FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION BASED ON FUSION OF CNN-AE AND BEATs-KNN

Technical Report

Wang Xiaoliang[1], Ming Ao[1]

Zhejiang University, Hangzhou, 310027
22431047@zju.edu.cn
22431156@zju.edu.cn

ABSTRACT

This technical report presents our submission to the DCASE 2025 Challenge Task 2. We propose a fusion-based system combining a CNN-BiGRU-Attention Autoencoder with a BEATs-KNN model to improve unsupervised anomalous sound detection (ASD). Both models are independently trained and then combined at the score level using a weighted average strategy. The fusion weights are optimized using the development dataset. Results show that this hybrid approach improves the robustness of anomaly detection across multiple machine types. Through the fusion of various models and methods, we have achieved a hmean of 66.00% on the development dataset.

Index Terms— Anomalous sound detection, CNN, BEATs, autoencoder, KNN, model fusion

1. INTRODUCTION

Anomalous Sound Detection (ASD) has become a critical component in predictive maintenance for industrial systems, especially in environments where anomalous data are scarce or unavailable [1, 2, 3]. The DCASE 2025 Task 2 aims to address this challenge by evaluating systems on unseen machine types using only normal sound samples for training. Task 2 of the DCASE 2025 Challenge [4, 5, 6, 7] targets the problem of detecting abnormal acoustic events from various machine types in an unsupervised manner, using a first-shot learning setting for condition monitoring.

To tackle this, we propose a hybrid model named CBFusion, which combines a CNN-BiGRU-Attention-based Autoencoder and a pre-trained BEATs model [8, 9]. The goal is to exploit complementary characteristics—CNN-BiGRU learns local temporal-frequency patterns, while BEATs captures global audio representations pre-trained on large-scale datasets.

This fusion strategy is motivated by prior research showing that diverse representation learning significantly improves robustness under domain-shift conditions. By linearly combining anomaly scores from both models and optimizing weights using development data, the proposed system aims to enhance anomaly discrimination.

2. SYSTEM OVERVIEW

2.1. CNN-BiGRU-Attn Autoencoder

Our first backbone is a convolutional autoencoder extended with BiGRU layers and an attention mechanism. The CNN layers extract spatial features from spectrograms; BiGRU captures temporal dependencies; and the attention mechanism focuses on informative time steps. This network reconstructs the input, and the mean square error (MSE) between input and output is used as the anomaly score.

2.2. BEATs Pre-trained Embedding + KNN

We use a pre-trained BEATs encoder to extract global embeddings for input waveforms. A 1-nearest-neighbor (1-NN) search is then conducted against the training set in the embedding space using cosine distance. This distance serves as the anomaly score for each test input [10, 11].

2.3 Score Fusion Strategy

Let s_{cnn} and s_{beats} denote the normalized anomaly scores from the CNN and BEATs models, respectively. The final fused score is:

$$s_{\text{fusion}} = \alpha \cdot s_{\text{cnn}} + (1-\alpha) \cdot s_{\text{beats}} \quad (1)$$

We use the labeled development set to perform grid search over $\alpha \in [0,1]$ to find the optimal fusion weight that maximizes AUC or pAUC.

3. EXPERIMENTS

We evaluate the system on the DCASE2025 Task 2 development dataset, which contains machine sounds from both source and target domains across multiple machine types. The CNN-BiGRU-Attn model is trained using only normal samples, and BEATs uses pre-trained weights without finetuning. The fusion system does not require retraining of individual models.Based on different fusion weights, we designed two systems.

Table 1: Detection results on the development set

Machine	Metric	System1	System2	System3	System4
bearing	AUC-S	66.00	66.46	63.78	62.57
	AUC-T	50.16	49.60	51.64	50.16
	pAUC	52.88	52.88	60.16	52.88
fan	AUC-S	66.16	72.56	70.76	78.80
	AUC-T	51.43	48.82	54.66	37.74
	pAUC	53.25	48.42	60.16	60.16
gearbox	AUC-S	70.10	48.20	62.66	73.04
	AUC-T	51.64	39.30	49.32	51.20
	pAUC	49.32	48.26	54.57	54.57
slider	AUC-S	63.14	72.54	68.78	72.54
	AUC-T	54.66	37.74	48.40	50.74
	pAUC	50.37	53.21	51.84	51.84
Toycar	AUC-S	63.78	62.52	70.10	62.56
	AUC-T	49.82	50.72	51.64	48.20
	pAUC	60.16	54.57	60.16	60.16
ToyTrain	AUC-S	70.76	78.80	63.14	48.20
	AUC-T	38.18	51.12	54.66	39.30
	pAUC	49.32	53.21	50.37	50.37
valve	AUC-S	62.66	57.56	63.90	57.56
	AUC-T	49.32	51.12	67.92	51.12
	pAUC	52.05	54.68	64.68	64.68
hmean	AUC-S	66.00	64.77	65.99	64.78
	AUC-T	50.16	48.19	50.16	48.19
	pAUC	52.88	52.85	60.47	60.47

4. CONCLUSION

By fusing local reconstructions (via CNN-BiGRU-Attn) and global embeddings (via BEATs), our system CBFusion improves anomaly detection performance under the first-shot, label-scarce condition of DCASE2025 Task 2. The fusion weights are optimized based on the development set, and the resulting anomaly scores are directly used for the final evaluation.

5. REFERENCES

[1] N. Harada, D. Niizumi, D. Takeuchi, et al., "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," DCASE Workshop, Barcelona, Spain, 2021.

[2] T. Nishida, N. Harada, D. Niizumi, et al., "Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2406.07250, 2024.

[3] K. Dohi, T. Nishida, H. Purohit, et al., "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," DCASE Workshop, Nancy, France, 2022.

[4] N. Harada, D. Niizumi, D. Takeuchi, et al., "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," Proc. 31st European Signal Processing Conference (EUSIPCO), pp. 191−195, 2023.

[5] Z. Lv, B. Han, Z. Chen, et al., "Unsupervised anomalous detection based on unsupervised pretrained models," DCASE2023 Challenge, Tech. Rep., June 2023.

[6] 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 sound detection for machine condition monitoring under domain shifted conditions, ″ arXiv preprint arXiv:2106.04492, 2021.

[7] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2506.10097, 2025.

[8] L. Alzubaidi, J. Zhang, A. J. Humaidi, et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data, vol. 8, pp. 1−74, 2021.

[9] S. Chen, Y. Wu, C. Wang, et al., "BEATs: Audio pre-training with acoustic tokenizers," arXiv preprint arXiv:2212.09058, 2022.

[10] S. Zhang, X. Li, M. Zong, et al., "Learning k for KNN classification," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 8, no. 3, pp. 1−19, 2017.

[11] D. Niizumi, Y. Koizumi, N. Harada, ″ DCASE 2020 Challenge Task 2 Technical Report,″ 2020.