MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

# Explainable Unsupervised Sound Anomaly Detection via Sparse Feature Decomposition

Paissan, Francesco; Wichern, Gordon; Masuyama, Yoshiki; Le Roux, Jonathan

TR2026-166 October 29, 2026

**Abstract**

Anomaly detection systems are increasingly embedded in real-world monitoring pipelines, yet they typically produce only a scalar score, leaving operators without the context needed to act. We propose an explanation framework that attributes anomaly scores to interpretable acoustic features. Given an anomalous sample, a trained sparse autoencoder decomposes its anomaly score into per-feature contributions; these features are then linked to semantic attributes via linear probes trained on machine operating condition metadata and validated through counterfactual interventions in the latent space. We demonstrate the approach for unsupervised machine sound anomaly detection on the DCASE2025 Task 2 benchmark, achieving performance comparable to state-of-the-art detection systems and providing human-interpretable explanations of anomalous behaviour.

*Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) 2026*

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

# Explainable Unsupervised Sound Anomaly Detection via Sparse Feature Decomposition

*Francesco Paissan*[1,2]*, Gordon Wichern*[2]*, Yoshiki Masuyama*[2]*, Jonathan Le Roux*[2]

[1] Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA
[2] Université Laval, Québec City, Canada

***Abstract*—Anomaly detection systems are increasingly embedded in real-world monitoring pipelines, yet they typically produce only a scalar score, leaving operators without the context needed to act. We propose an explanation framework that attributes anomaly scores to interpretable acoustic features. Given an anomalous sample, a trained sparse autoencoder decomposes its anomaly score into per-feature contributions; these features are then linked to semantic attributes via linear probes trained on machine operating condition metadata and validated through counterfactual interventions in the latent space. We demonstrate the approach for unsupervised machine sound anomaly detection on the DCASE2025 Task 2 benchmark, achieving performance comparable to state-of-the-art detection systems and providing human-interpretable explanations of anomalous behaviour.**

***Index Terms*—anomaly detection, explainability, sparse autoencoders, acoustic monitoring**

## 1. INTRODUCTION

Anomaly detection in audio is the task of distinguishing normal operational states from anomalous or faulty conditions [1]. In industrial settings, early detection can prevent costly equipment failures, reduce downtime, and improve safety, motivating sustained research in anomalous sound detection (ASD) for machine condition monitoring, for which the DCASE challenge [2]–[5] has become the primary benchmark. ASD presents several inherent challenges. First, training data typically consists only of normal recordings, as deliberately inducing anomalies requires damaging machinery or waiting for natural failures, and natural occurrences are too infrequent to cover the full diversity of failure modes. Second, real-world factory audio is inherently noisy, requiring embeddings that are simultaneously sensitive to subtle machine failures and insensitive to changes in acoustic conditions [6].

Recent advances in self-supervised learning have produced powerful pre-trained audio foundation models (e.g., BEATs [7], [8], EAT [9], CLAP [10]) that encode rich acoustic representations and generalize across tasks without task-specific fine-tuning [11]. For ASD, such models offer clear advantages: robustness to domain shifts and noise through large-scale pretraining on diverse audio corpora, and reduced reliance on labelled target-domain data [12]–[15]. However, this generality comes at a cost: pre-trained embeddings serve as black-box representations, collapsing complex acoustic information into fixed-dimensional vectors. Anomalies are flagged as distance scores between embeddings, but operators receive no insight into *which acoustic properties* deviated from normal operation or *why* the system raised an alert. In safety-critical industrial settings where costly maintenance decisions follow anomaly alerts, this opacity is untenable. The disconnect between model decisions and human understanding undermines operator trust and complicates root-cause diagnosis of failures.

To address this gap, we investigate sparse autoencoders (SAEs) [16], which decompose learned representations into sparse, interpretable latent features. SAEs learn an overcomplete basis by mapping embeddings into a high-dimensional latent space, in which dimensions are encouraged to align with interpretable semantic or acoustic concepts via a sparsity constraint. By enforcing sparsity (generally via $\ell_1$ regularization or Top-K activations), SAEs encourage each latent feature to activate selectively in response to specific input patterns, making individual features human-interpretable [17], [18]. SAEs have proven effective for interpretability in vision [19], [20] and natural language processing [17], [21]–[23], with recent applications in audio understanding [24]–[26]. However, their systematic application to pre-trained audio embeddings for industrial anomaly detection, particularly in decomposing anomaly signals into actionable acoustic insights, remains unexplored. SAE latent dimensions, if properly learned from pre-trained embeddings of normal machine sounds, should align with meaningful acoustic phenomena (e.g., mechanical signatures, spectral shifts, frequency modulation patterns), consistent with findings in audio SAE [24], [25]. By decomposing anomalies through these interpretable features, operators could understand not only *that* an anomaly occurred but also *which* acoustic changes triggered the alert, enabling verification, diagnosis, and justified confidence in automated decisions.

In this paper, we apply sparse autoencoders to pre-trained audio embeddings for interpretable anomaly detection in industrial machine condition monitoring. Our approach (i) learns a sparse, interpretable basis from pre-trained embeddings of normal machine sounds, (ii) decomposes the anomaly score into semantic components, revealing which sparse features deviate from normal operation, and (iii) validates identified features through counterfactual interventions, confirming their functional role in anomaly detection.

## 2. BACKGROUND

### 2.1. Anomaly Detection via Pretrained Audio Encoders

We follow the GenRep [13] pipeline, in which audio samples $\mathbf{x} \in \mathbb{R}^T$ are converted to mel spectrograms, divided into non-overlapping square patches of size $p$, and fed through a BEATs [7] encoder. We extract latent representations from the intermediate layers and pool over the temporal axis to obtain time-invariant, per-frequency embeddings:

$$\mathbf{h} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}_t \in \mathbb{R}^{F\times C}, \tag{1}$$

where $\mathbf{g} = \mathrm{BEATs}(\mathbf{x}) \in \mathbb{R}^{T\times F\times C}$, and in $\mathbf{h}$ each row corresponds to a frequency band and $C$ is the channel dimension.

During inference, an anomaly score for test sample embedding $\mathbf{h}$ is computed by comparing against a reference set $\mathcal{H}_{\text{ref}}$ of normal training embeddings. We use the density-normalized nearest-neighbor score of [12], which has proven effective at overcoming domain shift in recent DCASE challenges:

$$\mathcal{A}(\mathbf{h}_{\text{test}} \mid \mathcal{H}_{\text{ref}}) := \min_{\mathbf{h}\in\mathcal{H}_{\text{ref}}} \frac{d(\mathbf{h}_{\text{test}}, \mathbf{h})}{\sum_{\mathbf{h}_r\in\mathcal{N}_R(\mathbf{h})} d(\mathbf{h}, \mathbf{h}_r)}, \tag{2}$$

This work was done while F. Paissan was an intern at MERL.

where $d(\mathbf{h}, \mathbf{h}') := \|\mathrm{vec}(\mathbf{h}) - \mathrm{vec}(\mathbf{h}')\|_2^2$ is the squared Euclidean distance and $\mathcal{N}_R(\mathbf{h})$ denotes the $R$-nearest neighbors of $\mathbf{h}$ in the reference set. The denominator estimates local density: a test sample scores high if it is far from the training cloud *relative* to how densely packed that cloud is around its nearest neighbor.

### 2.2. SAE Design and Training

We train a $k$-sparse autoencoder (SAE) [16] with a group structure on the pooled representations $\mathbf{h}$. The SAE processes each frequency band as a group. Each group is independently processed through a shared encoder $\mathbf{W}_E \in \mathbb{R}^{L\times C}$ and decoder $\mathbf{W}_D \in \mathbb{R}^{C\times L}$. Weight-sharing across $F$ bands substantially reduces the parameter count relative to a monolithic SAE, enabling stable training in the low-data regime of anomalous sound detection. For each machine type, we have only 1000 samples, and the input embedding dimension would be 6144 for the flattened BEATs representation, resulting in SAE models with millions of parameters. With the shared encoder–decoder architecture, we reduce the parameter count by a factor of $F = 8$ and effectively increase the number of training vectors by the same amount, since each channel vector is processed independently. Critically, per-band sparse codes also preserve the natural structure of spectrograms: because acoustic anomalies could manifest as frequency-localized deviations, separate codes per band allow the SAE to isolate *which* frequencies drive an anomaly.

For each frequency band $i \in \{1, \ldots, F\}$ with channel vector $\mathbf{h}_i \in \mathbb{R}^C$:

$$\mathbf{f}_i = \mathrm{TopK}_k\left([\mathbf{W}_E\mathbf{h}_i + \mathbf{b}_{\text{enc}}]_+\right) \in \mathbb{R}^L, \tag{3}$$

$$\hat{\mathbf{h}}_i = \mathbf{W}_D\mathbf{f}_i \in \mathbb{R}^C, \tag{4}$$

where $L \gg C$, $\mathrm{TopK}_k(\cdot)$ retains only the $k$ largest activations (zeroing the rest), and $[\cdot]_+$ denotes ReLU.

We minimize squared Euclidean reconstruction error, keeping the SAE loss consistent with the distance metric $d$ used in anomaly scoring. To prevent the issue known as "dead latents," a common failure mode under top-$k$ sparsity where some features are never selected in the entire dataset, we add the ghost-gradients-like auxiliary loss presented in [17]. This loss component routes gradients through the top $k_{\text{aux}}$ latents that are not in the top-$k$, thereby preventing collapse. The full objective is:

$$J(\mathbf{h}, \mathbf{W}_E, \mathbf{W}_D) := \sum_{i=1}^{F} d(\mathbf{h}_i, \hat{\mathbf{h}}_i) + \alpha \sum_{i=1}^{F} d(\mathbf{e}_i, \hat{\mathbf{e}}_i)\,, \tag{5}$$

where $\mathbf{e}_i := \mathbf{h}_i - \hat{\mathbf{h}}_i$ is the per-band reconstruction residual, $\mathbf{f}_i^{\text{aux}} \in \mathbb{R}^L$ is the sparse code restricted to the $k_{\text{aux}}$ highest activations outside the top-$k$ for band $i$, $\hat{\mathbf{e}}_i := \mathbf{W}_D\mathbf{f}_i^{\text{aux}}$ is the reconstruction of the sparse code, and $\alpha$ balances the auxiliary term. Encoder weights are initialized as $\mathbf{W}_E = \mathbf{W}_D^\top$; decoder columns are kept at unit norm via gradient projection during optimization.

## 3. GENERATING EXPLANATIONS

### 3.1. Feature Attribution via Distance Decomposition

A natural quantity to explain is the distance between an anomalous sample and its closest normal embedding, which directly reflects the anomaly score defined in Section 2. Given an anomalous embedding $\mathbf{h}_{\text{anom}}$ and its closest normal embedding $\mathbf{h}_{\text{norm}} \in \mathcal{H}_{\text{ref}}$, we encode both through the group-structured SAE. For each frequency band $i$, this yields sparse codes $\mathbf{f}_i^{\text{anom}}$, $\mathbf{f}_i^{\text{norm}}$ and reconstructions $\hat{\mathbf{h}}_i^{\text{anom}}$, $\hat{\mathbf{h}}_i^{\text{norm}}$. We define per-band differences:

$$\Delta\mathbf{h}_i := \mathbf{h}_i^{\text{anom}} - \mathbf{h}_i^{\text{norm}}, \quad \Delta\hat{\mathbf{h}}_i := \hat{\mathbf{h}}_i^{\text{anom}} - \hat{\mathbf{h}}_i^{\text{norm}}, \tag{6}$$

$$\Delta\mathbf{e}_i := \Delta\mathbf{h}_i - \Delta\hat{\mathbf{h}}_i, \tag{7}$$

where we also have $\Delta\hat{\mathbf{h}}_i = \mathbf{W}_D\Delta\mathbf{f}_i$ and $\Delta\mathbf{f}_i := \mathbf{f}_i^{\text{anom}} - \mathbf{f}_i^{\text{norm}}$. The total anomaly score decomposes additively over frequency bands as

$$d(\mathbf{h}_{\text{anom}}, \mathbf{h}_{\text{norm}}) = \sum_{i=1}^{F} \|\Delta\mathbf{h}_i\|_2^2. \tag{8}$$

To attribute this distance to individual SAE features, we expand the reconstructed difference $\Delta\hat{\mathbf{h}}_i$ as a sum of rank-one terms,

$$\Delta\hat{\mathbf{h}}_i = \mathbf{W}_D\Delta\mathbf{f}_i = \sum_{j=1}^{L} \Delta f_{i,j}\, \mathbf{w}_{D,j}, \tag{9}$$

where $\mathbf{w}_{D,j}$ is the $j$-th column of $\mathbf{W}_D$. Taking the squared norm and collecting terms gives

$$\|\Delta\hat{\mathbf{h}}_i\|_2^2 = \sum_{j=1}^{L} \Delta f_{i,j}\, \mathbf{w}_{D,j}^\top \Delta\hat{\mathbf{h}}_i =: \sum_{j=1}^{L} c_{i,j}, \tag{10}$$

which defines the per-feature contribution $c_{i,j}$. By construction, $\{c_{i,j}\}$ exactly partitions $\|\Delta\hat{\mathbf{h}}_i\|_2^2$: a large positive $c_{i,j}$ indicates that feature $j$ pushes the anomalous reconstruction away from the normal one along $\mathbf{w}_{D,j}$, and a negative $c_{i,j}$ indicates that its decoder direction points away from the anomalous deviation, reducing the overall reconstructed distance.

The relationship between the reconstructed and raw distances follows from expanding the residual:

$$\|\Delta\mathbf{h}_i\|_2^2 = \|\Delta\hat{\mathbf{h}}_i\|_2^2 + 2\,\Delta\hat{\mathbf{h}}_i^\top \Delta\mathbf{e}_i + \|\Delta\mathbf{e}_i\|_2^2, \tag{11}$$

so the contributions $\{c_{i,j}\}$ explain $\|\Delta\mathbf{h}_i\|_2^2$ only up to the cross-term $2\,\Delta\hat{\mathbf{h}}_i^\top \Delta\mathbf{e}_i$ and the squared residual $\|\Delta\mathbf{e}_i\|_2^2$. While these terms are small for normal samples, which lie near the training distribution, their magnitude for anomalous samples depends on how well the SAE generalizes beyond it.

To measure how well the reconstructed-space decomposition approximates the raw distance, we use

$$R^2_{\text{decomp}} = 1 - \frac{\sum_i \left(\|\Delta\mathbf{h}_i\|_2^2 - \sum_j c_{i,j}\right)^2}{\sum_i \|\Delta\mathbf{h}_i\|_2^4}\,. \tag{12}$$

High $R^2_{\text{decomp}}$ indicates that the cross-term and residual are negligible. Therefore, the contributions $\{c_{i,j}\}$ faithfully approximate $\|\Delta\mathbf{h}_i\|_2^2$, and the identified features can be trusted as genuine drivers of the anomaly.

### 3.2. Assigning Attributes to SAE Features

To associate SAE features with interpretable acoustic attributes, we train a separate linear probe per acoustic attribute on SAE activations. For each sample, we parse recording metadata into discrete concept labels: for each attribute $c$, we have a set of classes $\mathcal{Y}_c$. Each class denotes an attribute realization, such as a specific car ID, operating speed, or domain. For every concept $c$, we fit an $\ell_1$-regularized multinomial logistic regression probe:

$$p_c(y \mid \mathbf{f}_i) = \mathrm{Softmax}(\mathbf{W}_c\mathbf{f}_i + b_c)_y, \quad y \in \mathcal{Y}_c, \tag{13}$$

where $\mathbf{f}_i \in \mathbb{R}^L$ is the normalized SAE activation for each frequency band of a given recording. We train the probes on normal recordings, fit them separately for each concept, and evaluate them on a held-out test set.

We rank SAE features within each $(c, y)$ pair by weight magnitude in $\mathbf{W}_c$, yielding for each feature index $j$ a set of concepts and classes for which it is most predictive. This mapping is used downstream to attach semantic labels to the highest-contributing features identified by the attribution scores $c_{i,j}$. Within a specific concept, the same

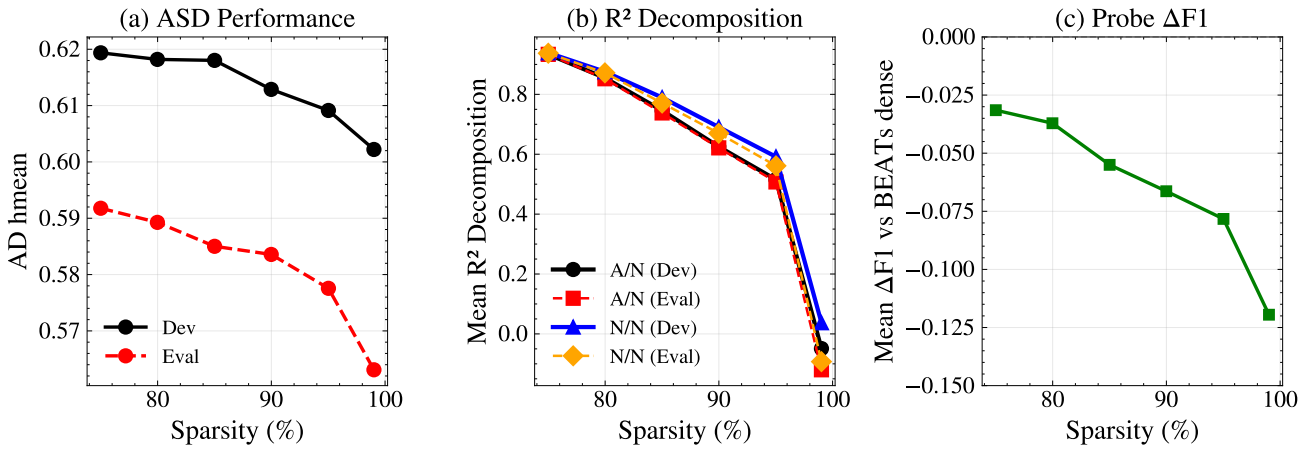


**Fig. 1**: Tradeoff between sparsity and relevant metrics computed for Layer 11. From left to right: anomaly detection performance, $R^2_{\text{decomp}}$, and average F1 drop across machines. In the central panel, "A" and "N" indicate Anomaly and Normal, respectively.

feature can correlate with different classes. In this case, we use the class with the highest weight.

### 3.3. Counterfactual Interventions

To validate that the attributed features are functional drivers of the anomaly, we perform targeted interventions in the sparse latent space. Given codes $\mathbf{f}^{\text{anom}}$ and $\mathbf{f}^{\text{norm}}$, we identify a top-$k'$ feature set $\mathcal{I}_{k'}$ based on $|c_{\cdot,j}|$. We construct a counterfactual code by replacing those features with their normal-sample values:

$$\mathbf{f}^{\text{cf}}_{i,j} = \begin{cases} \mathbf{f}^{\text{norm}}_{i,j} & j \in \mathcal{I}_{k'}, \\ \mathbf{f}^{\text{anom}}_{i,j} & \text{otherwise.} \end{cases} \tag{14}$$

Decoding $\mathbf{f}^{\text{cf}}$ through the decoder yields $\hat{\mathbf{h}}_{\text{cf}}$. We measure the resulting distance reduction:

$$\Delta d_{k'} = \frac{d(\hat{\mathbf{h}}_{\text{anom}}, \hat{\mathbf{h}}_{\text{norm}}) - d(\hat{\mathbf{h}}_{\text{cf}}, \hat{\mathbf{h}}_{\text{norm}})}{d(\hat{\mathbf{h}}_{\text{anom}}, \hat{\mathbf{h}}_{\text{norm}})}. \tag{15}$$

If the identified features are functional drivers of the anomaly, $\Delta d_{k'}$ should increase monotonically with $k'$: replacing more top-contributing features with their normal values should progressively close the gap between the counterfactual and normal embeddings.

## 4. EXPERIMENTS

We run our evaluation on the DCASE 2025 Task 2 benchmark [5], which comprises 15 machine types, split into source and target domains that differ in acoustic conditions. Models are evaluated using AUC aggregated according to the official protocol. Audio is resampled to 16 kHz and encoded through a pretrained BEATs model[1]; the reference set $\mathcal{H}_{\text{ref}}$ contains embeddings for all training samples of the matching machine type. We use $R$=16 nearest neighbours for density normalization [13], SAE latent dimension $L$=2048, auxiliary loss weight $\alpha$=0.1, and $k_{\text{aux}}$=128. We evaluate on BEATs layers 11 and 12 following [14], varying sparsity $k \in [20, 512]$ to study its joint effect on detection performance, explanation reliability, and probe semantic quality.

**Anomaly detection.** We compare our pipeline against the raw BEATs baseline to confirm that the SAE explanation layer introduces no degradation in detection performance.

**Explained anomaly score and counterfactual interventions.** We report $R^2_{\text{decomp}}$ across anomalous test samples to assess decomposition faithfulness, then verify contribution rankings via counterfactual interventions against a random-ranking baseline.

**Semantic quality of the probes.** We report the F1 gap between probes trained on SAE activations versus raw BEATs embeddings; a modest drop would confirm that top-contributing features carry genuine acoustic meaning.

[1] `BEATs_iter3_plus_AS2M_finetuned_on_AS2M_cpt1.pt`

## 5. RESULTS

### 5.1. Anomaly Detection

Fig. 1 (left panel) shows mean ASD performance as a function of sparsity for layer 11. On the development set, performance decreases steadily with increasing sparsity: at 85% and 80% sparsity, the overall hmean reaches $\approx 0.618$, then declines modestly but consistently at higher sparsity levels. At 99% sparsity (i.e., only 20 features are non-zero), performance drops to 0.602, consistent with the reconstruction error increase at high sparsity. The evaluation set shows a similar, slightly more pronounced trend: the overall hmean peaks at $\approx 0.594$ for 80%–85% sparsity and then gradually degrades, falling to 0.566 when all coefficients are retained. Crucially, peak performance remains comparable to the raw BEATs baseline (0.594 on eval, 0.6239 on dev), confirming that the SAE explanation layer introduces no meaningful performance degradation.

The gap between development and evaluation remains relatively stable across sparsity levels ($\Delta \approx 0.03$) and matches the baseline gap, suggesting the SAE does not introduce additional dataset sensitivity beyond that of the baseline. Beyond 85% sparsity, both curves degrade gradually without sharp discontinuities, indicating that moderate-to-high sparsity levels remain informative for anomaly detection.

A similar pattern emerges on layer 12: development performance reaches $\approx 0.627$ overall hmean at sparsity coefficients between 80% and 90%—higher than layer 11, though slightly below the layer 12 baseline of 0.6293. The development-to-evaluation gap is again approximately 0.03. Overall, layer 12 underperforms layer 11 on the evaluation set, with a baseline score of 0.5776.

### 5.2. Explained Distance

We evaluate how SAE sparsity affects representational fidelity by measuring mean $R^2_{\text{decomp}}$ across a range of sparsity levels (75–99%), shown in Fig. 1 (central panel). Reconstruction quality degrades gracefully across the moderate sparsity regime. At 75% sparsity, the sparse SAE accounts for approximately 90% of the distance between embeddings, and this fidelity is broadly maintained up to 95% sparsity, where $R^2_{\text{decomp}}$ remains above 0.6. Beyond 95% sparsity,

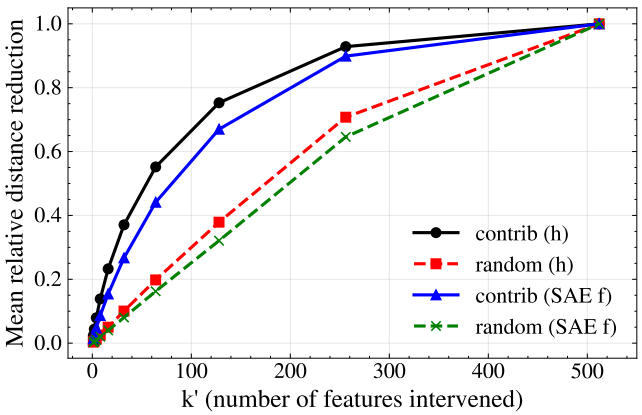


**Fig. 2**: Counterfactual intervention analysis. Contribution-based ranking (solid lines) closes the anomaly-to-normal distance substantially faster than random feature selection (dashed lines), in both dense reconstruction ($h$) and sparse embedding spaces (SAE $f$).

the explained distance collapses sharply to values slightly below zero. Negative $R^2_{\text{decomp}}$ values indicate the decomposition performs worse than predicting the mean, rendering attributions unreliable.

The four evaluation conditions, anomaly-vs-normal and normal-vs-normal pairs on both Dev and Eval splits, track one another closely across all sparsity levels, confirming that observed degradation reflects general reconstruction quality rather than sparsity-induced disruption specific to anomalous structure. Results are qualitatively similar across layers, though layer 11 exhibits slightly steeper mid-range degradation than layer 12 (not shown).

Having established 90% sparsity as a reasonable operating point, we now validate the feature-attribution rankings using counterfactual interventions. We find that a sparsity value of $90\%$ strikes a reasonable balance across all three criteria. At this level, Fig. 2 plots the mean relative distance reduction as a function of $k'$, the number of features intervened upon. Across both the dense reconstruction space and the sparse embedding space, contribution-based ranking substantially outperforms random selection. With contribution-based ranking in the reconstruction space, roughly $75\%$ of the distance is closed by intervening on only $\sim$125 features ($\sim$60% of active features), compared to under $40\%$ for random selection in the same space. Distance reduction proceeds faster in the reconstruction space than in the sparse-embedding space for both strategies, mainly because contributions are computed directly in the dense latent space, making them more informative for ranking there. In fact, when contributions are randomly sampled, the gap between reconstruction and sparse latent space becomes smaller. All four curves converge to a reduction of 1.0 by $k' \geq 2k$, confirming that full distance closure is always achievable.

### 5.3. Semantic Probing

Figure 1 (right panel) shows the probing F1 drop relative to the dense BEATs baseline, averaged across all machines. Probes are trained on normal recordings using majority voting across frequency bands; concepts with fewer than 2 classes and classes with fewer than 2 samples are excluded; class-weighted cross-entropy is used to address label imbalance. We observe a steady degradation as sparsity increases, from $-3\%$ at $75\%$ sparsity to $-7.7\%$ at $95\%$ sparsity, followed by a sharp drop to $-12\%$ at $99\%$ sparsity. The loss in the final interval alone is comparable in magnitude to the cumulative degradation across all preceding sparsity levels. The $75\%$–$95\%$ range thus constitutes a graceful degradation regime in which substantial compression is achievable with only modest semantic cost. This contrasts with [24], where sparsity did not affect probing performance in vocal-technique classification. Whether this reflects task-specific factors, domain characteristics, or differences in the underlying encoder warrants further investigation.

## 6. CONCLUSION

We proposed a framework for interpretable anomaly detection in industrial machine sounds that attributes anomaly scores to sparse, semantically meaningful features. Experiments on the DCASE 2025 data show that the SAEs are comparable to state-of-the-art pipelines in detection performance. In the 75%–95% sparsity regime, $R^2_{\text{decomp}}$ remains above 0.6, and counterfactual interventions confirm that contribution-ranked features close roughly $75\%$ of the anomaly-to-normal distance by modifying only $\sim 61\%$ of active features, substantially outperforming random selection. Linear probes further confirm that sparse features preserve the semantic structure of BEATs embeddings with only modest degradation. Future work includes extending the framework to other encoders and anomaly scoring functions, and evaluating explanation quality through human-centered studies with industrial operators.

## REFERENCES

[1] C. C. Aggarwal, *Outlier Analysis*. Springer, 2013. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-6396-2

[2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda *et al.*, "Description and discussion on DCASE2020 Challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2020.

[3] 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. DCASE*, 2023.

[4] 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 2024 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2024.

[5] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit *et al.*, "Description and discussion on DCASE 2025 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2025.

[6] K. Wilkinghoff, T. Fujimura, K. Imoto, J. Le Roux, Z.-H. Tan, and T. Toda, "Handling domain shifts for anomalous sound detection: A review of DCASE-related work," in *Proc. DCASE*, 2025.

[7] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023.

[8] S. Bharadwaj, S. Cornell, K. Choi, S. Fukayama, H.-j. Shim, S. Deshmukh, and S. Watanabe, "Openbeats: A fully open-source general-purpose audio encoder," in *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2025, pp. 1–5.

[9] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. IJCAI*, 2024.

[10] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 608–622, 2023.

[12] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Keeping the balance: Anomaly score calculation for domain generalization," in *Proc. IEEE ICASSP*, 2025.

[13] P. Saengthong and T. Shinozaki, "Deep generic representations for domain-generalized anomalous sound detection," in *Proc. IEEE ICASSP*, 2025.

[14] P. Saengthong and T. Shinozaki, "Genrep for first-shot unsupervised anomalous sound detection of dcase 2025 challenge," DCASE2025 Challenge, Tech. Rep., June 2025.
[15] L. Wang, "Pre-trained model enhanced anomalous sound detection system for dcase2025 task2," DCASE2025 Challenge, Tech. Rep., June 2025.
[16] A. Makhzani and B. Frey, "K-sparse autoencoders," 2014.
[17] L. Gao, T. D. l. Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu, "Scaling and evaluating sparse autoencoders," in *ICLR*, 2025.
[18] D. Braun, J. Taylor, N. Goldowsky-Dill, and L. Sharkey, "Identifying functionally important features with end-to-end sparse dictionary learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 286–107 325, 2024.
[19] M. Pach, S. Karthik, Q. Bouniot, S. Belongie, and Z. Akata, "Sparse autoencoders learn monosemantic features in vision-language models," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
[20] S. Han, Y. Kim, and N. Kwak, "Causal interpretation of sparse autoencoder features in vision," *arXiv preprint arXiv:2509.00749*, 2025.
[21] C.-T. Chou, G. Liu, J. Sun, C. Blondin, K. Zhu, V. Sharma, and S. O'Brien, "Causal Language Control in Multilingual Transformers via Sparse Feature Steering," Jul. 2025, arXiv:2507.13410 [cs].
[22] S. Chalnev, M. Siu, and A. Conmy, "Improving Steering Vectors by Targeting Sparse Autoencoder Features," 2024.
[23] S. Soo, C. Guang, C. Balaganesh, W. Teng, T. Guoxian, and Y. Ming, "INTERPRETABLE STEERING OF LARGE LANGUAGE MODELS WITH FEATURE GUIDED ACTIVATION ADDI-," 2025.
[24] D. Pluth, Y. Zhou, and V. K. Gurbani, "Sparse Autoencoder Insights on Voice Embeddings."
[25] T. Mariotte, M. Lebourdais, A. Almudévar, M. Tahon, A. Ortega, and N. Dugué, "Sparse Autoencoders Make Audio Foundation Models more Explainable," Sep. 2025, arXiv:2509.24793 [cs]. [Online]. Available: http://arxiv.org/abs/2509.24793
[26] G. Aparin, T. Sadekova, A. Rukhovich, A. Yermekova, L. Kushnareva, V. Popov, K. Kuznetsov, and I. Piontkovskaya, "AudioSAE: Towards Understanding of Audio-Processing Models with Sparse AutoEncoders," Feb. 2026, arXiv:2602.05027 [cs]. [Online]. Available: http://arxiv.org/abs/2602.05027