AITHU SUBMISSION FOR DCASE 2026 TASK 2: ROBUST MACHINE-WISE SCORING WITH BEATS REPRESENTATIONS

Technical Report

Anbai Jiang[1], Xinhu Zheng[2], Wenrui Liang[1], Shuwei Zhang[1], Tianyu Liu[1]*
Jia Liu[1,3], Pingyi Fan[1], Wei-Qiang Zhang[1], Cheng Lu[4], Xie Chen[2], Yanmin Qian[2]

[1] Tsinghua University, Beijing, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Huakong AI Plus Company Limited, Beijing, China
[4] North China Electric Power University, Beijing, China
*Email: jab22@mails.tsinghua.edu.cn

ABSTRACT

This report describes the AITHU submission to DCASE 2026 Challenge Task 2 on noise-aware anomalous sound detection. The central component of our system is a score-level fusion that exploits the complementarity of multiple independently trained systems. All systems are built on BEATs-based audio representations with distance-based, primarily Mahalanobis, anomaly scoring, and BEATs is the only external pre-trained model, so the diversity required for an effective fusion comes from heterogeneous scoring branches, training seeds, and generative augmentation rather than from additional backbones. The branches are combined by Bayesian-optimized score-level selection. Four systems are submitted, including one single scoring system and three ensemble systems. The best submitted system achieves a development-set harmonic mean of 68.20%.

Index Terms— Anomalous Sound Detection, Score Fusion, Complementarity, Ensembling

1. INTRODUCTION

The DCASE 2026 Challenge Task 2 [1] studies noise-aware anomalous sound detection for machine condition monitoring, continuing the first-shot ASD task series of recent years [2]. The task builds on a line of machine-sound anomaly detection datasets, including the original MIMII [3] and ToyADMOS [4] corpora and their domain-shift successors MIMII DUE [5], ToyADMOS2 [6], and MIMII DG [7], together with the first-shot domain-generalization baseline for machine condition monitoring [8]. The evaluation setting emphasizes robust scoring under machine-dependent acoustic variation and channel-dependent noise.

Recent advances in anomalous sound detection have been largely driven by self-supervised learning and audio foundation models. Large-scale pre-trained models such as BEATs [9], EAT [10], and other SSL-based encoders have demonstrated strong robustness under domain shift and limited-data conditions. Systematic studies further show that SSL-derived representations consistently outperform conventional handcrafted features across diverse machine categories [11]. More recently, industrial foundation models have extended representation learning by leveraging multimodal and large-scale industrial signals, providing richer semantic information for machine condition modeling [12].

Building upon these representations, extensive efforts have focused on adapting foundation models to ASD tasks. LoRA-based fine-tuning has been shown to improve adaptation efficiency while preserving model capacity [13]. AnoPatch enhances anomaly detection through patch-level consistency learning [14], whereas adaptive prototype learning improves robustness when machine attributes are partially available or missing [15]. In parallel, data-centric approaches have been explored, in particular generative augmentation: diffusion-based generation of rare samples has proven effective for ASD [16, 17], and RefGen introduces reference-guided synthetic data generation to alleviate data scarcity and domain mismatch under diverse operating conditions [18].

Beyond improving individual models, recent top-performing DCASE challenge systems have demonstrated the effectiveness of exploiting complementary information from multiple pre-trained models. Several leading submissions achieved strong performance through the fusion of heterogeneous foundation models, adaptation strategies, and training paradigms [19, 20, 21, 22]. These systems combine representations derived from different self-supervised and foundation models, allowing them to capture complementary acoustic characteristics and improve robustness under domain shifts. These findings suggest that model diversity plays a crucial role in anomalous sound detection and motivate further investigation into effective strategies for exploiting complementary information across foundation models.

The AITHU submission is built around BEATs-based audio representations [9] and machine-wise scoring. In contrast to our previous submissions [21, 22], which fused several distinct SSL backbones to obtain diversity, the present submission deliberately fixes the backbone to BEATs and instead extracts complementarity from the scoring and fusion stage: heterogeneous distance-based detectors, multiple independently trained systems, and a score-level fusion. Rather than presenting a detailed recipe for each branch, we describe the four submissions at the system level and focus on how complementary scoring branches are combined. Each system maps machine sounds into an audio representation space and produces anomaly score and binary decision outputs.

2. SYSTEM OVERVIEW

All systems operate on 16 kHz audio and use log-mel spectral representations as input to a BEATs-based feature extractor [9]. BEATs is the only external pre-trained model in the submission, so the four systems share the same backbone but differ in how that backbone is fine-tuned, augmented, scored, and fused. The pipeline has two stages—representation and scoring, and score-level fusion. To keep the report concise, implementation details such as branch-specific checkpoints and search ranges are summarized only at a high level.

2.1. Representation and Scoring

The backbone is adapted to machine sounds by fine-tuning it as an attribute classifier, where the label combines machine type, section, and the available operating-condition attributes, and an utterance-level embedding is obtained by attentive statistical pooling. For each machine type, the embeddings of the normal training samples define a reference score space, with source-domain and target-domain samples kept in separate memory banks so that the scarcity of target-domain data does not bias the source-domain statistics. Test recordings are scored using distance-based evidence against this reference space. The primary detector is a Mahalanobis detector, which scores a clip against a normal distribution that accounts for the covariance of the normal class; a nearest-neighbor variant, which scores a clip by its distance to the closest stored normal embedding, is also available. The single-system submission uses one such scoring path, while the ensemble systems combine multiple paths. Complementarity then arises mainly from the diverse fine-tuning configurations of the branches: different classification objectives, fine-tuning strategies, and training seeds. Deatils are described in Section 3, since branches trained under different recipes make decorrelated errors even when they share the Mahalanobis scoring rule.

2.2. Fusion and Complementarity

Score fusion is performed by linear combination of the anomaly scores. Following our previous mega-ensembling practice [21], the fusion adopts a two-step hierarchy: scores of homogeneous branches that differ only in the fine-tuning seed are first aggregated, and the resulting per-configuration scores are then combined across heterogeneous configurations and independently developed sub-systems. The combination weights are selected on the development set using Bayesian optimization in the provided ensemble runner, which searches the weight space at finer granularity and lower cost than grid search and thus scales to many branches. Evaluation labels are not used for weight selection. The motivation for fusion is complementarity rather than redundancy: branches that disagree on which recordings look anomalous—because they use different detectors, different seeds, or are trained independently—tend to make decorrelated errors, so their combination is more robust than any single branch. As shown in Section 4, no single system is best on every machine type, and the fusion exploits exactly this machine-dependent complementarity by letting the most reliable branches dominate where they are strong.

3. SUBMITTED SYSTEMS

We submit four systems that share the BEATs backbone but are deliberately built differently, each adopting a distinct source of diversity rather than being a scaled version of one recipe.

- **System 1** is a single-branch baseline—one fine-tuned BEATs backbone scored by a Mahalanobis detector, with no fusion—and reaches an overall harmonic mean of 64.04%.
- **System 2** is a single-backbone ensemble of ten BEATs branches that differ in classification objective (ArcFace, softmax, sub-center), fine-tuning strategy (full, partial, parameter-efficient), and dual-tokenizer training variants, and reaches 65.47%.
- **System 3** fuses three independently developed sub-systems—a large BEATs ensemble, a Mahalanobis scoring system, and a patch-pooling fusion system—whose independent recipes make their errors the most decorrelated; it is the best submitted system, reaching 68.20%.
- **System 4** is an ensemble of seven Mahalanobis branches, all trained on a generatively augmented set that synthesizes rare working conditions [16, 17, 18] to close the source–target domain gap, and reaches 67.70%.

Because the systems are built differently, they are also complementary, and no single entry is best on every machine type (Section 4). System 3 maximizes the overall mean through heterogeneous independent sub-systems, whereas System 4, through generative augmentation, is strongest on the hardest machines; their fusion exploits exactly this complementarity.

Table 1: Development-set harmonic mean (%) of the AITHU submitted systems. The best result per row is in bold.

Machine	S1	S2	S3	S4
bearingEmu	58.80	60.65	**63.44**	61.08
fan	**67.59**	**67.59**	66.85	66.43
gearboxEmu	72.06	75.91	75.87	**77.29**
sliderEmu	55.42	57.74	**58.97**	57.44
ToyCar	66.26	69.73	**75.05**	70.13
ToyCarEmu	67.90	69.51	70.38	**81.08**
valveEmu	63.41	60.89	**70.17**	66.54
Overall	64.04	65.47	**68.20**	67.70

4. EXPERIMENT RESULTS

Performance is measured using source-domain AUC, target-domain AUC, partial AUC, and their harmonic mean. Table 1 reports the harmonic mean of the four submitted systems on the development set. Full metric values are provided in the corresponding meta files.

The results provide direct evidence of complementarity. No single system dominates across all machine types: System 3 is best on four of the seven machines and on the overall mean, yet System 4 is clearly stronger on gearboxEmu and ToyCarEmu (where it reaches 81.08%), and the single-branch System 1 already matches the best on fan. Because different branches and independently trained systems are strongest on different machines, combining them is beneficial rather than redundant. Accordingly, the larger fusion in System 3 attains the best overall harmonic mean of 68.20%, improving over the single-branch reference (System 1, 64.04%) by exploiting this machine-dependent complementarity.

5. CONCLUSION

This report presented the AITHU submission to DCASE 2026 Task 2. The core of the system is a score-level fusion that exploits the complementarity of multiple independently trained systems built on a shared BEATs backbone. With BEATs fixed as the only external backbone, the submission obtains its strength from complementary scoring branches—Mahalanobis-based branches trained under diverse fine-tuning recipes, generative augmentation, and independently developed sub-systems—combined through Bayesian-optimized score-level fusion. The per-machine results confirm that the branches are complementary, and the best submitted system achieves a development-set harmonic mean of 68.20%.

6. 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," *In arXiv e-prints: 2606.01578*, 2026.

[2] 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," *In arXiv e-prints: 2506.10097*, 2025.

[3] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, November 2019, pp. 209–213.

[4] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, November 2019, pp. 308–312.

[5] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25, 2021.

[6] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[7] 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 *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[8] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, 2023.

[9] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 5178–5193.

[10] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 3807–3815.

[11] B. Han, A. Jiang, X. Zheng, W.-Q. Zhang, J. Liu, P. Fan, and Y. Qian, "Exploring self-supervised audio models for generalized anomalous sound detection," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[12] P. Fan, A. Jiang, S. Zhang, X. Zheng, Z. Lv, B. Han, W. Liang, J. Li, W.-Q. Zhang, Y. Qian, X. Chen, and J. Liu, "Fisher: A foundation model for multimodal industrial signal comprehensive representation," *IEEE Transactions on Industrial Informatics*, pp. 1–12, 2026.

[13] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 969–974.

[14] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Interspeech 2024*, 2024, pp. 107–111.

[15] A. Jiang, X. Zheng, B. Han, Y. Qiu, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Adaptive prototype learning for anomalous sound detection with partially known attributes," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[16] H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1271–1275.

[17] J. Yin, Y. Gao, W. Zhang, T. Wang, and M. Zhang, "Diffusion augmentation sub-center modeling for unsupervised anomalous sound detection with partially attribute-unavailable conditions," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[18] W. Liang, Y. Qiu, A. Jiang, B. Han, T. Liu, X. Zheng, P. Fan, C. Lu, J. Liu, and W.-Q. Zhang, "Refgen: Reference-guided synthetic data generation for anomalous sound detection," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 15 877–15 881.

[19] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[20] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "THUEE system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[21] A. Jiang, W. Liang, S. Feng, Y. Qiu, Y. Zhao, J. Li, P. Fan, W.-Q. Zhang, C. Lu, X. Chen, Y. Qian, and J. Liu, "Thuee system for dcase 2025 anomalous sound detection challenge," DCASE2025 Challenge, Tech. Rep., June 2025.

[22] X. Zheng, A. Jiang, B. Han, S. Zhang, W.-Q. Zhang, X. Chen, C. Lu, P. Fan, J. Liu, and Y. Qian, "Sjtu-aithu system for dcase 2025 anomalous sound detection challenge," DCASE2025 Challenge, Tech. Rep., June 2025.