%**
	ToyTrain	57.56%	59.91%	59.33%	**62.19%**
	fan	50.50%	60.79%	71.45%	**82.78%**
	gearbox	56.86%	61.56%	60.48%	**64.06%**
	pump	50.98%	**81.55%**	67.57%	77.32%
	slider	51.41%	63.60%	73.82%	**75.03%**
	valve	50.07%	60.15%	61.03%	**67.18%**
	harmonic mean	56.38%	66.80%	69.43%	**75.53%**

Table 4. Results for different machine types in Source Domain and Target Domain. h-mean means harmonic mean.

	Source domain				Target domain			
	Top1		ours		Top1		ours	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
ToyCar	**81.44%**	59.05%	78.63%	**68.94%**	69.97%	60.39%	**78.18%**	**68.53%**
ToyTrain	**77.56%**	62.21%	75.89%	**62.83%**	62.38%	57.78%	**71.12%**	**61.56%**
fan	51.45%	61.70%	**94.63%**	**85.43%**	74.93%	59.91%	**88.45%**	**80.29%**
gearbox	63.52%	61.38%	**76.13%**	**64.02%**	62.62%	61.75%	**70.87%**	**64.10%**
pump	**88.72%**	**82.19%**	86.39%	73.11%	84.88%	80.91%	**91.90%**	**82.05%**
slider	85.56%	66.10%	**91.10%**	**75.85%**	80.92%	61.28%	**87.20%**	**74.24%**
valve	69.56%	64.03%	**79.35%**	**68.06%**	61.64%	56.71%	**80.72%**	**66.33%**
h-mean	71.66%	64.56%	**82.59%**	**70.49%**	70.00%	61.91%	**80.45%**	**70.26%**

Table 2 shows the performance of the general anomalous sound detection model without using a high-pass filter. It is obvious that choosing different anomaly scores for the same model has a great impact on the results, which means that the model is often good enough, and the only thing to do is to choose an appropriate anomaly score. In order to better compare the performance of the model, we choose the anomaly scores that are more suitable for this machine type from the Mahalanobis distance and probability confidence in the following results.

In Table 3, baseline is built by Autoencoder, a classic method of anomalous sound detection. Top1 is the first-placed model in the DCASE 2021 challenge [7], which is an ensemble of three different models including two self-supervised classifier models and a probabilistic model. The harmonic mean obtained in the table is the average of AUC and pAUC. The harmonic mean of our general model reach 69.43%, which is a great improvement over previous methods. After fine-tuning the model parameters to fit the specific machine type, our method finally reached 75.53%. It can be clearly seen that our method outperforms the previous results.

In Table 4, the detection results in the source domain and the target domain are shown. The average performance of our model outperforms state-of-the-art models in both source and target domains. In particular, the performance of the target domain is only slightly degraded compared to the source domain, which means that our method is very robust.

Table 5. The impact of high-pass filtering on detection performance, we do not show the results of fan and pump after high-pass filtering because the model cannot converge.

	w/o hpss		hpss	
	AUC	pAUC	AUC	pAUC
ToyCar	62.92%	54.42%	**65.09%**	**58.24%**
ToyTrain	68.21%	56.04%	**71.24%**	**59.33%**
fan	**86.15%**	**71.45%**	-	-
gearbox	66.18%	57.68%	**71.09%**	**60.48%**
pump	**78.81%**	**67.57%**	-	-
slider	**89.15%**	**73.82%**	86.41%	69.16%
valve	59.23%	59.67%	**74.78%**	**61.03%**

To verify the effectiveness of the high-pass filter, we conducted ablation experiments to compare the performance without and with high-pass filtering. In Table 5, we found that the performance after the high-pass filter has been improved, which shows that filtering out a low-frequency noise is beneficial for detections. Specifically, for fan, pumps and slider, the detection performance will decrease, and we conjecture that it may be because the low-frequency signals of these machine types contain a lot of feature information. Table 6 shows the performance of our specfic model when using the centerloss. For fan, gearbox, pump and valve, centerloss can improve the performance. But the performance of other machine type has been decrease, we speculate that it may be because the centerloss destroys the details of features, so that the abnormal sound is mapped to the normal feature space. In Table 7, we summarize the best experimental setups in the fine-tuning stage.

Table 6. The impact of centerloss on detection performance.

	w/o centerloss		centerloss	
	AUC	pAUC	AUC	pAUC
ToyCar	**78.40%**	**68.73%**	65.12%	61.10%
ToyTrain	**73.43%**	**62.19%**	67.11%	56.97%
fan	89.70%	78.66%	**91.44%**	**82.78%**
gearbox	71.56%	62.59%	**73.40%**	**64.06%**
pump	80.35%	69.37%	**89.06%**	**77.32%**
slider	**89.11%**	**75.03%**	81.21%	66.48%
valve	76.20%	61.50%	**80.03%**	**75.53%**

Table 7. Experiment configurations, f_{min} represents the cut-off frequency of high pass filter.

	ToyCar	ToyTrain	Fan	Gearbox	pump	Slider	Valve
Centerloss	False	False	True	True	True	False	True
Highpass filter	True	True	False	True	False	False	True
f_{min}	1000	2000	0	2000	0	0	2000

4 Conclusion

We propose a self-supervised anomaly detection model that outperforms the DCASE 2021 first-placed Ensemble model. To achieve that, We first use all the training data to train a general anomalous sound detection model, and then fine tune the model parameters to better adapt to specific machine types. The findings are that for some machine types, the model will perform better by simply filtering out low frequency noise interference. Experiments show that the anomaly detection ability can be effectively improved by using our method. With less target domain data, our method achieves a superior performance in the target domain.

References

1. Koizumi, Y., Saito, S., Uematsu, H., Harada, N.: Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma. In: 25th European Signal Processing Conference (EUSIPCO), pp. 698–702. IEEE (2017)
2. Koizumi, Y., Saito, S., Uematsu, H., Kawachi, Y., Harada, N.: Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 212–224 (2018)
3. Koizumi, Y., et al.: Description and discussion on DCASE 2020 challenge task2: unsupervised anomalous sound detection for machine condition monitoring. arXiv preprint arXiv:2006.05822 (2020)

4. Kawaguchi, Y., et al.: Description and discussion on DCASE 2021 challenge task 2: unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions. arXiv preprint arXiv:2106.04492 (2021)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**(3), 1–58 (2009)
6. Dohi, K., Endo, T., Purohit, H., Tanabe, R., Kawaguchi, Y.: Flow-based self-supervised density estimation for anomalous sound detection. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 336–340. IEEE (2021)
7. Lopez, J., Stemmer, G., Lopez-Meyer, P., Singh, P.S., del Hoyo Ontiveros, J.A., Courdourier, H.A.: Ensemble of complementary anomaly detectors under domain shifted conditions. DCASE2021 Challenge, Technical report (2021)
8. Morita, K., Yano, T., Tran, K.: Anomalous sound detection using CNN-based features by self supervised learning. DCASE2021 Challenge, Technical report (2021)
9. Wilkinghoff, K.: Utilizing sub-cluster adacos for anomalous sound detection under domain shifted conditions. DCASE2021 Challenge, Technical report (2021)
10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
11. Chen, S., Liu, Y., Gao, X., Han, Z.: MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices. In: Zhou, J., et al. (eds.) CCBR 2018. LNCS, vol. 10996, pp. 428–438. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97909-0_46
12. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A comprehensive study on center loss for deep face recognition. Int. J. Comput. Vis. **127**(6), 668–683 (2019)
13. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The mahalanobis distance. Chemom. Intell. Lab. Syst. **50**(1), 1–18 (2000)
14. Xiang, S., Nie, F., Zhang, C.: Learning a mahalanobis distance metric for data clustering and classification. Pattern Recogn. **41**(12), 3600–3612 (2008)
15. Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., Michalak, S.: On mixup training: improved calibration and predictive uncertainty for deep neural networks. arXiv preprint arXiv:1905.11001 (2019)
16. Harada, N., Niizumi, D., Takeuchi, D., Ohishi, Y., Yasuda, M., Saito, S.: ToyADMOS2: another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions. arXiv preprint arXiv:2106.02369 (2021)
17. Tanabe, R.: MIMII due: sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions. arXiv preprint arXiv:2105.02702 (2021)