The Anomaly Sound Detection Method Based on the Dual-Path CNN and the Autoencoder

Technical Report

Chao Chen[1], *Peng Wu*[2], *Pengqi Wang*[1], and *Bo Ma*[1]

[1]College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing, China
[2]College of Information Science & Technology, Beijing University of Chemical Technology, Beijing, China
(2024210473@buct.edu.cn;2023400229@buct.edu.cn;
PengQ@buct.edu.cn; mabo@mail.buct.edu.cn)

ABSTRACT

This report contains a description of the systems submitted to task 2 "First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring" of the DCASE2025 Challenge. The anomaly detection model based on the attention mechanism and the convolutional networks enhanced autoencoder (ACAE) is proposed. In addition, we introduced the DensitySoftmax and the dynamic topic mixture model (DtMM) into the previous unsupervised model to represent the distance between abnormal samples and normal samples. In experimental evaluations, it is shown that both modifications improve the resulting performance and that the proposed. By introducing domain generalization methods, our model achieved improved metrics on the target domain compared to the baseline model.

Index Terms— anomalous sound detection, unsupervised, domain generalisation, anomaly detection, threshold

1. INTRODUCTION

Task 2 of the DCASE2025 Challenge[1] is called "First-Shot Unsupervised Anomalous Sound Detection (ASD) for Machine Condition Monitoring". ASD focuses on identifying whether the sound emitted by the target machine is abnormal by solely relying on prior knowledge of normal sounds. This is the primary focus of Task 2 in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge [2-4].

The complexity of this task lies in distinguishing normal operational noise from genuine anomalies, requiring sophisticated algorithms capable of learning from diverse acoustic patterns. In practical production environments, the diversity of equipment types, complex surroundings, and challenges with sound data collection make it difficult to develop systems that can accurately identify and classify abnormal sounds across different devices and environments. The interpretations of the requirements for this task are as follows:

- Only the normal data is used to train the model. This method is suitable for scenarios where normal samples are abundant and abnormal samples are scarce. Through this unsupervised method, the model can learn the characteristics of normal data, thereby accurately identifying abnormal data when it encounters it.
- The model has domain generalisation capabilities. Aims to solve the problem of generalisation in target domains not seen during model training. Its core objective is to enable the model to learn features or patterns with generalisation capabilities from the source domain, thereby performing well in unknown but related domains.
- The model can be used for brand new machine types. In practical applications, ASD systems may need to be deployed on a variety of different types of machines or systems. These machine types may differ significantly in many ways. Through data preprocessing, selecting appropriate models, and using transfer learning and domain adaptation techniques, the generalisation ability of the model can be improved.
- The model can be trained with or without attribute information. The ability to be trained with or without attribute information is key to achieving model flexibility and adaptability. By designing models with optional feature inputs, performing multi-task learning, standardized data pre-processing, and conducting multi-scenario validation, it is possible to ensure that the model works effectively with or without attribute information.

This year, submitted systems not only had to be trained using only normal data and to be robust to possible acoustic domain shifts, which may for example be caused by changing machine parameters or the background noise, but also had to be capable to effectively handle completely novel machine types without having access to recordings of similar machines and not always having access to machine parameter settings.

2. SYSTEM DESCRIPTION

2.1. Models based on the dual-branch CNN

The systems described in this chapter is based on the dual-branch convolutional neural networks model [5]. We use the DensitySoftmax and the DtMM to replace the original K-means method to calculate the anomaly distance.

In terms of data preprocessing, two different input features are extracted. As a first input feature, a magnitude spectrogram with a window size of 1024 and a hop size of 512 is used. For each magnitude spectrogram, the temporal mean, calculated by using only the frames belonging to the non-padded signal values, is subtracted to remove constant frequency information. As a second input feature, the magnitude of the spectrum belonging to the entire waveform is used. Only frequencies up to 8 kHz are retained, which is equal to half of the sampling rate, are kept.

The embedding model consists of two different CNNs as sub-networks, one for each of the two feature branches, and has the same general architecture as the embedding model used in before[5]. The sub-network for the spectrogram branch has a modified ResNet architecture [6] with four residual blocks, a max-pooling operation over the time dimension in combination with a flattening operation and a linear layer. The sub-network for the spectrum branch consists of three one-dimensional convolutions with large strides are applied to downsample the input followed by a flattening operation and and five dense layers. In both networks, ReLU is chosen as an activation function and batch normalization [7] is applied. To obtain a single embedding for each input sample, the embedding of both sub-networks are concatenated. More details about the network architecture can be found in the paper[8].

2.1.1 The CNN-DS model

In terms of abnormal distance detection, Traditional methods typically rely on strategies such as distance metrics, reconstruction errors, or probability density modelling to determine whether test samples deviate from the normal distribution. In recent years, with the development of deep representation learning, researchers have increasingly favoured anomaly detection in embedding spaces. Among these, the DensitySoftmax method[9-11], as a probability density-based anomaly scoring formula, combines Gaussian density estimation with softmax classification principles, offering good interpretability and detection performance, particularly suitable for unsupervised or semi-supervised anomaly detection tasks.

The DensitySoftmax method assumes that in the feature space, the embedded features $f(x)$ of each category follow a multidimensional Gaussian distribution. Specifically, for each category c_i in the known training data (typically normal data), we can separately construct the distribution parameters of that category in the embedding space: the mean vector μi and the covariance matrix Σi. During the inference stage, for any test sample x, DensitySoftmax calculates its probability density $p(x \mid c_i)$ across all class distributions and normalises it via softmax to obtain its relative 'similarity' distribution across all classes. If this distribution is not concentrated (e.g., the density values are smoothly distributed or all densities are very small), the sample can be deemed to have a high degree of anomaly.

Strong probabilistic modelling capabilities: Compared to simple centre distance or reconstruction error, this method can model category boundaries and distribution structures, making it suitable for handling multimodal tasks with high intra-class variability. No need for training with outlier samples: Only normal samples are required for modelling, making it suitable for one-class learning scenarios. Simple structure, compatible with any encoder: The method relies solely on the embeddings of training samples and does not depend on network structure, giving it strong generalisability.

2.1.2 The CNN-DtMM model:

The DtMM is the dynamic topic model (DTM) combined with a *t*-distribution early warning model modelling method. This method fixes the order structure of sub-components within the mixture model and establishes the parameter evolution relationship between the mixture model and the baseline mixture model based on DTM modelling principles. The new mixture model evolves from the baseline mixture model combined with input data. By calculating the difference between the mixture model and the baseline mixture model, it characterises the difference between the real-time operating condition of the equipment and its normal state, thereby achieving condition monitoring of the equipment.

Construct the training sample set based on the reduced-dimensional feature vectors. The training sample set is divided into the model training sample set and the threshold training sample set. The specific steps for training the DtMM model are as follows:

1) Input the model training sample set and use the Dirichlet process to generate the prior parameters of the mixture model, i.e., the number of submodels and weight parameters in the baseline mixture model.

2) Based on the results from step 1), combine variational inference to calculate the posterior parameters of the baseline mixture model, obtain the baseline mixture model, and save it.

3) Input the threshold training sample set, train m normal state mixture models, and calculate the distance between each mixture model and the baseline mixture model, i.e., the KL divergence. Calculate the mean μ and standard deviation σ of the m KL divergences, and the threshold is set, as shown in Eq (1).

$$T = \mu + k\sigma \quad (1)$$

Input the reduced-dimensional feature vector of the sample to be tested, establish a mixed model, and calculate the KL divergence with the benchmark mixed model. If KL > T, it is determined to be abnormal, triggering a fault warning;if KL < T, it is determined to be normal.

2.2. The ACAE model

To effectively address the domain generalization problem in acoustic signal anomaly detection, The novel anomaly detection model based on the attention mechanism and the convolutional networks enhanced autoencoder (ACAE) is proposed, where the noisy-arcmix loss function and the minimum covering volume (MCV) method are introduced. This method primarily consists of four parts: the feature extraction, the anomaly detection model construction, the self-learning threshold, and the model testing.

1) The feature extraction: The short-time Fourier transform (STFT) is used to obtain the time-frequency domain feature, and the averaging strategy based on the time window is introduced, which not only effectively suppresses the transient noise interference but also enhances the robustness and expressiveness of features in the time dimension.

2) The anomaly detection model construction: The encoder is employed to extract the deep features through the convolution and pooling operations. Moreover, the lobal dependency modelling capabilities is improved through the multi-head attention module, the features independent of domain attributes are extracted. The decoder is used for signal reconstruction and assists in effectively constraining the latent space. During the training process, the mean squared error (MSE) and the noisy-arcmix are used as loss function. The noisy-arcmix is an improved version of the arcmix loss function, which integrates contrastive learning strategies with angular boundary constraints [15]. The embedding features that are both discriminative and robust are constructed by regulating the angles and performing mixed interpolation on the distributions of different class samples in the latent space.

3) The self-learning threshold: The threshold self-learning method based on the cosine similarity and the MCV is adopted. The principle the MCV is to take the mean of normal sample features as the center of the sphere and construct the smallest radius sphere that can cover all or most normal samples. Specifically, the cosine similarity of the normal samples is calculated, and the minimum sphere coverage radius corresponding to the farthest normal sample is obtained through the MCV with the cosine centre as the sphere centre, which is set the threshold.

4) The model testing: The test sample are processed using the STFT to extract time-frequency features, and the potential features are extracted using the trained ACAE model. The cosine similarity between the test sample and the reference normal sample is calculated, and the relative position of the test sample in the feature space is obtained in combination with the MCV and compared with the threshold.

3. EXPERIMENT RESULTS

In the released DCASE2025 development set [15,16], we compare our systems with the baseline systems of the DCASE 2025 Challenge Task 2, i.e., AE-MSE and AE-MAHALA [17]. The results are given in Table 1, where we can see that Our system outperforms the baseline system in the target domain.

4. CONCLUSION

In this report, we submitted three ASD systems for Task 2 of the DCASE2025 competition. Systems 1 and 2 were obtained by improving the anomaly distance calculation model on the existing unsupervised system, and System 3 proposes an anomaly detection model based on attention mechanisms and convolutional neural network-enhanced autoencoders. The experimental results show that our system significantly outperforms the baseline system.

Table 1: Detection results of three submitted systems and baseline systems on the development set

Machine	Metric	Baseline_mse	Baseline_MAHALA	CNN-DS	CNN-DtMM	ACAE
bearing	AUC-s	66.5%	63.6%	53.0%	55.0%	55.0%
	AUC-t	53.2%	59.0%	54.4%	57.0%	61.7%
	pAUC	61.1%	61.9%	54.4%	55.7%	58.1%
fan	AUC-s	80.0%	78.0 %	56.6%	56.0%	47.4%
	AUC-t	38.8%	38.6%	52.0%	50.0%	57.7%
	pAUC	49,5%	50.8%	50.2%	50.3%	51.9%
gearbox	AUC-s	64.8%	73.3%	60.3%	43.0%	41.4%
	AUC-t	50.5%	51.6%	77.0%	56.0^	57.0%
	pAUC	52.5%	55.1%	47.7%	49.9%	47.8%
slider	AUC-s	70.1%	73.8%	54.2%	50.0%	46.3%
	AUC-t	48.8%	50.3%	55.5%	48.0%	59.1%
	pAUC	52.3%	53.6%	50.4%	49.1%	52.7%
ToyCar	AUC-s	71,1%	73.2%	42.5%	58.0%	50.0%
	AUC-t	53.5%	50.9%	54.4%	53.0%	46.9%
	pAUC	49.7%	49.1%	49.5%	50.4%	42.7%
ToyTrain	AUC-s	61.8%	50.9%	36.6%	47.0%	55.7%
	AUC-t	56.5%	46.2%	61.6%	50.0%	52.6%
	pAUC	50.2%	48.3%	50.9%	48.7%	53.7%
valve	AUC-s	63.5%	56.2%	46.3%	85.0%	57.6%
	AUC-t	67.2%	61.0%	72.0%	69.0%	47.3%
	pAUC	57.4%	52.5%	47.7%	62.9%	52.6%

5. REFERENCES

[1] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. *Description and discussion on DCASE 2025 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring. In arXiv e-prints: 2506.10097*, 2025.

[2] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 Challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *in Proc. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2023, pp. 31–35.

[3] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi,
K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 Challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Barcelona, Spain, November 2021, pp. 186–190.

[4] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura,Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo,M. Yasuda, and N. Harada, "Description and discussion on DCASE 2020 Challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc.of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, November 2020, pp. 81–85.

[5] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2024, pp. 276–280.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, 2016, pp. 770–778.

[7] S. Ioffe and C. Szegedy, "Batch normalization: acceleratingdeep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning (ICML)*, vol. 37, 2015, pp. 448–456.

[8] K. Wilkinghoff, "Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[9] K. Lee, H. Lee, K. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7167–7177, 2018.

[10] Y. Yang, Z. Wang, X. Zhao, and D. Lin, "Generalized energy-based open-set recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7463–7472, 2021.

[11] Y. Sun, L. Zhang, X. Liang, X. Zhao, and D. Lin, "OpenGAN: Open-set recognition via open generative adversarial network," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, no. 2, pp. 2118–2126, 2022.

[12] BLEI D M, LAFFERTY J D. “Dynamic topic models” in *23rd International Conference on Machine Learning. Pittsburgh*, 2006: pp. 113-120.

[13] Zhang Mingguang, Luo Xuelin, Jia Deng, et al. “Early warning method for reciprocating machinery failures based on dynamic topic models,”. *Journal of Beijing University of Chemical Technology (Natural Science Edition)*, vol. 50, 2023, pp. 88-97.

[14] S. Choi and J.-W. Choi, "Noisy-arcmix: Additive noisy angular margin loss combined with mixup for anomalous sound detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1-10.

[15] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro Saito. *ToyADMOS2: another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions.* In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 1–5. Barcelona, Spain, November 2021.

[16] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei Kawaguchi. *MIMII DG: sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task.* In Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022). Nancy, France, November 2022.

[17] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda. *First-shot anomaly detection for machine condition monitoring: a domain generalization baseline. Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, pages 191–195, 2023.