JOINT DOMAIN-ADVERSARIAL AND CONTRASTIVE LATENT OPTIMIZATION FOR UNSUPERVISED AUDIO ANOMALY DETECTION

Technical Report

Taharim Rahman Anon[†], *Jakaria Islam Emon*

Hokkaido Denshikiki Co., Ltd.
Sapporo, Hokkaido, Japan.
{tahrim.anon21@gmail.com, emon_j@hdks.co.jp}

ABSTRACT

This paper presents a unified framework for Unsupervised Anomaly Sound Detection (UASD) that combines Convolutional Autoencoders (CAE) with Domain-Adversarial Neural Networks (DANN) and Deep Support Vector Data Description (Deep SVDD). Our approach addresses the critical challenges of domain shift and first-shot generalization in the DCASE 2025 Task 2 challenge. The proposed architecture employs a CAE to learn compact latent representations while a domain classifier with gradient reversal enforces domain-invariant features. The latent space is simultaneously optimized using Deep SVDD to create a tight hypersphere around normal samples. Unlike traditional reconstruction-based methods, our approach leverages both reconstruction loss and a contrastive SVDD loss that pushes generated pseudo-outliers from the normal data boundary, combined with adversarial domain adaptation. Our system demonstrates superior performance over the DCASE 2025 autoencoder baseline, with achieving a total score of 0.77 (versus baseline 0.65). The domain-adversarial training significantly improves target domain generalization, establishing the efficacy of joint optimization for robust anomaly detection in dynamic acoustic environments.

***Index Terms*—** audio anomaly detection, domain-adversarial training, convolutional autoencoder, deep SVDD, unsupervised learning

1. INTRODUCTION

Predictive maintenance using acoustic signals is a cornerstone of modern industrial monitoring and Industry 4.0 [1, 2]. Although Unsupervised Anomaly Sound Detection (UASD) systems are designed to be effective in real-world settings, their performance is hampered by two persistent and intertwined challenges: the rarity of fault data and the problem of domain shift.

First, anomalous sounds corresponding to machine faults are, by nature, infrequent and diverse. This scarcity makes it impractical to collect a comprehensive dataset of all possible failure modes, rendering traditional supervised classification methods ineffective. Second, a model trained in a controlled source environment will invariably encounter domain shift when deployed in a new target environment. Variations in machine load, operating speed, component mounting, and ambient noise create distributional shifts in the acoustic data that can drastically degrade a model's performance.

The DCASE Task 2 challenge series reflects the research community's progressive efforts to tackle these issues, evolving from plain inlier modeling (2020) [3, 4], through domain adaptation (2021) [5, 6, 7], to the current, more demanding machine generalization [8] and first-shot detection scenarios (2022-2025) [9, 8, 10]. Prevailing strategies in the field, including successful approaches like [11, 12], generally follow three main lines of work. The foundational approach relies on autoencoders (AEs) [13, 14]. To improve upon this, Outlier Exposure (OE) [15] explicitly trains a model to recognize anomalousness by using proxy outliers, often sourced from large, external audio corpora. Finally, hybrid ensembles [16, 17] have shown success by fusing the scores of multiple, diverse subsystems.

However, each of these strategies carries inherent limitations. Autoencoders are also sensitive to domain shift. Outlier Exposure, conversely, is heavily dependent on the quality and relevance of the chosen outlier dataset [18]. While ensembles are powerful, they increase computational complexity and can be challenging to deploy and maintain. A framework that can learn domain-invariant features while simultaneously creating a compact, well-defined boundary for normal data within a single model remains a key research gap.

This paper closes that gap by proposing a unified architecture that combines a Convolutional Autoencoder (CAE) with Domain-Adversarial Neural Networks (DANN) and the one-class objective of Deep SVDD. We tackle the challenging machine generalization task by training a single, robust model on a diverse set of machine data.

Our main contributions are as follows:

1. We propose a unified framework for a single, machine-generalized model that is jointly regularized by DANN for domain invariance and Deep SVDD for creating a compact, hyperspherical class boundary for normal data.
2. We demonstrate that our hybrid-objective CAE,leveraging a stable and structured alternating training strategy, substantially outperforms both a machine-specific baseline and the embedding-centric generalized architecture (AudioMamba)[19].
3. We validate that the hybrid-objective CAE-DANN model achieves superior generalization performance in the machine-agnostic setting. With a TOTAL (Ω) score of **0.77**, it significantly outperforms both the embedding-centric AudioMamba [20] model (0.61) and the machine-specific *DCASE 2025* autoencoder baseline (0.65).

[†]Work done during internship at Hokkaido Denshikiki Co., Ltd.

The rest of this paper is organized as follows: Section 2 details the proposed method, Section 3 describes the experimental configuration, and Section 4 presents our main results and discussion. Finally, Section 5 concludes the paper and suggests future directions.

2. METHOD

Our method employs a unified architecture that combines three complementary objectives: reconstruction-based representation learning, pseudo Outlier Exposure(pOE), and domain-invariant feature learning. Given log-mel spectrogram $X \in \mathbb{R}^{C\times H\times W}$ (C=1), we process it through a convolutional autoencoder that simultaneously optimizes for accurate reconstruction of normal samples and compact latent representations.

2.1. Convolutional Autoencoder Architecture

The Convolutional autoencoder (CAE) is composed of a symmetric encoder-decoder structure, mapping an input spectrogram X to a latent representation z and then to a reconstructed spectrogram $\hat{X}$.

The encoder, denoted as a function $\text{Enc}(\cdot)$, consists of four convolutional blocks. Each block, indexed by $l \in \{1, 2, 3, 4\}$, applies a 2D convolution (Conv_l), followed by batch normalization (BN_l) and a ReLU activation function. This sequence progressively downsamples the spatial dimensions while increasing the channel depth ($1 \to 16 \to 32 \to 64 \to 128$). The transformation can be expressed as:

$$H_l = \text{ReLU}(\text{BN}_l(\text{Conv}_l(H_{l-1}))) \tag{1}$$

where $H_0 = X$ is the input spectrogram, and the final encoder output is the feature map $H_4 = \text{Enc}(X)$.

The bottleneck operates on the flattened output of the final convolutional layer, $h_{\text{flat}} = \text{Flatten}(H_4)$. A fully connected linear layer then maps this high-dimensional feature map to the latent representation $z \in \mathbb{R}^d$:

$$z = W_{\text{enc}} h_{\text{flat}} + b_{\text{enc}} \tag{2}$$

where W_{enc} and b_{enc} are the weights and bias of the encoding linear layer.

The decoder, denoted $\text{Dec}(\cdot)$, mirrors the encoder's architecture. It first projects the latent vector z back to the dimensionality of the flattened feature map, which is then reshaped to its 2D spatial form $\hat{H}_4$:

$$\hat{h}_{\text{flat}} = W_{\text{dec}} z + b_{\text{dec}} \quad \text{and} \quad \hat{H}_4 = \text{Reshape}(\hat{h}_{\text{flat}}) \tag{3}$$

The core of the decoder is a sequence of four transposed convolutional blocks that progressively upsample the feature map and reduce its channel depth ($128 \to 64 \to 32 \to 16 \to 1$). The final layer employs a 'Sigmoid' activation function to produce the reconstructed spectrogram $\hat{X}$:

$$\hat{X} = \text{Dec}(\hat{H}_4) = \text{Sigmoid}(\text{final\_block}(\dots)) \tag{4}$$

Each transposed convolutional block comprises a 2D transposed convolution, batch normalization, and a ReLU activation, ensuring a symmetric reconstruction of the original input's shape.

2.2. Deep SVDD Optimization in Latent Space

Following the Deep SVDD framework [21], our training objective includes a loss term designed to map the latent representations z of normal samples into a minimal-volume hypersphere, defined by a center $c \in \mathbb{R}^d$. The center c is initialized as the mean of embeddings from normal training samples in an initial forward pass.

The Deep SVDD loss term consists of two parts. The first part penalizes the distance of normal samples from the center, encouraging them to be compact. The second part, acting as a form of in-domain outlier shaping, leverages the known abnormal samples from the training data. It pushes their representations away from the center c, beyond a specified margin ν. This enforces a structured latent space where the boundary between normal and anomalous classes is explicitly defined.

This is formally expressed as:

$$\mathcal{L}_{\text{SVDD}} = \frac{1}{|I_n|}\sum_{i\in I_n} \|z_i - c\|^2 + \frac{1}{|I_a|}\sum_{i\in I_a} \max(0, \nu - \|z_i - c\|^2) \tag{5}$$

where I_n and I_a are the sets of normal and abnormal samples in the batch, respectively. This joint objective trains the encoder to not only learn the distribution of normal data but also to actively separate it from the distributions of known fault conditions, leading to a more discriminative latent space.

2.3. Domain-Adversarial Feature Learning

To improve the model's generalization capability between the source and target domains, we incorporate a domain adaptation mechanism based on the Domain-Adversarial Neural Network (DANN) framework.

The domain classifier, D, is a feed-forward network composed of fully connected layers, batch normalization, ReLU activation, and dropout. Crucially, a Gradient Reversal Layer (GRL) [22] is placed between the encoder's output and the domain classifier's input. The forward pass through the classifier, which takes the GRL-modified latent vector as input, is given by:

$$D(z) = W_2 \cdot \text{ReLU}(\text{BN}(W_1 \cdot \text{GRL}_\lambda(z) + b_1)) + b_2 \tag{6}$$

The GRL is the key to adversarial training. In the forward pass, it acts as an identity function. In the backward pass, it multiplies the incoming gradient by a negative scalar, $-\lambda$. This behavior is formally defined by its gradient with respect to the adversarial loss, $\mathcal{L}_{\text{adv}}$:

$$\frac{\partial \mathcal{L}_{\text{adv}}}{\partial z} = -\lambda \frac{\partial \mathcal{L}_{\text{adv}}}{\partial \text{GRL}(z)} \tag{7}$$

This gradient reversal creates a conflicting objective: the domain classifier D is trained to correctly predict the domain, while the encoder is simultaneously trained to produce features that confuse the classifier. This adversarial dynamic encourages the encoder to learn features that are not only useful for the primary anomaly detection task but are also indistinguishable between the source and target domains, thus promoting robust generalization.

2.4. Joint Training and Inference

The training of our unified model follows the principles of Domain-Adversarial Neural Networks, employing an alternating, two-step update scheme within each training batch. This training procedure, detailed in Algorithm 1, is essential for stabilizing the adversarial dynamic between the feature encoder and the domain classifier.

In the first step, we update only the parameters of the domain classifier, θ_D. The classifier is trained to minimize the standard cross-entropy loss for domain prediction, $\mathcal{L}_D$. Crucially, this is done using latent features, z, that are detached from the encoder's

Algorithm 1 Joint Training Procedure

Require: Dataset $\mathcal{D}$, margin ν, weights $\alpha_{\text{recon}}, \lambda$
Initialize CAE parameters θ_{CAE}, domain classifier θ_D
Initialize SVDD center c using normal sample embeddings from $\mathcal{D}$
for epoch = 1 to N_{epochs} **do**
  for each batch (X, d, y) in $\mathcal{D}$ **do**
    // — Step 1: Update Domain Classifier —
    $z, \_ \leftarrow \text{CAE}(X; \theta_{\text{CAE}})$ ▷ Get latent embeddings
    $\mathcal{L}_D \leftarrow \text{CrossEntropy}(D(z.\text{detach}(); \theta_D), d)$
    Update θ_D via gradient descent on $\mathcal{L}_D$
    // — Step 2: Update Autoencoder —
    $z, \hat{X} \leftarrow \text{CAE}(X; \theta_{\text{CAE}})$ ▷ Forward pass with gradients
    $\mathcal{L}_{\text{recon}} \leftarrow \|X_y - \hat{X}_y\|^2$ for samples where $y = 0$
    $\mathcal{L}_{\text{SVDD}} \leftarrow$ Eq. 5 using z and y
    $\mathcal{L}_{\text{adv}} \leftarrow \text{CrossEntropy}(D(\text{GRL}_\lambda(z)), d)$
    $\mathcal{L}_{\text{total}} \leftarrow \alpha_{\text{recon}}\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{SVDD}} + \mathcal{L}_{\text{adv}}$
    Update θ_{CAE} via gradient descent on $\mathcal{L}_{\text{total}}$
  end for
  Evaluate on validation set and save best model based on performance
end for
return Best performing model parameters θ_{CAE} and SVDD center c

computation graph, ensuring that the gradients are only used to improve the classifier and do not affect the encoder.

In the second step, the domain classifier's parameters are frozen, and the autoencoder's parameters, θ_{CAE}, are updated. The encoder is trained to minimize a hybrid objective function while simultaneously trying to maximize the domain classifier's loss via the Gradient Reversal Layer. The complete objective function for this step is a weighted sum of the reconstruction, Deep SVDD, and adversarial losses:

$$\mathcal{L}_{\text{total}} = \alpha_{\text{recon}}\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{SVDD}} + \mathcal{L}_{\text{adv}} \tag{8}$$

Here, $\mathcal{L}_{\text{recon}}$ is the mean squared error computed only for normal samples, $\mathcal{L}_{\text{SVDD}}$ is the one-class hypersphere loss from Eq. 5, and $\mathcal{L}_{\text{adv}}$ is the adversarial domain classification loss regularized by the GRL's hyperparameter λ. The weights α_{recon} and λ control the trade-off between reconstruction fidelity and the two regularization terms.

An important aspect of our methodology is the distinction between the objectives for training and the metric for inference. While the Deep SVDD and DANN losses are crucial regularizers, our empirical validation showed that the mean squared reconstruction error provided a more stable and effective anomaly score for the final evaluation. Therefore, at inference time, the anomaly score for a test sample X is computed as:

$$\text{score}(X) = \frac{1}{C \cdot H \cdot W}\|X - \hat{X}\|^2 \tag{9}$$

3. EXPERIMENTAL SETUP

3.1. Data

Our experiments utilize the DCASE 2022-2025 Task 2 corpus [4, 3]. The training dataset for our single, machine-generalized model was constructed by creating a comprehensive pool of normal operational data. This pool includes:

1. **Core Normal Data:** All normal audio clips from the seven primary machine types (*ToyCar, Fan, etc.*), combining samples from both source and target domains across the official 'train' directories.
2. **Additional Training Data:** All clips from the provided "additional_training_data" set, which contains further examples of source-domain operational sounds.

From this large pool of normal data, we generated our training anomalies using Synthetic Anomaly Augmentation. To do this, we followed the principles of pseudo Outlier Exposure [15] by taking a subset of the normal clips from both source and target domains and applying corrupting transformations, such as random noise bursts and frequency shifts. These augmented clips were then labeled as "anomaly" and constitute the entire set of anomalies used during training.

The final training dataset, consisting of the original normal data pool and the synthetic anomalies, then undergoes stratified oversampling. This final step balances the classes (normal, synthetic anomaly) and domains (source, target), ensuring each category is equally represented to prevent model bias. Model performance was evaluated using the official DCASE 2025 Task 2 test sets.

3.2. Implementation Details

Our framework was implemented in PyTorch and trained on NVIDIA T4 and RTX 3090 GPUs. The specific hyperparameters for the proposed CAE-DANN model are as follows:

- **Architecture**: The latent dimension was set to $d = 128$. The domain classifier used a hidden layer of 64 units with a dropout rate of 50%.
- **Optimizer**: We used the AdamW optimizer for both the main model and the domain classifier, with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}.
- **Loss Weights**: The hyperparameters for the joint loss function (Eq. 8) were set to $\alpha_{\text{recon}} = 0.5$ for the reconstruction weight, $\nu = 3.0$ for the Deep SVDD margin, and $\lambda = 0.1$ for the GRL adversarial weight.
- **Training**: Models were trained for 20 epochs with a batch size of 256. We employed a learning rate scheduler that reduced the learning rate by a factor of 0.5 if the validation AUC did not improve for 5 consecutive epochs.

4. RESULTS AND DISCUSSION

This section presents a comprehensive performance analysis of our proposed systems against the official baseline. We evaluate three distinct models:

1. Our primary Proposed (CAE-DANN) system, which uses a convolutional autoencoder with a hybrid training objective to learn a single, generalized model.
2. An embedding-based AudioMamba (AuM) system. This was also trained as a single, machine-generalized model incorporating GRL and an SVDD Outlier Exposure (OE) strategy.
3. The official DCASE 2025 Autoencoder (AE) baseline.

Table 1: Performance comparison of the proposed machine-generalized CAE-DANN system, a machine generalized AudioMamba (AuM) system, and DCASE 2025 AE baseline.

	Proposed (CAE-DANN)			**AudioMamba (AuM)**			**Baseline (AE)**		
Machine	**AUC-S**	**AUC-T**	**pAUC**	**AUC-S**	**AUC-T**	**pAUC**	**AUC-S**	**AUC-T**	**pAUC**
ToyCar	0.781	0.670	0.654	0.820	0.810	0.630	0.790	0.725	0.741
ToyTrain	0.891	0.910	0.836	0.870	0.870	0.740	0.673	0.598	0.602
Fan	0.653	0.694	0.675	0.580	0.500	0.510	0.644	0.338	0.503
Gearbox	0.844	0.883	0.508	0.610	0.490	0.490	0.702	0.653	0.664
Bearing	0.950	0.934	0.858	0.710	0.570	0.540	0.711	0.600	0.643
Slider	0.906	0.937	0.703	0.500	0.680	0.530	0.703	0.575	0.621
Valve	0.868	0.851	0.614	0.570	0.510	0.500	0.650	0.611	0.593
hmean	0.836	0.821	0.688	0.643	0.603	0.552	0.681	0.614	0.628
amean	0.842	0.840	0.693	0.666	0.633	0.563	0.702	0.631	0.646
TOTAL (Ω)		**0.778**			**0.610**			**0.650**	

A comprehensive analysis of performance is presented in Table 1.

4.1. Analysis of Generalized Models

A key outcome of our comparative analysis is the superior performance of our Proposed (CAE-DANN) architecture in the machine-generalized setting. With a TOTAL (Ω) score of 0.778, it not only surpasses the more AudioMamba architecture (0.610) but also significantly outperforms the Autoencoder baseline (0.650). This indicates that the CAE-DANN, when trained on a diverse dataset from all machines, successfully learns robust, generalizable features that allow it to effectively detect anomalies across different machine types. Its strong and consistent scores across nearly all machines validate the power of its hybrid training objective for generalization.

In contrast, the AudioMamba (AuM) system exhibits struggles in this generalized setting. Despite being a modern architecture combined with established GRL and OE techniques, its overall performance is the lowest of the three. Its inconsistency performing well on 'ToyTrain' but poorly on 'Fan', 'Gearbox', and 'Slider' suggests that its learned embedding space does not generalize as effectively across the high variance of the different machine acoustics.

The key takeaway is that for this task, the training strategy and scoring method are more critical than the raw complexity of the model backbone. The success of the CAE-DANN architecture demonstrates that a carefully designed hybrid loss is highly effective at producing a single, robust model for multiple machine types.

4.2. Discussion on Scoring Method

The decision to use reconstruction error for inference, despite the heavy use of a latent-space SVDD loss during training, is a critical element of our system's success. We suggest two reasons for this. First, the reconstruction score is a global property of the entire model and can be more robust than a distance metric tied to a single center point 'c'. Second, and more importantly, the SVDD and DANN losses act as powerful regularizers that force the autoencoder to learn an exceptionally precise manifold of normal data. Consequently, the model's reconstruction fidelity becomes an even stronger and more discriminative signal of normalcy, which proves highly effective in a generalized setting.

4.3. Complexity Analysis

The inference-time complexity of our proposed CAE-DANN system is highly efficient. Since the domain classifier and SVDD loss are only used during training, inference requires only a single forward pass through the lightweight CAE. With 17.04 M parameters, the model requires approximately 0.61 GMACsfor a standard 10-second input (a 128×1024 spectrogram). This computational efficiency makes the model well-suited for deployment on edge devices and embedded GPUs for real-time monitoring applications.

5. CONCLUSION AND FUTURE WORK

In this work, we introduced a unified framework combining a Convolutional Autoencoder (CAE) with Domain-Adversarial Neural Networks (DANN) and a Deep SVDD training objective. Our key finding is that this approach, when trained as a single machine-generalized model, achieves state-of-the-art performance, outperforming not only the Auto Encoder baseline but also more complex embedding centric architectures. This success highlights that a carefully designed hybrid training objective which jointly optimizes for reconstruction fidelity, latent space structure, and domain invariance can be more critical for generalization than the raw complexity of the model backbone. Furthermore, we validated that using reconstruction error as the final inference score is a robust and effective strategy, particularly when the model has been regularized with powerful latent-space objectives during training.

Building directly on our findings, several avenues for future research appear promising. The most immediate direction is the fusion of the two available anomaly scores. Since our model computes both the reconstruction error and the latent-space SVDD distance, combining these, for instance through a normalized weighted average, could yield a more robust detector by capturing different facets of anomalousness. Additionally, while our CAE captures spatial patterns in the spectrogram, explicitly modeling temporal dependencies with lightweight attention or recurrent mechanisms could further enhance performance on anomalies that manifest as evolving patterns.

Acknowledgment

This work was supported by Hokkaido Denshikiki Co., Ltd.

6. REFERENCES

[1] A. Ucar, M. Karakose, and N. Kırımça, "Artificial intelligence for predictive maintenance applications: Key components, trustworthiness, and future trends," *Applied Sciences*, vol. 14, no. 2, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/2/898

[2] M. Islam, J. Emon, K. Ng, A. Asadpour, M. Aziz, M. Baptista, and J. Kim, *Artificial Intelligence in Smart Manufacturing: Emerging Opportunities and Prospects*, ser. Springer Series in Advanced Manufacturing. United States: Springer Nature, Mar. 2025, pp. 9–36, publisher Copyright: © The Author(s), under exclusive license to Springer Nature Switzerland AG 2025.

[3] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, November 2019, pp. 209–213. [Online]. Available: http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Purohit_21.pdf

[4] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, November 2019, pp. 308–312. [Online]. Available: https://ieeexplore.ieee.org/document/8937164

[5] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 186–190.

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

[7] 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 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 1–5.

[8] 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.

[9] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2506.10097*, 2025.

[10] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," 09 2023, pp. 191–195.

[11] K. Wilkinghoff, T. Fujimura, K. Imoto, and J. Le Roux, "Handling domain shifts for anomalous sound detection: A review," in *Proc. 51st Annual Meeting on Acoustics*. Deutsche Gesellschaft für Akustik, DEGA, Apr. 2025, pp. 101–104.

[12] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 276–280.

[13] NTT Corporation, "DCASE 2025 challenge task 2 and DCASE 2023 challenge task 2 baseline auto encoder: dcase2023_task2_baseline_ae," GitHub Repository, 2024. [Online]. Available: https://github.com/nttcslab/dcase2023_task2_baseline_ae

[14] K. Rai, F. Hojatpanah, F. Ajaei, and K. Grolinger, "Deep learning for high-impedance fault detection: Convolutional autoencoders," *Energies*, vol. 14, 06 2021.

[15] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *Proceedings of the International Conference on Learning Representations*, 2019.

[16] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

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

[18] Y. Tachioka, "Outlier exposure with efficient division of positive and negative examples for anomalous sound detection," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024, pp. 76–80.

[19] M. H. Erol, A. Senocak, J. Feng, and J. S. Chung, "Audio mamba: Bidirectional state space model for audio representation learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.03344

[20] M. H. Erol, A. Şenocak, J. Feng, and J. S. Chung, "Audio mamba: Bidirectional state space model for audio representation learning," *IEEE Signal Processing Letters*, vol. 31, pp. 2215–2219, 2024.

[21] Z. You, Y. Zhou, T. Yang, and W. Fan, "Anomaly-injected deep support vector data description for text outlier detection," *arXiv preprint arXiv:2110.14729*, 2021.

[22] J. Guan, J. Tian, Q. Zhu, F. Xiao, H. Zhang, and X. Liu, "Disentangling hierarchical features for anomalous sound detection under domain shift," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.