NOISE-AWARE FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION VIA FROZEN ENCODERS, WIENER DENOISING, AND SHIFTALL-LDKNN SCORING

Technical Report

Yasaman Shokriazar[1]*, Arian Moradi*[1]*,*
Felix Leber[1]*, Alfiia Ziganshina*[1]

[1] Johannes Kepler University Linz, Linz, Austria,
{K12245261, K12240625, K11919221, K12456306}@students.jku.at

ABSTRACT

We describe a system for DCASE 2026 Challenge Task 2: first-shot noise-aware unsupervised anomalous sound detection for machine condition monitoring. The system uses two frozen pre-trained audio encoders—BEATs and EAT-base—without any fine-tuning or gradient updates. Two-channel audio is processed via stereo Wiener denoising (using the far microphone as a noise reference) and a raw near – far difference view to suppress environmental noise. For anomaly scoring, we use local-density k-nearest-neighbour (LDKNN) heads. To overcome the extreme scarcity of target-domain training data, we apply a mean-shift augmentation (*shiftall*) that projects source-domain embeddings into the target domain, effectively expanding the small ten-sample target memory bank. The final anomaly score is a linear fusion of four complementary components using globally fixed weights, with no per-machine tuning or evaluation-score normalization. To ensure generalizability to unseen machines, validation follows a strict leave-one-machine-type-out (LOMO) cross-validation protocol. We submit two weight configurations. Our primary submission utilizes *LOMO-consensus* weights, obtained by averaging the optimal weights across all LOMO folds, ensuring weights are selected without ever seeing the full held-out fold scores (dev $\Omega = 0.6235$). A secondary backup submission uses weights optimized directly on the full development set ($\Omega = 0.6256$). The strict LOMO out-of-fold average harmonic mean is 0.6147 (worst fold 0.5604).

Index Terms— anomalous sound detection, machine condition monitoring, domain generalization, BEATs, EAT, LDKNN, Wiener filtering, shiftall augmentation

1. INTRODUCTION

Anomalous sound detection (ASD) for industrial machines is a central problem in acoustic condition monitoring. DCASE 2026 Challenge Task 2 [1, 2] builds on the series of DCASE ASD tasks and datasets [3, 4, 5, 6] by introducing two key difficulties.

First, only a single *normal* section of each machine is provided at training time (*first-shot* setup). In the DCASE terminology, a *section* denotes the machine-specific subset used for training and testing under a given setup or operating condition. Having only one normal section prohibits multi-section averaging. Second, recordings are made with two microphone channels: a near microphone capturing a stronger machine signal and a far microphone capturing a stronger environmental noise component. Robust systems must handle both the small training set (1 000 training clips per machine: 990 source-domain and 10 target-domain normals) and the channel-wise noise mismatch.

The main contribution of our system is the combination of several train-only and frozen components for this first-shot noise-aware setting, rather than a single isolated module. Specifically, we combine two-channel denoising, frozen pretrained audio encoders, source-to-target embedding-space augmentation, local-density KNN scoring, and fixed global score fusion.

Our approach avoids any per-machine hyperparameter selection or model fine-tuning. The complete pipeline is illustrated in Fig. 1. We exploit the dual-channel signal as a noise reference for Wiener filtering and as a near–far channel-difference view, extract embeddings with two frozen AudioSet-pre-trained encoders, and score with two frozen AudioSet-pre-trained encoders, and score with a local-density k-nearest-neighbour (LDKNN) head augmented by a train-only embedding-space shift that stabilizes the sparse target-domain memory bank. LDKNN modifies a standard KNN anomaly score by normalizing query–neighbour distances with the local density of the corresponding training neighbour. Here, the difference view denotes time-domain channel differencing, $x_{\mathrm{diff}} = x_{\mathrm{near}} - x_{\mathrm{far}}$, rather than spectral-domain subtraction.

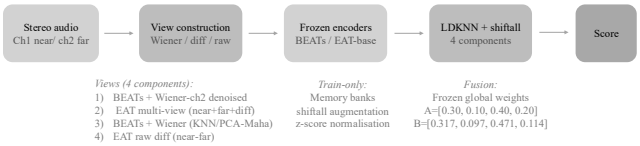

Figure 1: System pipeline. Two-channel audio is processed into four complementary views, encoded with two frozen pre-trained encoders, scored with LDKNN + shiftall heads, and fused with frozen global weights. All memory banks, shift vectors, and normalization statistics are computed from training normals only.

2. TASK AND DATASET

The *development* (seven machine types) and *additional training* sets each provide 1,000 normal training clips per machine (990 source-domain, 10 target-domain). The development set also provides 200 labelled test clips per machine for validation. The *evaluation set* provides 200 blind test clips per unseen machine type, which carry no labels, domain information, or metadata.

A key challenge is the severe 99:1 source–target training imbalance. To prevent detectors from biasing entirely toward source-

domain normality (yielding poor target AUC), our method separates domain memory banks and stabilizes the sparse target bank using a train-only source-to-target shift augmentation (detailed in Sec. 3.5).

The official evaluation metric is the harmonic mean (Ω) of the source-domain AUC, target-domain AUC, and partial AUC (maximum false-positive rate 0.1), averaged across all machine types:

$$\Omega = \mathrm{HM}\big(\mathrm{AUC}_{\mathrm{src}},\ \mathrm{AUC}_{\mathrm{tgt}},\ p\mathrm{AUC}_{0.1}\big). \tag{1}$$

3. PROPOSED METHOD

3.1. Two-Channel Preprocessing and Views

Two complementary audio views are extracted from the stereo pair $(x_{\text{near}},\ x_{\text{far}})$.

Wiener-denoised view. Channel 2 (far microphone) serves as a noise reference in a Wiener filter applied to channel 1 (near microphone) [7]. This suppresses broadband environmental noise while retaining the machine-specific signal captured at close range.

Raw difference view. The per-sample difference $x_{\text{diff}} = x_{\text{near}} - x_{\text{far}}$ subtracts correlated ambient content. This view is particularly effective for machines (e.g. fan) where correlated far-field content in the near channel creates a spurious source-domain structure that degrades target-domain AUC.

3.2. Frozen Encoders

To prevent overfitting to the extremely sparse target domain, all encoder weights are strictly frozen, meaning no gradient updates are performed. We extract 768-dimensional embeddings using two AudioSet-pretrained transformers, with AudioSet providing large-scale weakly labelled audio-event supervision [8]:

BEATs [9] (`BEATs_iter3_plus`) serves as the backbone for Components 1 and 3, while **EAT-base** [10] (`worstchan/EAT-base_epoch30_finetune_AS2M`) is used for Components 2 and 4.

Our scoring approach builds upon the frozen EAT-base KNN backend of Wang [11], extending it via two-channel preprocessing, local-density scoring, *shiftall* augmentation, and multi-component fusion. We employ dual encoders because they exhibit highly complementary failure modes. While BEATs excels at capturing general machine structure (yielding high performance on *valveEmu* and *bearingEmu*), it struggles with correlated background noise. Conversely, EAT-base effectively processes the raw near–far difference view to uniquely resolve the *fan* bottleneck. Fusing these representations provides broad, robust coverage without requiring task-specific fine-tuning.

3.3. Train-Only Memory Banks

For each machine, separate source- and target-domain memory banks are built from training normals. Source bank: 990 embeddings. Target bank: 10 embeddings (the entire target-domain training set).

No test embeddings enter the memory banks at any stage.

3.4. LDKNN Scoring Head

Standard k-nearest-neighbour (KNN) scoring often triggers false positives in sparse regions of the normal data manifold. Local-density KNN (LDKNN) [12] mitigates this by dividing each query–neighbour distance by the neighbour's local neighbourhood density. This ensures that larger distances are tolerated in sparse normal regions, while tighter distances are required in dense regions:

$$s_j = \frac{d(q,\, b_j)}{\rho(b_j)}, \qquad s = \min\big(\bar{s}_{\mathrm{src}},\ \bar{s}_{\mathrm{tgt}}\big), \tag{2}$$

where d is the cosine distance, $\rho(b_j)$ is the mean distance from b_j to its k_s (source, k_s=16) or k_t (target, k_t=9) nearest bank neighbours, and $\bar{s}$ is the mean over the q_k=1 closest bank entries per domain. The domain-minimum combiner in (2) selects the more confident domain prediction per clip. Scores are z-normalized using training-set statistics only.

3.4.1. PCA-Mahalanobis Backend (used in component 3)

While Components 1, 2, and 4 use the LDKNN scoring heads described above, Component 3 (`bw_v1`) uses a different scoring method. This component is an earlier BEATs + Wiener branch that relies on a PCA-Mahalanobis backend applied to the frozen BEATs embeddings, following the common use of Mahalanobis distance for anomaly detection in high-dimensional feature spaces [13].

For each machine, section, and domain group, we first collect the corresponding training-normal embeddings. This group is used in two ways. First, it defines a PCA-Mahalanobis model: embeddings are projected to a d=256-dimensional PCA space fitted on the training normals, and the Mahalanobis distance to the projected training mean is computed:

$$D_{\mathrm{M}}(x) = \sqrt{(x-\boldsymbol{\mu})^{\top}\,\boldsymbol{\Sigma}^{-1}\,(x-\boldsymbol{\mu})}, \tag{3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the projected training-normal embeddings of the same machine, section, and domain group.

Second, the same projected training-normal embeddings are kept as a cosine KNN memory bank. For a test embedding, we compute both the PCA-Mahalanobis distance to the fitted normal distribution and the cosine k=3 KNN distance to the stored projected training embeddings. The two distances are blended with equal weight, yielding the component 3 score. As with the LDKNN heads, all PCA, covariance, mean, and KNN memory-bank statistics are computed from training data only.

3.5. Shiftall Target-Domain Augmentation

A naive pooling of the training data creates a 99:1 domain imbalance, effectively forcing the model to learn source-domain characteristics. This typically results in excellent source AUC but poor target AUC, as the sparse target normals are statistically treated as outliers. Furthermore, the 10-sample target bank is too small for reliable PCA-Mahalanobis covariance estimation or LDKNN local-density calculation.

To resolve this imbalance, we apply a train-only embedding-space mean-shift augmentation:

$$\Delta = \mu_{\mathrm{tgt}} - \mu_{\mathrm{src}}, \qquad \tilde{x}_i = \mathrm{L2norm}(x_i^{\mathrm{src}} + \Delta), \tag{4}$$

where μ are mean embeddings computed from training normals. All 990 shifted source embeddings $\{\tilde{x}_i\}$ are appended to the target bank, yielding a 1 000-sample effective target bank. This shift uses only training-set statistics and introduces no label information.

While estimating a local density manifold or covariance matrix from only 10 samples is highly unstable, estimating their simple geometric centroid (μ_{tgt}) is statistically much more robust.

3.6. Four Scoring Components

The final submission linearly fuses the scores of four complementary components. In our system, a component refers to an individual end-to-end anomaly detector branch consisting of a specific preprocessing view, a frozen encoder, and a dedicated scoring backend. The configurations of these four individual detectors are detailed in Table 1.

Table 1: Comparison of the four scoring components. Component numbers correspond to the fusion weights w_1–w_4 in Table 3.

Component	Backbone	View	Scoring Method	Notable Characteristics
bw_v3 [1]	BEATs + Wiener	Wiener-denoised	LDKNN + shiftall	Primary contributor; improves *sliderEmu* and *ToyCarEmu*.
eat_diff [2]	EAT-base	Multi-view (near + far + diff)	Plain LDKNN	Low fusion weight (0.097/0.10) but consistent fan coverage.
bw_v1 [3]	BEATs + Wiener	Wiener-denoised	Cosine-KNN + PCA-Mahalanobis ($k = 3$, $d = 256$)	Highest LOMO-preferred component; stabilizes fusion.
eat_diff_v3 [4]	EAT-base	Raw diff view ($x_{\text{near}} - x_{\text{far}}$)	LDKNN + shiftall	Both fan source AUC and fan target AUC above chance (0.599 / 0.615).

Component 4 addresses the fan bottleneck that persisted across all BEATs-based candidates. Table 2 shows the fan per-candidate source/target AUC.

Table 2: Fan machine source and target AUC per component. The bold numbers indicate the best value in each AUC column. Component 4 provides the most balanced source–target AUC performance among the four components.

Component	Src AUC	Tgt AUC	Balanced?
bw_v3 [1]	0.713	0.406	No
eat_diff [2]	0.517	0.610	Partial
bw_v1 [3]	**0.721**	0.443	No
eat_diff_v3 [4]	0.599	**0.615**	**Yes**

3.7. Frozen Global Fusion

Four component scores (s_1, s_2, s_3, s_4) are linearly fused with frozen global weights applied identically to all machines:

$$s_{\text{final}} = \sum_{i=1}^{4} w_i \cdot s_i. \tag{5}$$

Two weight sets are provided (Table 3). The decision threshold is $s_{\text{final}} > 0$, derived from the z-normalization at the component level; no evaluation-score distribution is used at fusion time.

Table 3: Frozen global fusion weight sets. No per-machine tuning.

Set	w_1	w_2	w_3	w_4	Selection
A	0.300	0.100	0.400	0.200	Full-dev Ω max.
B	**0.317**	**0.097**	**0.471**	**0.114**	LOMO-consensus (primary)

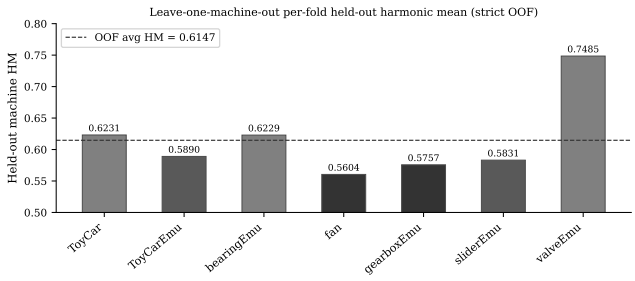

Figure 2: Strict LOMO per-fold held-out harmonic mean. Each bar is the HM evaluated on the held-out machine using weights selected from the remaining six machines. The fan fold (0.5604) is the worst; valveEmu (0.7485) is the best. OOF average = 0.6147.

4. VALIDATION PROTOCOL AND COMPLIANCE

4.1. Leave-One-Machine-Out Protocol

For each of the seven development machine types we hold out that machine, select weights using only the remaining six machines, evaluate on the held-out machine, and report the held-out harmonic mean (HM). To prevent information leakage, the implementation excludes all data and embeddings from the held-out machine during weight selection.

4.2. Result Tiers

We report two different validation results.

Primary submitted system—LOMO-consensus, weight set B (dev $\Omega = 0.6235$): Fusion weights are the mean of per-fold LOMO-best weights, clipped and renormalized. Because weights are derived without simultaneous access to all seven development machine labels, this is the primary official submission. **Conservative generalization estimate—LOMO cross-validation** (average held-out HM $= 0.6147$, worst fold $= 0.5604$): Each machine type is held out in turn; the held-out machine is scored with weights selected on the *remaining* machine types only. No labels from the held-out machine influence weight selection. This is the most reliable estimate of unseen-machine performance.

For reference, an internal development run that applied a transductive per-CSV z-score normalization at fusion time reached $\Omega = 0.6277$; that result is not evaluation-clean and was not used for submission.

4.3. Compliance and Deploy-Config Verification

The submitted outputs were generated under the following constraints: no per-machine hyperparameters, no anomaly labels inside any scoring function, no per-machine fusion weights, no threshold tuning on evaluation data, and no evaluation-score normalization.

5. DEVELOPMENT RESULTS

Table 4 and Fig. 2 detail our system performance. While performance on the blind evaluation data remains to be seen, our primary submission ($\Omega = 0.6235$, strict OOF $= 0.6147$) significantly outperforms the official DCASE 2026 Task 2 MAHALA baseline ($\Omega \approx 0.577$) on the development set.

Table 4: Per-machine results on the development set using weight set B ($\Omega = 0.6235$, full-dev evaluation, not OOF).

Machine	Src AUC	Tgt AUC	pAUC	HM
ToyCar	0.6868	0.7256	0.5347	0.6377
ToyCarEmu	0.5436	0.8568	0.4911	0.5949
bearingEmu	0.6472	0.6200	0.6221	0.6295
fan	0.7200	0.4736	0.5858	0.5761
gearboxEmu	0.6140	0.6464	0.5289	0.5922
sliderEmu	0.6472	0.6096	0.5474	0.5985
valveEmu	0.9196	0.7064	0.7242	0.7724
Overall Ω		**0.6235**		

While all other machines achieve HM > 0.59, *fan* remains the hardest case. Because *fan* and *ToyCar* are the only real (non-emulated) two-channel recordings in the development set, *fan* acts as the worst-case LOMO fold (HM $= 0.5604$). However, introducing the raw-diff EAT-v3 component successfully raised the final-fusion *fan* HM to 0.5761.

Generalization to the Evaluation Set. Because all five evaluation machines are real recordings, *fan* serves as the closest development proxy. Specifically, the evaluation set includes *BlowerDustCollector* and *ToyDrone*—machines that generate high-RPM aerodynamic broadband noise highly analogous to *fan*. Solving the *fan* bottleneck is therefore strictly necessary for evaluation generalization, directly motivating the label-free transfer diagnostics in Section 6 (Fig. 4).

6. EVALUATION SUBMISSION AND TRANSFER DIAGNOSTICS

6.1. Evaluation Submissions

Two frozen submissions are provided for the five evaluation machine types (ToyDrone, ToothBrush, SewingMachine, Sander, BlowerDustCollector):

Primary submission (weight set B, LOMO-consensus): Weights derived via LOMO without simultaneous access to all seven development machine labels. Corresponds to weight set B in Table 3 ($[0.317, 0.097, 0.471, 0.114]$, dev $\Omega = 0.6235$).

Backup submission (weight set A, full-dev no-z-score): Weights selected by maximising Ω over all seven development machines. Competitive but uses full development label set for selection; provided as an alternative. Corresponds to weight set A in Table 3 ($[0.30, 0.10, 0.40, 0.20]$, dev $\Omega = 0.6256$).

Each submission contains five `anomaly_score_*.csv` and five `decision_result_*.csv` files (200 rows each). No evaluation labels, no evaluation-score normalization, and no per-CSV z-score were used at any stage of evaluation submission generation.

6.2. Label-Free Transfer Diagnostics

Since evaluation labels are hidden, we characterize estimated transfer risk using two label-free diagnostics.

A-vs-B score stability (Fig. 3). All five evaluation machines show Spearman correlation $\rho \geq 0.983$ between weight sets A and B. Decision disagreement is at most 5% (Sander: 10 of 200 clips). The two submissions are largely equivalent in ranking quality.

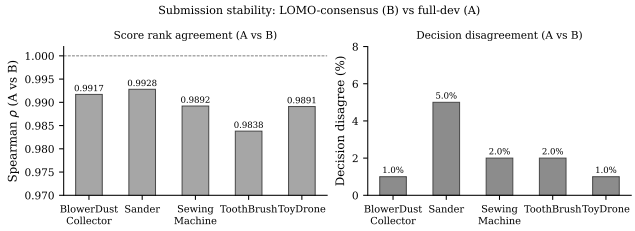

Figure 3: Left: Spearman score-rank correlation between submissions A and B for each evaluation machine (all ≥ 0.983). Right: percentage of clips where the binary decision differs between A and B (at most 5%). Diagnostics are label-free; no evaluation labels were used.

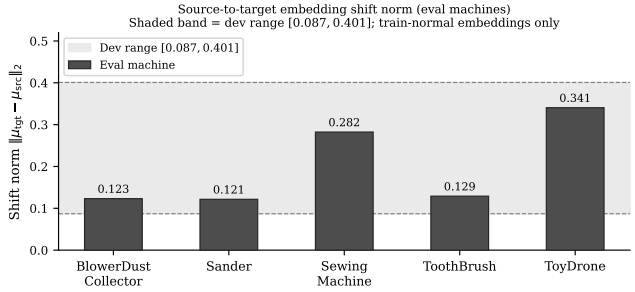

Figure 4: Source-to-target embedding shift norm per evaluation machine (bars) compared with the development range (shaded band, $[0.087, 0.401]$). All evaluation machines are within range. ToyDrone (shift 0.341, cosine distance 0.296 to nearest dev machine) is the watch item. Shift vectors are computed from training normals only.

Source-to-target embedding shift norm (Fig. 4). We measure $\|\mu_{\text{tgt}} - \mu_{\text{src}}\|_2$ in BEATs+Wiener embedding space from training normals. All five evaluation shift norms fall within the development range $[0.087,\ 0.401]$, indicating that the shiftall augmentation has precedent from development machines. ToyDrone has the largest shift norm (0.341) and the largest embedding distance from any development machine (cosine 0.296 to nearest sliderEmu); it is the evaluation machine most likely to behave differently from development expectations.

These diagnostics were computed from training data only and were not used to tune any weights or thresholds.

7. CONCLUSION

We presented a first-shot anomalous sound detection system based on frozen BEATs and EAT-base encoders with stereo Wiener denoising, raw near-minus-far difference views, LDKNN scoring, and shiftall target-domain augmentation. The system is designed to satisfy the first-shot constraints of DCASE 2026 Task 2. Development results (LOMO-consensus weight set B, $\Omega = 0.6235$; strict LOMO OOF average HM $= 0.6147$) indicate robust generalization across seven heterogeneous development machine types. Label-free transfer diagnostics confirm low estimated transfer risk for all five evaluation machines.

8. REFERENCES

[1] T. Nishida, K. Dohi, K. Imoto, N. Harada, Y. Koizumi, R. Tanabe, R. Yamamoto, T. Nitta, Y. Kawaguchi, K. Mishima, and S. Nakamura, "Description and Discussion on DCASE 2026 Challenge Task 2: Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," *arXiv preprint*, 2026, arXiv:2606.01578.

[2] DCASE Challenge Organizers, "DCASE 2026 Challenge Task 2 Official Page," https://dcase.community/challenge2026/, 2026.

[3] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "Description and discussion on DCASE 2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Obata, and Y. Kawaguchi, "MIMII DG: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022.

[5] 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.

[6] 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.

[7] Y. Ephraim and D. Malah, "Optimal filtering with signal-to-noise ratio constraints," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1984, pp. 266–269.

[8] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] 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. [Online]. Available: https://arxiv.org/abs/2212.09058

[10] W. Chen, Y. Zhang, Z. Li, Z. Wang, and Y. Wu, "EAT: Self-Supervised Pre-Training with Efficient Audio Transformer," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[11] Y. Wang *et al.*, "Anomalous sound detection using EAT fine-tuned with K-nearest neighbours," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2025, dCASE Challenge Technical Report.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, 2000, pp. 93–104.

[13] K. Lee, K. Lee, H. Lee, and J. Shin, "Mahalanobis distance for anomaly detection in high-dimensional feature spaces," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2018.