Multi-System Framework with Heterogeneous Feature Extractors for DCASE 2026 Task 2

Technical Report

Wen Zeng[1], Jianxia Liao[2,3], Ting Wu[2,3], Zhaoli Yan[1], Fusheng Sui[2,3]*

[1]Beijing University of Chemical Technology, Beijing, China
[2]Institue of Acoustics, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
2824095672@qq.com
{liaojianxia, wuting, sui}@mail.ioa.ac.cn
yanzl@mail.buct.edu.cn

Abstract

This report describes our submission to DCASE 2026 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. We propose a multi-system framework built upon heterogeneous feature extractors, including a BEATs-based audio encoder, a convolutional neural network (CNN) feature extractor adapted from the DCASE2023 FKIE system, and a ResNet-based audio encoder. Stereo information is exploited through both independent channel modeling and joint two-channel representation learning. Domain-wise Local Density Normalization (DLDN) is employed in the BEATs- and FKIE-based systems for anomaly score computation. Finally, anomaly scores from different channels and feature extractors are fused using optimized weights determined on the development dataset. Experimental results demonstrate that exploiting stereo-channel information and combining heterogeneous feature representations effectively improves robustness and detection performance under domain-shift conditions.

Index Terms— Anomalous Sound Detection, DCASE2026, BEATs, Domain Generalization

1. INTRODUCTION

This report presents our solution for the DCASE 2026 Challenge Task 2, which focuses on noise-aware unsupervised anomalous sound detection for machine condition monitoring [1]. Our experiments are conducted on ToyADMOS2 [2] and MIMII DG [3], two benchmark datasets designed for domain generalization, and are compared with the first-shot anomaly detection baseline [4].

DCASE 2026 Task 2 addresses first-shot unsupervised anomalous sound detection under domain generalization, aiming to detect anomalies from unseen machine types without using evaluation labels during training.

Recent studies have shown that large-scale pretrained audio foundation models, such as BEATs, learn highly transferable audio representations and achieve strong anomaly detection performance. Meanwhile, task-specific convolutional models developed for previous DCASE challenges remain competitive for machine condition monitoring.

Motivated by these advances, we propose a multi-system framework that combines heterogeneous feature extractors and score fusion. Our submitted systems employ three feature extractors: a BEATs-based encoder, a CNN adapted from the DCASE2023 FKIE system [5], and a ResNet-based encoder. To utilize the stereo recordings in DCASE2026, the two audio channels are processed either independently or jointly, depending on the feature extractor. Different channel-level and model-level score fusion strategies are adopted for different system configurations.The main contributions of this work are summarized as follows:

- Exploitation of stereo audio information through both independent channel modeling and joint two-channel representation learning;
- Application of Domain-wise Local Density Normalization (DLDN) [6] for anomaly score computation;
- Multi-level score fusion across channels and feature extraction models.

* Corresponding author: yanzl@mail.buct.edu.cn

2. SYSTEM 1 DESCRIPTION

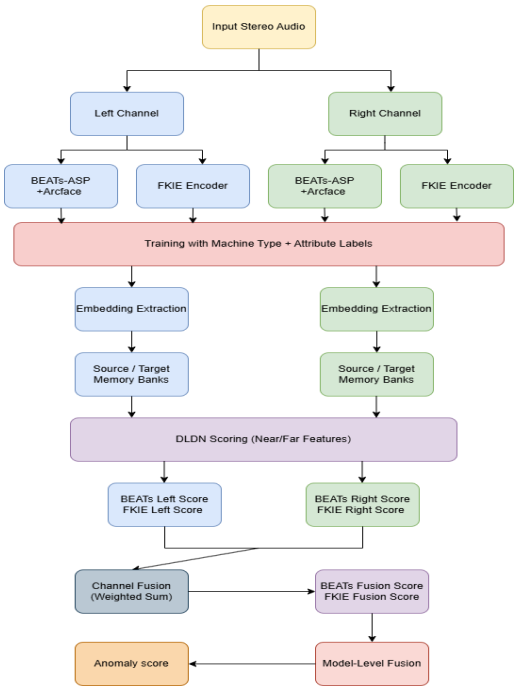

Figure 1: Anomaly detection pipeline of System 1.

2.1. Overall Framework

The proposed system consists of four independent branches:

- BEATs model – Left channel
- BEATs model – Right channel
- FKIE model – Left channel
- FKIE model – Right channel

In the stereo recordings, the left and right channels correspond to the near-field and far-field microphones, respectively.

For each branch, audio embeddings are extracted and stored separately in source-domain and target-domain memory banks. During inference, anomaly scores are computed using Domain-wise Local Density Normalization (DLDN).

The score fusion process is conducted in two stages. First, anomaly scores from the left and right channels are fused within each model. Then, the fused scores obtained from the BEATs and FKIE models are combined to generate the final anomaly score.

2.2. BEATs-based Feature Extractor

We employ the pretrained BEATs_iter3_plus_AS2M_finetuned_on_AS2M_cpt2 model as the backbone and further fine-tune it on the DCASE2026 Task 2 training set. Instead of directly using frozen BEATs representations, an Attentive Statistics Pooling (ASP) layer and an ArcFace [7] classification head are introduced to learn more discriminative embeddings for anomalous sound detection.

Input audio is first encoded by BEATs, and the output of the 11th Transformer layer is extracted as the frame-level representation. The ASP layer aggregates temporal information by modeling the mean and standard deviation of frame-level features, producing an utterance-level embedding. During training, the embeddings are optimized with ArcFace using labels constructed from machine types and attribute information, enhancing inter-class separability and embedding compactness [8,9].

After training, the ArcFace head is removed, while the BEATs backbone and ASP layer are retained for feature extraction. The extracted embeddings are stored in separate source- and target-domain memory banks and used for DLDN-based anomaly score computation.

2.3. FKIE Feature Extractor

To provide complementary feature representations, we adopt the CNN-based feature extractor proposed in the DCASE2023 FKIE submission.

The feature extractor consists of two parallel branches:

An FFT-based feature extraction branch;

A residual convolutional branch operating on Mel-spectrograms.

The outputs of the two branches are concatenated to form the final audio embedding.

While the original feature extraction architecture is preserved, the adaptive projection (AdaProj) scoring mechanism employed in the original FKIE system is removed. Instead, anomaly scores are uniformly computed using the DLDN framework.

2.4. Domain-wise Local Density Normalization

For each domain, a dedicated memory bank is constructed using the embeddings extracted from the training samples.

Given a test embedding, its anomaly score is calculated by comparing it with neighboring embeddings stored in the corresponding memory bank.

DLDN estimates local density statistics separately for the source and target domains. The final anomaly score is computed as

$$score(y) = \min\left(score_{source}(y), score_{target}(y)\right) \quad (1)$$

where both source-domain and target-domain scores are normalized according to their local neighborhood densities.

This strategy improves robustness against domain shifts and alleviates score-scale mismatches between different domains.

2.5. Dual-Channel Processing

Most previous DCASE Task 2 datasets consisted of monaural recordings. In contrast, the DCASE2026 dataset introduces stereo audio, which provides additional spatial acoustic information.

Instead of converting stereo recordings into monaural signals, the proposed system processes the left and right channels independently.

For each model, separate memory banks are constructed for the left and right channels. The corresponding anomaly scores are fused through weighted averaging:

$$S_{model} = \alpha * S_{left} + (1 - \alpha) * S_{right} \quad (2)$$

where the weighting coefficient α is determined through grid search on the development dataset.

This fusion strategy enables the system to effectively exploit complementary information captured by different channels.

2.6. Model Fusion

The final anomaly score is obtained by combining the outputs of the BEATs and FKIE models:

By integrating heterogeneous feature representations, the proposed model fusion strategy further improves the robustness and generalization capability of the overall system.

3. SYSTEM 2 DESCRIPTION

3.1. Pre-processing

For each channel, the raw audio is first normalized to zero mean and unit variance and resampled to 16 kHz. A power spectrogram is computed using a 1024-point FFT and a hop length of 512 samples. The spectrogram is then mapped to 128 Mel filter banks and converted to the logarithmic scale.

To obtain a fixed-size network input, a temporal segment containing 256 consecutive frames is used. During training, the starting position of the segment is randomly selected, serving as a form of data augmentation. During testing, the anomaly score for each sample is generated by averaging the scores of multiple 256-frame windows with a hop size of one frame.

3.2. ResNet-based Feature Extractor

We adopt a residual convolutional neural network (ResNet) [10] following the architecture described in [11]. According to the DCASE 2026 Task 2 setup, each recording contains a near-channel signal (Channel 1) and a far-channel signal (Channel 2). Log-Mel spectrograms are extracted independently from both channels and concatenated along the channel dimension to form a two-channel input for the ResNet encoder.

Exponential moving average (EMA) of model parameters is used during training to obtain a more stable model for inference.

3.3. Classification

To learn discriminative machine representations, we train or finetune our models using a supervised auxiliary classification task. For machine types with attribute annotations, samples are classified according to joint machine-domain-attribute classes. For machine types without attribute information, machine-domain labels are used.

ArcFace loss [7] is adopted to enhance feature discriminability by encouraging intra-class compactness while increasing inter-class separability in the embedding space.

3.4. Data Augmentation

The data augmentation technique mixup [12] is applied to improve feature discriminability and generalization performance. Mixup generates new training samples by linearly interpolating spectrograms and the corresponding labels from different classes. The model is trained to predict both the constituent classes and their corresponding mixing coefficients.

To improve robustness, random time and frequency masking is applied to the input spectrograms during training.

4. EXPERIMENTAL SETUP

4.1. Dataset

Experiments were conducted using the development dataset provided for DCASE2026 Challenge Task 2.

The dataset contains normal machine sounds collected under various operating conditions from both source and target domains. No evaluation labels were used during the training process.

4.2. Evaluation Metrics

Following the official evaluation protocol of DCASE2026 Task 2, system performance is evaluated using the following metrics:

Area Under the Receiver Operating Characteristic Curve (AUC);

Partial Area Under the ROC Curve (pAUC);

Official Challenge Score.

4.3. Submitted Systems

Table 1: Description of the submitted systems.

System	Model	Normalization
System 1	BEATs + FKIE	DLDN
System 2	ResNet	-
System 3	System1 + System2	-
System 4	FKIE	DLDN

We submitted four systems to the challenge, as summarized in Table 1. System 1, the primary submission, ensembles BEATs and FKIE feature extractors with DLDN-based anomaly scoring. System 2 employs a ResNet-based feature extractor trained with an auxiliary classification task. System 3 fuses the normalized anomaly scores of Systems 1 and 2 through z-score normalization followed by score averaging. System 4 uses the FKIE feature extractor with DLDN-based anomaly scoring.

For BEATs and FKIE, embeddings are extracted independently from the left and right audio channels, whereas the ResNet model processes stacked two-channel log-Mel spectrograms to learn joint stereo representations. In Systems 1 and 4, both channel-level and model-level fusion weights are determined using the development dataset.

5. RESULTS

Table 2: Official scores of the baseline and submitted systems on the development set.

System	Official Score
Baseline - MSE	57.06
Baseline - MAHALA	57.71
System 1	64.67
System 2	61.55
System 3	65.86
System 4	62.96

Table 3: Machine-wise AUC and pAUC results of the submitted systems on the development set.

Machine	System 1	System 2	System 3	System 4
Bearing Emu				
AUC Source	56.68	68.42	59.64	51.64
AUC Target	62.08	64.04	67.80	56.20
pAUC	56.79	58.84	61.53	54.21
Fan				
AUC Source	55.92	86.64	79.22	63.68
AUC Target	62.44	73.98	63.20	48.84
pAUC	52.26	65.37	63.71	53.63
Gearbox Emu				
AUC Source	73.24	81.34	73.09	72.80
AUC Target	73.44	70.58	79.24	76.40
pAUC	60.11	61.47	56.96	62.32
Slider Emu				
AUC Source	61.68	72.26	71.05	65.84
AUC Target	55.36	47.54	56.96	52.92
pAUC	52.63	51.63	52.37	50.74
Toycar_r2				
AUC Source	71.40	60.40	55.77	68.44
AUC Target	87.04	76.58	90.72	82.36
pAUC	60.74	58.89	65.12	58.26
Toycar Emu				
AUC Source	59.12	39.92	53.73	58.76
AUC Target	86.48	90.46	89.64	81.32
pAUC	62.53	48.37	53.24	62.42
Valve Emu				
AUC Source	99.80	56.96	94.00	99.40
AUC Target	70.48	61.06	70.08	70.04
pAUC	80.05	66.78	63.64	77.79

Table 2 presents the official scores of the baseline and submitted systems on the development set. Table 3 reports the machine-wise AUC and pAUC results on the development set. The experimental results lead to the following observations:

- All submitted systems outperform the official baseline.
- Dual-channel processing consistently outperforms single-channel processing;
- The BEATs-based model provides highly discriminative and expressive semantic audio representations;
- The FKIE-based model captures machine-specific acoustic characteristics and serves as a complementary

representation to BEATs;

- The proposed score fusion strategy further improves the overall performance and robustness of the system.
- The ResNet-based model also improves over the baseline, suggesting that the joint two-channel input helps exploit both channel-specific and inter-channel information.
- By averaging the normalized scores of System 1 and System 2, System 3 outperforms both individual systems, indicating the complementarity of the two systems.

6. CONCLUSION

This report presented a multi-system framework for DCASE2026 Task 2 based on heterogeneous feature extractors, including BEATs-, FKIE-, and ResNet-based audio encoders. For the BEATs- and FKIE-based systems, Domain-wise Local Density Normalization (DLDN) was adopted for anomaly scoring to reduce domain-wise score discrepancies. Stereo information was exploited through two complementary strategies: independent channel modeling with score fusion for the BEATs- and FKIE-based systems, and joint two-channel representation learning for the ResNet-based system. By combining channel-level and model-level score fusion, the proposed framework effectively leverages complementary information from both stereo channels and heterogeneous feature representations. Experimental results on the development set demonstrate its robustness under domain generalization and its effectiveness for first-shot unsupervised anomalous sound detection.

7. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and Discussion on DCASE 2026 Challenge Task 2: Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," *arXiv e-prints: 2606.01578*, 2026.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection under Domain Shift Conditions," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021, pp. 1-5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task," in *Proc. 7th Detection and Classification of Acoustic Scenes and Events Workshop (DCASE2022)*, 2022.

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

[5] K. Wilkinghoff, “Fraunhofer FKIE submission for Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring,” DCASE2023 Challenge Technical Report, 2023.

[6] P. Saengthong and T. Shinozaki, “GenRep for first-shot unsupervised anomalous sound detection of DCASE 2025 challenge,” in Proc. Detect. Classif. Acoust. Scenes Events (DCASE), 2025.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019.

[8] L. Wang, “Pre-trained model enhanced anomalous sound detection system for DCASE2025 Task2,” in Proc. Detect. Classif. Acoust. Scenes Events (DCASE), 2025.

[9] J. Yang, “A two stage fusion anomaly detection approach for Task2,” in Proc. Detect. Classif. Acoust. Scenes Events (DCASE), 2025.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.

[11] P. Primus, “Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers,” DCASE2020 Challenge Technical Report, 2020.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019.