THE WISTLAB SYSTEM FOR DCASE 2026 TASK 2: FINE-GRAINED SCORING FOR NOISE-AWARE MACHINE SOUND ANOMALY DETECTION

Technical Report

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

[1] Tsinghua University, Beijing, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] North China Electric Power University, Beijing, China
*Email: {jab22,zhangsw25}@mails.tsinghua.edu.cn

ABSTRACT

This report outlines the WISTLAB submission to DCASE 2026 Challenge Task 2 on noise-aware machine sound anomaly detection. The system uses task-adapted BEATs representations together with sub-band scoring, fine-grained localized scoring, and local-density normalization. The sub-band component extends AdaBEAM with a learned attentive-pooling view, while the fine-grained component retains local time-frequency evidence for normal-reference scoring. We submit one single scoring system and three score-level fusion systems, with a best development-set harmonic mean of 67.20%.

Index Terms— DCASE Challenge, anomalous sound detection, localized scoring, local-density normalization, score fusion

1. INTRODUCTION

The DCASE 2026 Challenge Task 2 [1] studies noise-aware anomalous sound detection (ASD) for machine condition monitoring. The task builds on prior machine-sound datasets and domain-generalization baselines, including ToyADMOS2 [2], MIMII DG [3], and first-shot anomaly detection for machine condition monitoring [4]. The 2026 setup introduces paired near-field and far-field recordings. Systems must identify unknown faults from normal-only training data while remaining robust to machine-dependent source–target shifts and channel-dependent background noise. These characteristics make the task particularly challenging, because channel noise and domain shifts may affect broad acoustic statistics while abnormal evidence can remain localized in time or frequency.

Current state-of-the-art (SOTA) embedding-based ASD systems represent an audio recording as an utterance-level embedding and compare it with normal training samples. Although this design is efficient, a fault may occupy only a short temporal interval or a narrow spectral region and can therefore be weakened by global pooling. The WISTLAB submission instead retains local temporal-spectral evidence until the anomaly-scoring stage. Structured representations are compared with aligned normal references, calibrated using the neighborhood structure of normal data, and then aggregated into file-level scores.

This perspective is related to patch-memory anomaly detection [5], machine-sound patch modeling [6], and density-aware normal-reference scoring [7]. A practical challenge is that no single granularity is uniformly reliable. Coarse descriptors are stable for sustained or broadband deviations, whereas fine descriptors are more sensitive to short and localized events. We therefore treat scores computed at different temporal-spectral granularities as complementary views and combine their outputs only at the score level.

In implementation, this scoring framework is built on a task-adapted acoustic encoder. The encoder is adapted to machine sounds using classification-oriented objectives, and its structured outputs support anomaly scoring at multiple temporal-spectral granularities. The remainder of this report describes the representation interface and scoring components, defines the four submitted systems, and analyzes their development-set performance.

2. METHODS

2.1. Acoustic Representation Extraction

All branches use BEATs [8] as the external pre-trained acoustic encoder. Before anomaly scoring, the encoder is adapted to normal machine operating conditions through task-oriented proxy classification. The training variants use angular-margin objectives, following ArcFace [9], to make condition-dependent embedding clusters more compact. These choices follow the broader finding that task adaptation of self-supervised audio models improves generalized ASD [10, 11, 12, 13] and have also been used in recent DCASE systems [14, 15]. More broadly, this use of pre-trained representations is aligned with recent foundation representation learning for industrial signals, as exemplified by FISHER [16].

The fine-tuned encoders are then employed as fixed feature extractors. Input waveforms are converted to time-frequency features and passed through the encoder. Besides an utterance-level embedding, the interface exposes structured encoder outputs before complete temporal-spectral collapse. Consequently, the same backbone supports utterance-level descriptors, sub-band descriptors, and fine-grained localized descriptors under a shared scoring interface. Encoder parameters remain fixed during memory construction and anomaly scoring.

2.2. Localized Scoring Views

We organize anomaly scoring into multiple file-level views before score-level fusion, with two localized views serving as the main

Table 1: Development-set results (%) of the WISTLAB submitted systems.

Machine	Metric	System 1	System 2	System 3	System 4
bearingEmu	AUC_s	63.38	64.44	64.54	64.86
	AUC_t	68.30	69.14	68.78	65.82
	pAUC	61.26	60.95	60.84	60.32
	hmean	64.18	**64.67**	64.56	63.57
fan	AUC_s	75.24	75.58	75.42	75.08
	AUC_t	60.92	63.46	63.26	62.50
	pAUC	60.58	58.63	59.11	60.42
	hmean	64.92	65.15	65.24	**65.40**
gearboxEmu	AUC_s	75.26	78.00	77.80	80.94
	AUC_t	68.58	75.28	74.98	80.42
	pAUC	54.63	57.26	57.11	64.32
	hmean	64.97	68.86	68.65	**74.37**
sliderEmu	AUC_s	59.28	58.64	59.32	63.04
	AUC_t	66.68	67.58	67.52	69.74
	pAUC	53.32	53.05	53.21	51.32
	hmean	59.26	59.17	59.45	**60.38**
ToyCar	AUC_s	71.38	76.02	76.16	80.18
	AUC_t	81.42	84.24	83.96	82.70
	pAUC	66.42	66.21	66.05	67.11
	hmean	72.56	74.76	74.66	**76.02**
ToyCarEmu	AUC_s	60.30	54.30	54.00	60.36
	AUC_t	79.42	79.36	79.64	83.70
	pAUC	51.89	52.26	52.47	53.32
	hmean	61.92	59.82	59.84	**63.47**
valveEmu	AUC_s	75.28	77.58	77.46	77.20
	AUC_t	79.70	83.38	83.06	82.96
	pAUC	58.05	59.37	58.95	56.63
	hmean	69.67	**71.90**	71.58	70.31
Overall	AUC_s	68.59	69.22	69.24	**71.67**
	AUC_t	72.15	74.63	74.46	**75.41**
	pAUC	58.02	58.25	58.25	**59.06**
	hmean	65.10	65.89	65.86	**67.20**

AUC_s and AUC_t denote source- and target-domain AUC, respectively. Overall AUC_s, AUC_t, and pAUC are arithmetic averages over machine types; overall hmean follows the official harmonic-mean criterion.

structured scoring components. The first view performs sub-band scoring and is derived from AdaBEAM [17] and AnoPatch [6]. AdaBEAM provides complementary temporal-mean and temporal-max scoring streams, while AnoPatch motivates the use of learned attentive temporal summarization. We combine these ingredients to obtain a band-preserving score that retains temporal-spectral structure before file-level anomaly scoring, rather than collapsing the representation into a single global descriptor.

The second view performs fine-grained localized scoring on structured encoder outputs before complete temporal-spectral collapse. This branch is designed to preserve short-duration or spatially concentrated evidence that may be weakened by coarser pooling. It converts localized discrepancies into a calibrated file-level anomaly score through an aggregation procedure. The resulting score remains a branch-level output and is not merged with other views until the score-fusion stage.

This design also reduces the dependence on any single pooling rule. Sustained deviations, impulsive events, and narrow-band changes may lead to different score patterns, and their relative importance varies across machine types. We therefore keep the localized scoring views as separate branches and combine them only at the score level. This makes the final fusion depend on complementary anomaly evidence rather than on a single pooled representation.

This merged design keeps the scoring interface simple: coarse utterance-level descriptors, sub-band scores, and fine-grained local scores are treated as complementary file-level scoring views. The temporal-spectral granularity and aggregation parameters of the localized scoring views were optimized on the development set and fixed for all submitted systems. This provides a unified interface for combining complementary evidence across different temporal-spectral granularities at the score level.

This separation is especially useful in the noise-aware setting. Far-field recordings may introduce broad background changes, while machine faults may appear as short localized events or as

narrow-band spectral deviations. A single pooled representation tends to mix these effects before scoring, making it difficult to distinguish channel-related variation from abnormal machine evidence. By keeping the sub-band and localized scores separate until the final fusion stage, the system can preserve both coarse file-level robustness and fine temporal-spectral sensitivity. The fusion step then combines already calibrated evidence streams rather than raw intermediate representations.

2.3. Local-Density Calibration and Aggregation

Following prior local-density normalization (LDN) approaches [18], we calibrate nearest-neighbor distances using the local neighborhood scale of the matched normal reference. The LDN-calibrated distance is written in generic form as

$$\widetilde{d}(x) = \frac{d(x)}{\epsilon + \rho\big(m^*; \mathcal{M}\big)}, \tag{1}$$

where m^* is the matched reference and $\rho(\cdot)$ is the mean distance from that reference to its K nearest normal neighbors. The neighborhood size K is selected on the development set and then fixed for submitted scoring.

Calibrated local evidence must then be mapped to a file-level score. The submitted branches include both broad aggregation, which is stable for sustained deviations, and selective aggregation, which is more sensitive to concentrated high-score regions. These alternatives are kept as separate branches rather than combined inside the representation. This separation makes their complementarity directly accessible to score-level fusion.

3. SUBMITTED SYSTEMS

We submit four systems. They share the same BEATs representation family but use different scoring branches. For each fusion system, branch selection and fusion settings are determined on the development set and then fixed for submission, so that all evaluation clips are scored with the same predefined system configuration.

System 1	A single fine-grained localized scoring branch with local-density calibration.
System 2	A six-branch fusion combining the fine-grained score, the three-view sub-band extension, AdaBEAM, attentive-pooled KNN, and the temporal-mean and temporal-max band-matching streams.
System 3	A seven-branch fusion that augments System 2 with a second file-level aggregation of fine-grained localized evidence.
System 4	A compact fusion of AdaBEAM and a band-preserving KNN score.

Systems 2 and 3 test whether a broad collection of scoring granularities gives stable gains. System 4 instead selects a small complementary subset. The four submissions therefore compare a single fine-grained detector, two broad multi-view ensembles, and a compact fusion under the same development protocol.

4. EXPERIMENT RESULTS

The four systems are evaluated on the DCASE 2026 Task 2 development set. Following the official protocol, we report source- and target-domain AUC (AUC_s and AUC_t), partial AUC (pAUC) in the low false-positive-rate range, and their harmonic mean. No anomalous training recordings are used to construct the normal-reference memories. Development labels are used only for branch selection, fixed calibration settings, and fusion weights. Table 1 gives the complete per-machine breakdown.

System 1 achieves an overall harmonic mean of 65.10%. The six- and seven-branch fusions improve this score to 65.89% and 65.86%, respectively. Their near-identical performance indicates that adding another fine-grained aggregation view mainly introduces redundant evidence. By contrast, the compact two-branch System 4 reaches 67.20%, improving over System 1 by 2.10 points and over the best larger ensemble by 1.31 points. System 4 also gives the highest overall AUC_s, AUC_t, and pAUC, so its gain is not caused by a single metric component.

The machine-level results expose where this improvement comes from. System 4 obtains the best hmean on fan, gearboxEmu, sliderEmu, ToyCar, and ToyCarEmu. The largest gain over System 1 occurs on gearboxEmu, from 64.97% to 74.37%, together with large increases in all three component metrics. System 4 also improves the target-domain AUC of ToyCarEmu to 83.70% and raises the ToyCar hmean to 76.02%. These cases support the use of complementary band-preserving views when abnormal evidence is not equally expressed across temporal-spectral regions.

The compact fusion is not uniformly best. System 2 remains strongest on bearingEmu and valveEmu, and System 4 reduces the bearingEmu hmean relative to the single localized branch. This machine dependence explains why the larger ensembles remain useful alternative submissions. More generally, the results suggest that fusion quality depends on error complementarity rather than branch count: a carefully chosen pair can outperform six or seven partially redundant views, but no fixed subset dominates every machine.

5. CONCLUSION

This report presented the WISTLAB submission to DCASE 2026 Task 2. The system uses task-adapted BEATs representations and concentrates on downstream anomaly scoring. Local evidence is retained at multiple temporal-spectral granularities, calibrated against the neighborhood structure of normal references, and combined through score-level fusion. Among the four submissions, the compact two-branch fusion performs best and achieves a development-set harmonic mean of 67.20%. The comparison with the larger ensembles shows that carefully selected complementary scoring views can be more effective than simply increasing the number of branches.

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] 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.

[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 *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 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," *Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, 2023.

[5] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 318–14 328.

[6] 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 *Proceedings of Interspeech 2024*, 2024, pp. 107–111.

[7] P. Saengthong and T. Shinozaki, "GenRep for first-shot unsupervised anomalous sound detection of machine condition monitoring," DCASE 2025 Challenge Task 2 Technical Report, Tech. Rep., 2025.

[8] 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*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[10] 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.

[11] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[12] 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*, vol. 33, pp. 4126–4141, 2025.

[13] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 15 877–15 881.

[14] 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," DCASE 2025 Challenge, Tech. Rep., June 2025.

[15] 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," DCASE 2025 Challenge, Tech. Rep., June 2025.

[16] 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.

[17] P. Saengthong and T. Shinozaki, "Sub-band spectral matching with localized score aggregation for robust anomalous sound detection," *arXiv preprint arXiv:2603.13749*, 2026.

[18] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Local Density-Based Anomaly Score Normalization for Domain Generalization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4642–4652, Jan. 2026.