A FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION SYSTEM FOR DCASE2026 TASK 2

Technical Report

Peiwei Chang, Yuelan Cheng, Yongqiang Chen, Philip J.B. Jackson, Wenwu Wang

Centre for Vision, Speech, and Signal Processing (CVSSP)
University of Surrey
Guildford, UK

ABSTRACT

We present our submissions to Detection and Classification of Acoustic Scenes and Events (DCASE) 2026 Challenge Task 2 (Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring). We combine up to ten complementary anomaly signals using a per-machine, label-free adaptive router. Each component is weighted by the Kolmogorov–Smirnov distance between its test-clip and train-clip score distributions, requiring no anomaly labels and no cross-machine fitting. The signals span frozen AudioSet self-supervised embeddings (M2D, M2D-CLAP, BEATs, EAT) scored by cosine k-NN, classical condition-monitoring and inter-channel spatial features, a domain-adapted M2D obtained by further masked-modeling pre-training on machine audio (broadband and 4–7.9 kHz band), and a two-microphone spatial noise-cancellation channel that removes the coherent factory noise before re-embedding. Our system uses per-clip weighted-maximum fusion, treating anomaly detection as an existential event over components. On the development set, our submitted system achieves a source-domain AUC of 76.3% a target-domain AUC of 68.0%, and a pAUC of 56.9%.

***Index Terms*—** anomalous sound detection, first-shot, self-supervised embeddings

1. INTRODUCTION

DCASE 2026 Task 2 [1] addresses noise-aware first-shot unsupervised anomalous sound detection (ASD) for machine condition monitoring. The task is to decide in a short audio recording, whether a machine is operating normally or anomalously, training only on normal-operation sounds. And at test time, generalising to machine types never seen during development ("first-shot"). The five evaluation machine types are disjoint from the seven development types, and each clip is two-channel (near/far microphone) recorded with real factory noise. The data follow the ToyADMOS2 [2] and MIMII DG [3] conventions with source/target domain shifts. The official score is the harmonic mean of AUC and partial AUC (max FPR = 0.1) over machine types, sections and domains [4]. We report development performance as TOTAL, the harmonic mean across the seven machines of each machine's harmonic mean of source-domain AUC, target-domain AUC and pAUC, which closely tracks this official score.

Two empirical observations guided our design. First, per-machine variance: per-machine development total scores range from approximately 0.52–0.75 across the systems, which is much larger than the differences between normal systems, and no single representation wins everywhere (Fig. 2). Second, strong AudioSet self-supervised backbones are already noise-robust, making explicit noise modelling redundant. The more productive directions are therefore orthogonal signals, domain adaptation, and exploitation of the two microphone signals.

2. SYSTEM DESCRIPTION

Fig. 1 shows the pipeline of the primary system. All four submissions share the structure and differ only in the component pool and fusion operator (Table 1).

2.1. Component anomaly signals

Each component produces anomaly scores for the test clips and the training clips of a machine: **(1) fusion** – concatenated L2-normalised M2D-AS [5] and M2D-CLAP [6] clip embeddings (ch0), k=1 cosine k-NN distance to the machine's training memory; **(2) EAT** and **(3) BEATs-diff** – frozen EAT [7] embeddings and BEATs [8](ch0)−BEATs(ch1) spatial-difference embeddings, k-NN; **(4) CM** – 34 classical condition monitoring features (crest/impulse/clearance factors, spectral kurtosis, envelope-spectrum statistics), Ledoit–Wolf Mahalanobis distance; **(5) IC** – 19 inter-channel features (coherence spectrum, band-wise level differences, GCC-PHAT delay), Mahalanobis; **(6) FPT** – M2D further pre-trained for 200 epochs of masked spectrogram modelling on the pooled normal machine audio, k-NN; **(7) FPT-high** – the same backbone on 4–7.9 kHz band-passed audio; **(8) spatial-ANC** – a per-clip frequency-domain Wiener filter estimates the component of ch0 coherent with ch1 (transfer function $H(f) = S_{01}/S_{11}$) and subtracts it, removing the shared factory noise (cross-channel correlation drops from $\sim$0.88 to $\sim$0); the cleaned signal is re-embedded with M2D-AS, k-NN; **(9) clean→FPT** and **(10) clean→FPT-high** – the cancelled signal embedded with the domain-adapted backbone (broadband / high-band): the composition of the two strongest operations. A control replacing the cancelled input with plain ch0 hurts, isolating the gain to the cancellation itself. For the M2D-free System 3, the cancelled signal is instead embedded with EAT (clean→EAT), and EAT / BEATs-diff high-band variants replace the FPT channels.

2.2. Per-machine label-free adaptive routing

For machine m and component i, let s_i and t_i denote the test and training score distributions respectively. We define the reliability

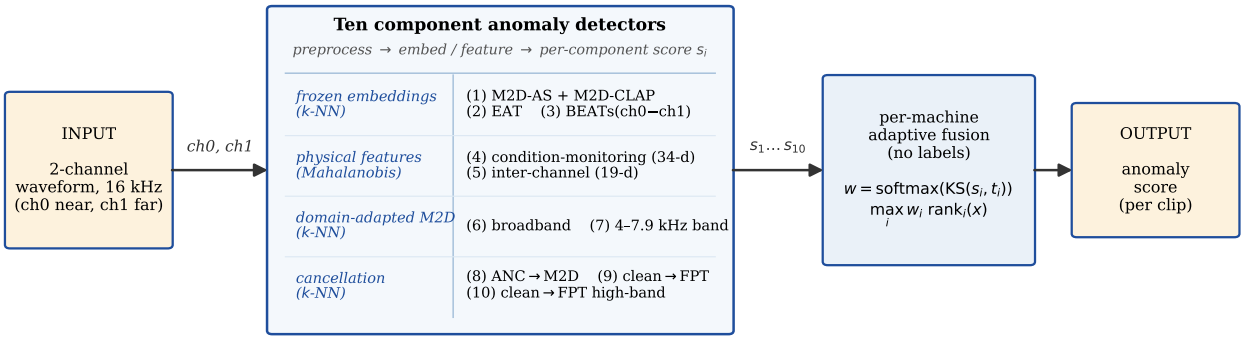

Figure 1: System overview (primary system). The two-channel 16 kHz clip is processed by ten complementary anomaly detectors, grouped by family with their scorers. Each giving one per-clip score. Detectors 6–10 add the domain-adapted backbone, two-microphone noise cancellation and band-pass front-ends. A per-machine, label-free router weights the ten scores by their test-vs-train distribution shift, and a per-clip weighted maximum gives the output anomaly score.

proxy as $p_i = \mathrm{KS}(s_i, t_i)$, the Kolmogorov–Smirnov statistic between the two distributions. Intuitively, a component that responds to anomalous behaviour should exhibit a larger shift from its own normal distribution, resulting in a higher KS value. We then compute per-machine weights $w = \mathrm{softmax}\big((p - \max p)/\mathrm{std}(p)\big)$, label-free and with no cross-machine fitting, and rank-normalise the component scores before fusion.

2.3. Fusion and decisions

Anomaly detection is an existential problem, a fault may be visible in only one view. Mean fusion can therefore dilute strong evidence from a single specialist. Our system instead uses a per-clip weighted maximum fusion rule, $\mathrm{score}(x) = \max_i w_i \, \mathrm{rank}_i(x)$ (with an ϵ-weighted-mean term), which raises $\mathrm{AUC}_{\mathrm{source}}$, $\mathrm{AUC}_{\mathrm{target}}$ and pAUC simultaneously over the weighted mean (TOTAL $0.653 \rightarrow 0.661$). In contrast, the unweighted maximum degrades to 0.595, so the adaptive router weights are essential. Binary decisions are obtained by thresholding each machine's test scores at the median. Every design decision was gated on robustness across three independent label-free proxies (KS, z-score, median-ratio) and by leaving one machine out re-evaluation, to avoid over-fitting the seven development machines.

3. SUBMITTED SYSTEMS

Table 1 summarises the four submitted systems. They were selected by a Monte-Carlo analysis of the expected best-of-four performance under correlated eval-score noise (correlations measured label-free on the blind evaluation scores). System 1 is the peak; System 2 removes only the max-fusion risk; System 3 removes the entire M2D/FPT family (hedging a backbone-family failure); System 4 uses no learned embedding at all (hedging a deep-representation failure, and immune to router pathologies of deep scores).

4. RESULTS

Table 2 reports per-machine development results (System 1 AUC decomposition and the per-machine TOTAL of all four submissions), Fig. 2 compares them to the official baseline, and Table 3 gives the cumulative ablation.

Table 1: The four submitted systems.

#	Pool / fusion	Comp.	TOTAL
1	full pool, weighted-**max**	10	**0.661**
2	full pool, weighted mean	10	0.653
3	non-M2D pool (EAT/BEATs/phys.), mean	7	0.633
4	physical features only (CM+IC), mean	2	0.579
	official AE baseline [9]	–	0.575

All four systems exceed the baseline, and System 1 improves it by 8.6 points overall (0.661 vs. 0.575). Per machine (Fig. 2), System 1 beats the baseline on six of seven machines, with the largest margins on valveEmu (+0.19), ToyCar (+0.14) and fan (+0.11), the exception is bearingEmu (0.621 vs. 0.626).

The AUC decomposition of System 1 locates the difficulty: source-domain AUC (76.3) clearly exceeds target-domain AUC (68.0), which in turn exceeds pAUC (56.9). The source/target gap reflects the domain shift. The target domain offers only a few normal training clips, so the k-NN memory is sparse and normal target clips can look anomalous. This is extreme on fan, whose 96.3 source AUC collapses to 54.0 in the target domain.

Contrasting the systems isolates each design choice. Per-clip weighted-max fusion (System 1 vs. 2) is the only change that raises all three metric columns at once (+0.008 overall). Removing all learned embeddings (System 3→4, −0.054) is uniformly damaging except on valveEmu, isolating the value of deep representations to the non-impulsive machines. These gains come from routing and combining orthogonal signals rather than refining one representation. Eight attempts to train a head on the frozen embeddings all under-performed plain k-NN, and five further backbones (AST [10], PANN [11], HTS-AT [12], WavLM [13], wav2vec2 [14]) were redundant with the deployed components or carried no complementary anomaly signal.

On the unseen evaluation machines, the router assigns mechanistically sensible weights without any labels. The Toy machine (ToyDrone) routes to EAT, while the mechanical machines rely

Table 2: Per-machine development results. Left: System 1 decomposition (%). Right: per-machine TOTAL of the four submitted systems. Last row: harmonic means.

	System 1 (%)			TOTAL			
Machine	AUC_s	AUC_t	pAUC	S1	S2	S3	S4
ToyCar	77.4	88.2	61.4	0.740	0.746	0.742	0.585
ToyCarEmu	73.0	84.2	63.4	0.726	0.679	0.646	0.567
bearingEmu	69.7	61.9	56.2	0.621	0.608	0.560	0.567
fan	96.3	54.0	52.7	0.627	0.614	0.570	0.522
gearboxEmu	73.6	64.0	54.7	0.632	0.618	0.610	0.597
sliderEmu	67.1	60.4	52.3	0.593	0.592	0.606	0.526
valveEmu	84.0	78.1	59.5	0.723	0.753	0.745	0.729
h-mean	76.3	68.0	56.9	**0.661**	0.653	0.633	0.579

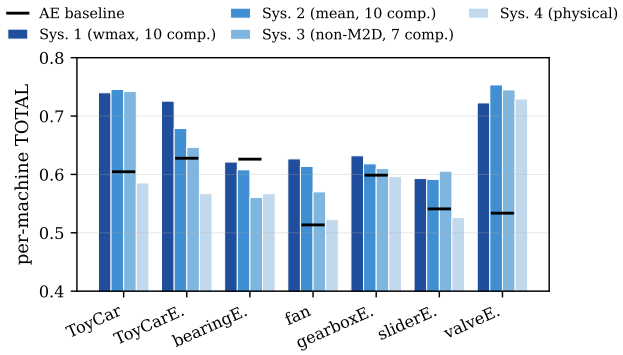

Figure 2: Per-machine development TOTAL of the four submitted systems vs. the official AE baseline (black ticks).

on the cancellation and domain-adapted channels (e.g. clean→FPT weight 0.34–0.38 on BlowerDustCollector, spatial-ANC 0.44 on SewingMachine), suggesting the routing strategy transfers beyond the development machine types.

5. CONCLUSION

We presented a system for noise-aware first-shot unsupervised ASD that detects abnormal machine sound from two-channel audio, using no anomalous examples and no per-machine training. It performs per-machine, label-free routing over orthogonal anomaly signals with frozen self-supervised embeddings, classical fault physics, inter-channel spatial cues, a domain-adapted backbone, explicit two-microphone cancellation of the coherent factory noise, and their compositions. Through a per-clip weighted maximum, we improve the official baseline by 8.6 TOTAL points on the development set.

Two observations stand out. First, the largest gains did not come from a stronger individual representation, but from introducing complementary signal operations, such as spatial cancellation and domain adaptation. In several cases, the fusion operator mattered more than the individual components. Second, the entirely training-free router assigned physically meaningful weights even on unseen evaluation machines. Together, these results suggest a practical recipe for first-shot ASD: an ensemble of frozen detectors can be made useful without labelled anomalies or per-machine training, provided that a simple distribution-shift proxy is available to weight the detectors. This gives a low-cost and easily reusable basis for real condition-monitoring deployments.

Table 3: Cumulative development-set ablation towards System 1.

System	TOTAL
Official baseline (AE, selective Mahalanobis) [9]	0.575
M2D-AS + k-NN	0.591
+ M2D-CLAP fusion	0.599
Adaptive routing (3 deep backbones)	0.610
+ CM + IC physical features	0.615
+ domain-adapted FPT (+ high-band)	0.631
+ spatial noise cancellation	0.635
+ clean→FPT (+ high-band) (= System 2)	0.653
+ per-clip weighted-max fusion (= System 1)	**0.661**

Several limitations remain. We did not domain-adapt the non-M2D backbones, nor did we learn the routing function. A meta-router that selected the fusion operator per machine was anti-correlated with the oracle choice, suggesting that operator selection is a nontrivial problem rather than a straightforward supervised target. The low-false-positive regime, measured by pAUC, also remained the bottleneck under every calibration strategy we tried.

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 *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 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. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2022.

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2023.

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

[6] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2D-CLAP: Masked modeling duo meets CLAP for learning general-purpose audio-language representation," in *Proc. Interspeech*, 2024.

[7] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. International Conference on Machine Learning (ICML)*, 2023.

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

[10] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[12] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.