NCUT SYSTEM FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Xichang Cai

North China University of Technology
School of AI and Computer Science
Beijing 100144, China
caixc_ip@126.com

Jiafeng Li, Shenghao Liu

North China University of Technology
School of AI and Computer Science
Beijing 100144, China
jiafeng.l@qq.com

ABSTRACT

Anomalous sound detection (ASD) plays a crucial role in machine condition monitoring, especially in scenarios where collecting anomalous data is impractical. In this report, we propose a First-Shot Unsupervised Anomalous Sound Detection method that requires only normal sound recordings during training. Our approach leverages multiple pre-trained audio embedding models to extract rich and diverse feature representations from machine sounds. Each embedding is evaluated using a K-Nearest Neighbors (KNN) algorithm to compute anomaly scores without supervision. To further improve detection performance and robustness, we perform model-level score fusion by combining the outputs from different embedding models. Experiments conducted on public datasets demonstrate that our method achieves competitive performance in first-shot and low-resource settings, with strong generalization capabilities across machine types and environments. This framework offers a practical and scalable solution for industrial anomaly detection applications.

***Index Terms*—** anomaly detection, pre-trained models, K-Nearest Neighbors

1. INTRODUCTION

Anomalous Sound Detection (ASD) is a key task in machine condition monitoring that aims to detect abnormal machine behavior by analyzing acoustic signals. Since mechanical failures often manifest as changes in running sound, ASD offers a non-invasive, cost-effective, and flexible approach for early fault detection. Particularly in industrial environments, ASD is well suited for monitoring a wide range of machine types without requiring physical contact or complex sensor setups. The unsupervised setting—where only normal sound data is available during training—is especially important in real-world scenarios, where collecting anomalous samples is challenging due to their rarity and unpredictability.

This task builds upon the series of DCASE challenges from 2020 to 2024, and introduces several important refinements in DCASE 2025 Task 2[1]–[3]. Like previous years, participants must build models using only normal sounds (unsupervised ASD, or UASD) and handle domain shifts caused by changes in machine operation or background noise. Since DCASE 2023, the challenge has also required models to adapt to completely new machine types, with no hyperparameter tuning allowed post-deployment. This year, a new element has been introduced: participants may optionally use additional clean machine sounds or noise-only recordings to improve performance[4]. This addition reflects real-world scenarios where such supplementary data may be easier to collect during machine idle times or factory downtime. Furthermore, systems are expected to function with or without auxiliary attribute information, ensuring broader applicability and robustness.

To meet these requirements, recent ASD approaches have moved beyond traditional reconstruction-based methods, such as autoencoders, toward embedding-based strategies.[5] These methods use powerful pre-trained audio models to extract high-level features from sound data, followed by similarity-based techniques such as K-Nearest Neighbors (KNN) or distance-based scoring for anomaly detection[6]. Embedding-based approaches have shown strong generalization ability, especially under domain shifts or unseen machine types, and offer a practical balance between performance and simplicity. In this work, we adopt such a strategy, leveraging multiple pre-trained embedding models and fusing their outputs to improve detection accuracy and robustness.

2. METHOD

In this section, we describe our proposed approach for first-shot unsupervised anomalous sound detection based on multiple pre-trained audio embeddings and K-Nearest Neighbors (KNN) anomaly scoring, followed by rank-based score fusion.

We employ a diverse set of pre-trained audio models to extract feature embeddings from machine sound recordings. These models cover different domains, including general audio classification, speech, and music, aiming to capture complementary acoustic characteristics. Specifically, we utilize the following models: BEATs[7], EAT[8], M2D[9], dasheng base, dasheng 06B, dasheng 12B[10], and MuQ[11]. Each model processes the input audio and outputs fixed-dimensional embeddings that represent the sound's high-level features.

For each embedding type, we build a reference feature set from the normal training data. During inference, anomaly scores are computed for each test sample by calculating its distance to its K nearest neighbors in the normal reference set within the embedding space. Typically, Euclidean or cosine distance metrics are used. The rationale is that anomalous samples will lie farther away from the cluster of normal embeddings, yielding higher anomaly scores.

To leverage the complementary strengths of different embedding models, we perform late fusion of their anomaly scores. Instead of directly averaging raw scores—which may be on

different scales—we convert scores into rankings for each model, reflecting the relative anomaly severity of samples. The final anomaly score for each test sample is obtained by aggregating these ranks using a weighted or unweighted combination strategy. This rank-based fusion enhances robustness and mitigates the effect of inconsistent score distributions across models without requiring additional training or hyperparameter tuning.

3. EXPERIMENT

We evaluated the performance of each pre-trained model using the official DCASE evaluation metrics, including AUC, pAUC, and average precision (AP), to ensure fair and consistent comparison across systems. The results indicate that models pre-trained on AudioSet—such as EAT, M2D, and MuQ—tend to outperform those trained on other domains across most machine types. We believe this advantage is not only due to the diversity of acoustic events in AudioSet, but more importantly because AudioSet contains some machine-related sounds that partially overlap with the target domain of DCASE Task 2. This domain overlap may help such models learn representations that are inherently more suitable for anomalous machine sound detection. In contrast, models pre-trained on speech or music datasets, such as Dasheng-06B, Dasheng-12B, and BEATs, generally show inferior overall performance, despite often having significantly larger model sizes. This is likely due to the domain gap between their training data and the target industrial sound domain. Nevertheless, we observed complementary behavior in some cases. For example, Dasheng-12B achieved the best results on specific machine types where EAT performed poorly, suggesting that speech-based models can still capture useful patterns under certain conditions. Interestingly, this implies that fine-tuning with section-level attribute information may be especially effective for speech or music models, as it could help reduce the domain mismatch and enhance their ability to detect anomalies in unfamiliar sound types.

Due to time constraints, we did not apply attribute-aware fine-tuning to all models. We only conducted full fine-tuning with section-level attributes on the BEATs model. While this helped improve domain-specific performance in some scenarios, the overall gain in the official pAUC score was limited. Therefore, we chose not to include those results in the final submission, nor to extend the fine-tuning process to other models. This decision was made as a practical trade-off durin during the challenge period, but we acknowledge it as a limitation of our system.

We submitted three systems in total. One was the official baseline_MSE to remain aligned with the DCASE baseline. The second was a standalone EAT model, which showed the best overall performance among all individual models. The third was an ensemble of EAT and M2D, using weighted averaging. Although this ensemble did not surpass EAT, it significantly outperformed M2D alone, indicating that the ensemble was effective in leveraging their complementarity. We also considered integrating all the evaluated pre-trained models, but this approach introduced substantial computational and memory overhead while failing to exceed the performance of the best single model. Notably, Dasheng-12B showed strong complementarity with EAT, achieving top results on machine types where EAT struggled. Despite this, the large-scale ensemble was not submitted due to practicality. The submitted system performances can be referenced directly from the results presented in Table 1

4. 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] 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.

[3] 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.

[4] 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 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring,” In arXiv e-prints: 2406.07250, 2024.

[5] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-Shot Anomaly Sound Detection for Machine Condition Monitoring: A Domain Generalization Baseline", in *2023 31st European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland: IEEE, Sep. 2023, pp. 191–195.

[6] A. Jiang, X. Zheng, and Y. Qiu, "Thuee System for First-Shot Unsupervised Anomalous Sound Detection", DCASE2024 Challenge, Tech. Rep., June 2024.

[7] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, “Beats: Audio pre-training with acoustic tokenizers,” arXiv preprint arXiv:2212.09058, 2022.

[8] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, “Eat: Self-supervised pre-training with efficient audio transformer,” arXiv preprint arXiv:2401.03497, 2024.

[9] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, “Masked modeling duo: Towards a universal audio pre-training framework,” Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2391–2406, 2024.

[10] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification", in *Interspeech 2024*, ISCA, Sep. 2024, pp. 547–551.

[11] H. Zhu *et al.*, "MuQ: Self-Supervised Music Representation Learning with Mel Residual Vector Quantization," In arXiv e-prints: 2501.01108, 2025.

		Baseline MSE	Baseline MAHALA	BEATs	EAT	M2D	Dasheng base	Dasheng 06B	Dasheng 12B	MuQ
ToyCar	AUC(Source)	71.05%	73.17%	75.76%	**83.4%**	71.08%	73.38%	75.5%	75.38%	76.74%
	AUC(Target)	53.52%	50.91%	58.98%	**59.74%**	58.56%	58.5%	56.46%	47.98%	52.66%
	pAUC	49.7%	49.05%	52.11%	**55.63%**	51.95%	51.68%	53.15%	51.47%	53.05%
ToyTrain	AUC(Source)	61.76%	50.87%	85.04%	79.74%	80.72%	77.08%	76.88%	75.58%	**85.56%**
	AUC(Target)	56.46%	46.15%	**69.8%**	66.6%	67.06%	54.1%	52.76%	45.98%	56.94%
	pAUC	50.19%	48.32%	53.84%	**55.1%**	52.05%	48.79%	49.68%	49.36%	52%
Bearing	AUC(Source)	66.53%	63.63%	60.56%	**75.14%**	64.12%	70.38%	66.8%	67.18%	58.26%
	AUC(Target)	53.15%	59.03%	48.02%	**61.28%**	57.58%	56.86%	57.1%	56.28%	51.8%
	pAUC	61.12%	**61.86%**	57.56%	61.05%	60.58%	61.37%	60.53%	61.36%	56.42%
Fan	AUC(Source)	70.96%	77.99%	61.2%	57.82%	65.83%	63.06%	57.04%	**66.54%**	59.46%
	AUC(Target)	38.75%	38.56%	41.2%	47.72%	45.58%	49.64%	**51.16%**	46.96%	46.87%
	pAUC	49.46%	50.82%	48.79%	50.37%	**51.05%**	50.53%	50.58%	49.84%	48.42%
Gearbox	AUC(Source)	64.8%	**73.26%**	62.5%	64.04%	64.7%	66.54%	66.68%	67.9%	64.4%
	AUC(Target)	50.49%	51.61%	51.24%	49.76	51.64%	53.02%	52.08%	**53.52%**	52.32%
	pAUC	52.49%	55.07%	55.05%	52.63%	52.53%	54.32%	52.53%	**55.63%**	53.21%
Slider	AUC(Source)	70.1%	73.79%	75.44%	**77.04%**	75.36%	75.1%	77.56%	76.76%	73.4%
	AUC(Target)	48.77%	50.27%	52.9%	**61.3%**	52.77%	53.56%	52.72%	53.56%	51.4%
	pAUC	52.32%	53.61%	51.21%	**54.16%**	52.58%	51.84%	52.37%	52.47%	50.8%
valve	AUC(Source)	63.53%	56.22%	72.8%	**76.32%**	74.3%	54.36%	52.32%	58.64%	71.11%
	AUC(Target)	67.18%	61.0%	76.7%	**86.74%**	77.96%	64.22%	65.2%	71.98%	57.86%
	pAUC	57.35%	52.53%	51.53%	**74.84%**	67.53%	54.47%	52.21%	58.37%	60.95%

Table 1 Detection performance (in %) of different pre-trained models on each machine type. The evaluation metrics include Area Under the Curve (AUC) for both source and target domains, as well as partial AUC (pAUC) at 10% false positive rate. Models include the official baselines (MSE and MAHALA), models pre-trained on AudioSet (EAT, M2D, MuQ), speech-based models (Dasheng-base, Dasheng-06B, Dasheng-12B), and a music-based model (BEATs). Bold values (if any) indicate the highest performance per row.