forum acusticum 2026
8-12 September · Graz, Austria

MULTI-CHANNEL ANOMALY SOUND DETECTION BASED ON SPATIAL-RESIDUAL LEARNING AND DUAL-ENCODER RECONSTRUCTION

Clara Luzon-Alvarez[1], Francesc J. Ferri[1], Maximo Cobos[1]

[1] *Computer Science Department, Universitat de València, Spain*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Anomalous sound detection (ASD) is a key task in industrial applications, where reliable detection of machine faults can prevent costly downtime and ensure operational safety. Although single-channel recordings simplify data acquisition and model design, they often do not capture spatial information that can be critical for distinguishing between normal and anomalous acoustic events in complex industrial environments. To address this, we propose a multichannel approach leveraging inter-channel correlations. Two autoencoder-based models are presented: in both, one channel serves as a reference, and differences with the remaining channels are computed. Two encoders process the reference and the remaining channels, respectively, while decoder designs vary—one model uses separate decoders for reference and remaining channels, and the other employs a single decoder to reconstruct all channels jointly. Our approach captures richer machine behavior representations, improving robustness and detection performance in real-world industrial settings.

Keywords: *anomalous sound detection, multichannel audio, autoencoders, industrial monitoring, spatial information*

1. Introduction

Anomalous Sound Detection (ASD) is the task of identifying whether an observed acoustic signal deviates from expected or normal behavior. Formally, given an input audio stream, an ASD system aims to determine whether the signal belongs to the distribution of normal operating sounds or corresponds to an abnormal event. ASD has found applications across a wide range of domains. In medical analysis, acoustic signals such as respiratory sounds can be monitored to detect early signs of disease or physiological abnormalities [1]. In traffic control, audio-based systems can identify unusual events such as accidents or honking anomalies in urban environments [2]. Despite these diverse applications, industrial monitoring remains one of the most impactful and well-studied use cases, where reliable detection of machine faults can prevent costly downtime and ensure operational safety [3]. In the context of industrial systems, ASD is typically applied to monitor machinery such as motors, pumps, and fans, where deviations in acoustic patterns often indicate mechanical degradation or failure. A key challenge in this domain is the scarcity of anomalous data: faults are rare by nature, and collecting labeled examples of all possible failure modes is often impractical or prohibitively expensive [4], [5]. As a result, the majority of existing approaches adopt an unsupervised or self-supervised paradigm, where models are trained exclusively on normal operating sounds and anomalies are detected as deviations from the learned distribution. This setup places significant importance on the model's ability to capture the intrinsic structure of normal acoustic behavior while remaining sensitive to subtle deviations. Autoencoders [3] have emerged as a prominent class of models for ASD under this paradigm. By learning to reconstruct normal input signals, autoencoders implicitly model the manifold of normal sounds; anomalous inputs, which do not conform to this manifold, typically result in higher reconstruction errors. Variants such as convolutional autoencoders [6], variational autoencoders [7], and recurrent architectures [8] have been widely explored to better capture temporal and spectral dependencies in audio signals. These models offer a flexible and scalable framework for unsupervised anomaly detection, particularly in scenarios with limited labeled data.

However, a notable limitation in much of the existing literature is the predominant focus on mono-channel audio. While single-channel recordings simplify data acquisition and model design, they often fail to capture

***Corresponding author**: clara.luzon@uv.es.*

spatial information that can be critical for distinguishing between normal and anomalous acoustic events in complex industrial environments.
Recent studies have explored multichannel approaches to address this limitation. For instance, [9] combines dictionary learning with beamforming techniques on microphone arrays to enhance anomaly detection in industrial machinery. Similarly, [10] utilizes multichannel recordings for the joint detection, extraction, and localization of anomalous sounds, while [11] proposes attention-based mechanisms to exploit spatial and temporal dependencies across channels, demonstrating improved performance over mono-channel models. Motivated by these advances, this work adopts a multichannel approach to ASD, leveraging spatial cues and inter-channel correlations to improve detection performance. By incorporating multichannel audio into an autoencoder-based framework, the proposed methods aim to provide a richer representation of machine behavior and enhance robustness in real-world industrial settings.

2. Proposed Method

Our proposed method builds upon our previous autoencoder-based approach [12] for anomalous sound detection, extending it to more effectively leverage multichannel audio. The model is trained exclusively on normal operating data, learning to reconstruct input signals and using reconstruction discrepancies as an indicator of abnormality.

2.1. Feature Extraction

Following our previous procedure, the input is constructed from multichannel audio waveforms. A single channel is selected as the reference (x_{ref}), while the remaining channels are used to compute difference signals (x_{diff}) relative to the reference in the waveform domain.
For both the reference and difference signals, 128-dimensional Mel-based spectral features are extracted using short-time analysis with fixed frame length and hop size. Five consecutive frames are concatenated to form the final input vectors, resulting in a fixed-dimensional representation per time step, both for the reference ($X_{ref} \in \mathbb{R}_+^L$) and difference channels ($X_{diff_n} \in \mathbb{R}_+^L, \quad n = 1, \ldots, N-1$).

2.2. Dual-Encoder Dual-Decoder Model

The first model uses two encoders and two decoders, as shown in Figure 1. One encoder processes the reference channel input vectors (X_{ref}) producing a latent representation z_{ref}, while the second encoder processes the difference input vectors (X_{diff_n}) producing a latent representation Z_{diff_n}. These latent representations from the difference channels are concatenated to form a joint embedding Z_{diff}, which is further concatenated to z_{ref} and provided to both decoders.
To efficiently handle multiple difference channels, the difference inputs are reshaped by merging the channel dimension with the batch dimension before encoding. After encoding, latent representations are reshaped back to recover the original batch and channel structure, preserving channel-wise information while sharing encoder parameters.
The first decoder reconstructs the reference input $\hat{X}_{ref}$, while the second reconstructs the difference input tensor $\hat{\mathbf{X}}_{diff} \in \mathbb{R}_+^{(N-1)\times L}$. The model is trained by minimizing the weighted sum of mean squared errors (MSE) for both outputs.

$$\mathcal{L} = \mathrm{MSE}(X_{ref}, \hat{X}_{ref}) + \alpha \mathrm{MSE}(\mathbf{X}_{diff}, \hat{\mathbf{X}}_{diff}). \quad (1)$$

Following prior work, we introduce a hyperparameter α to control the relative contribution of the reconstruction error associated with the difference signals. This weighting factor allows the model to balance the influence of the reference and difference channels during training. In particular, larger values of α emphasize accurate reconstruction of the difference signals, while smaller values prioritize the reference channel. This is especially important in the multichannel setting, where the number and variability of difference signals may otherwise dominate the optimization objective.

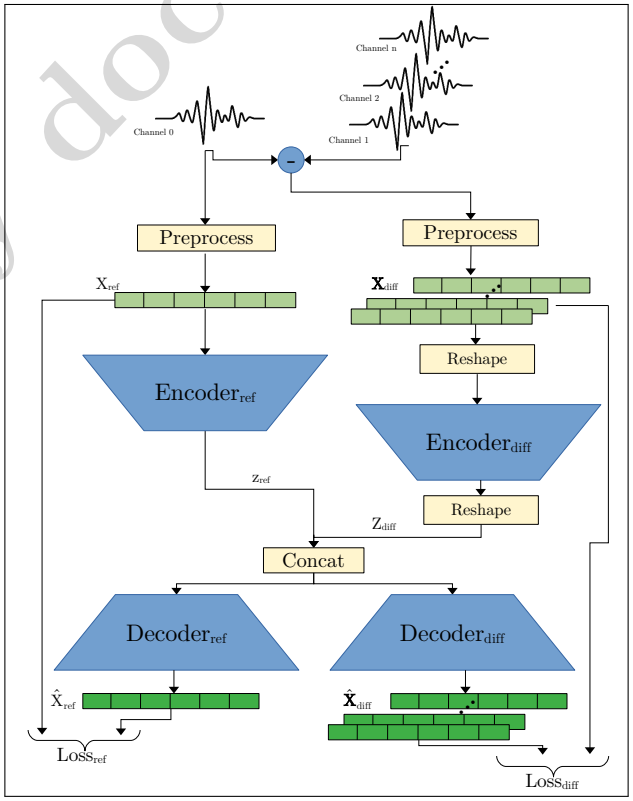

Figure 1: Structure of the dual-encoder dual-decoder model.

2.3. Dual-Encoder Single-Decoder Model

The second model uses two encoders but a single decoder, as shown in Figure 2. The encoder structure and input processing are identical to the dual-decoder model, producing latent representations z_{ref}

and Z_{diff_n} that are concatenated into a joint embedding. A single decoder then reconstructs both the reference and difference inputs jointly.
The training objective is the reconstruction error of all channels without any specific weighting:

$$\mathcal{L} = \text{MSE}(\mathbf{X}, \hat{\mathbf{X}}), \tag{2}$$

where $\mathbf{X} = [X_{ref}, X_{diff_1}, \ldots, X_{diff_{N-1}}] \in \mathbb{R}_+^{N \times L}$ is the tensor formed by concatenating the reference and all difference channels, and $\hat{\mathbf{X}}$ is its corresponding reconstruction.

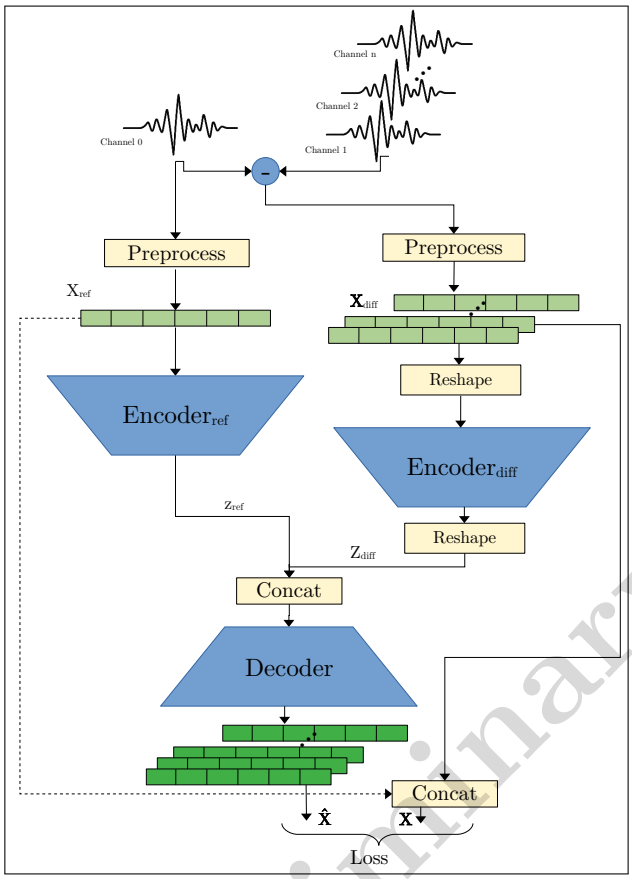

Figure 2: Structure of the dual-encoder single-decoder model.

2.4. Anomaly Scoring and Thresholding

For both models, anomaly detection is based on the reconstruction error, under the assumption that the models trained on normal data will yield higher reconstruction errors for abnormal samples.
Since both the dual-encoder and single-encoder models are trained on their corresponding reconstruction errors, we use their loss functions as an anomaly score $\mathcal{A}(x)$ for any new input x.
In both cases, higher values of $\mathcal{A}(x)$ indicate a greater deviation from normal behavior.
To determine a decision threshold, we assume that anomaly scores for normal data follow a Gamma distribution [3]. The parameters of this distribution are estimated using the anomaly scores computed on the training set. The decision threshold τ is then defined as the p-th percentile of the fitted distribution:

$$\tau = \text{Percentile}_p(A(x)), \tag{3}$$

where p is typically set to 90. A sample is classified as anomalous if $A(x) > \tau$.

3. Experiments and results

3.1. Dataset

We evaluated our method using the MIMII dataset [13], which contains recordings of four types of industrial machines: fans, slide rails, valves, and pumps. Each type includes recordings from multiple individual units, potentially representing different product models. The valves correspond to solenoid valves that are repeatedly opened and closed, the pumps are water pumps that continuously drain and discharge water, the fans are industrial units providing a continuous airflow, and the slide rails represent linear slide systems consisting of a moving platform and a stage base. Each machine type produces sounds with distinct characteristics, which may be stationary or non-stationary, and exhibit varying degrees of detection difficulty.
Recordings were captured as 16-bit audio signals sampled at 16 kHz following the setup of Fig. 3, and were mixed with real industrial background noise at multiple signal-to-noise ratio (SNR) levels. For each segment, background noise was selected randomly and scaled to achieve a target SNR relative to the machine signal. Eight channels were recorded separately using a circular microphone array. The microphones were positioned 50 cm from the machine for all types except the valves, which were recorded at 10 cm. Each recording session captured a single machine along with the ambient factory noise.
In total, the dataset contains 26,092 normal sound segments across all machines. To simulate real-life conditions, a limited number of anomalous segments were included, representing scenarios such as contamination, leakage, rotating unbalance, and rail damage. These anomalous segments were primarily intended for testing, reflecting an unsupervised learning scenario. For training, we randomly selected 1,000 normal samples, combining different individual units and noise levels, while the test set included 200 normal and anomalous audio files per machine type.

3.2. Baseline Model

Our baseline corresponds to our previous work [12], which extended the autoencoder-based approach introduced in the DCASE Task 2 challenge [3] to incorporate spatial information. While the original DCASE baseline operates on a single audio channel, our previous work leveraged two channels recorded from opposite-side microphones: a reference channel and a difference channel computed between the microphones. Both channels were processed using the same feature extraction pipeline as in DCASE, where 128 log-Mel bands were computed using STFT with 64 ms

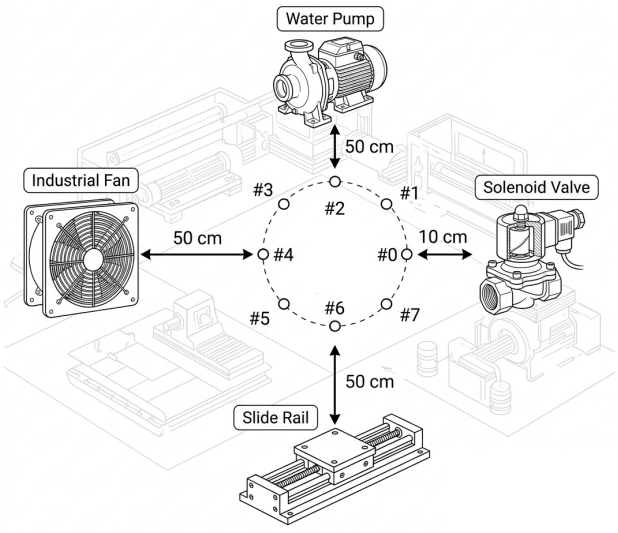

Figure 3: Recording setup of the dataset (as indicated in [13])

frames and 50% overlap, and five consecutive frames were concatenated into 640-dimensional input vectors. The model adopted a dual-branch autoencoder architecture, with separate encoders and decoders for each channel. The resulting latent representations were combined and used to reconstruct both inputs. Similar to our current proposal, training was performed using a weighted sum of reconstruction errors from the two channels, allowing the model to balance spectral and spatial information.

3.3. Evaluation Metrics

To evaluate our method, we adopt the same metrics used in the DCASE Challenge and in most related works [3], [4], [6], [12]. Specifically, we report the area under the receiver operating characteristic curve (AUC), which is computed as:

$$\text{AUC} = \frac{1}{N_- N_+} \sum_{i=1}^{N_-} \sum_{j=1}^{N_+} \mathcal{H}\Big(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)\Big), \quad (4)$$

where N_+ and N_- denote the number of anomalous and normal test samples, respectively, and $\mathcal{H}(\cdot)$ is the Heaviside step function, which returns 1 if its argument is positive and 0 otherwise. Here, $\mathcal{A}_\theta(x)$ is the anomaly score computed from the reconstruction error.

3.4. Implementation

Our proposed model is trained separately for each machine type, meaning a dedicated model is optimized using data specific to that machine category. For each machine type, both our model and the baseline model are trained on 1,000 raw waveform signals. To compute the log-Mel spectrograms, we use a frame size of 1,024 samples with 50% overlap, and apply a Mel filter bank with 128 filters. Model training is performed using the Adam optimizer [14] with a learning rate of 0.001. Training is carried out for 100 epochs with a batch size of 256.

3.5. Performance comparison

This section presents the experimental evaluation of the proposed models. We first compare the two architectural variants, followed by an analysis of the effect of the weighting parameter α. Finally, we study the impact of the number of input channels.

3.5.1. Comparison of Model Architectures

We begin by comparing the performance of the dual-encoder dual-decoder model and the dual-encoder single-decoder model. Tables 1 and 2 report the anomaly detection performance (AUC) for different machines and latent space sizes. All results in this section are obtained using the full set of available input channels (8 channels) and with α=1.

Overall, both architectures achieve competitive performance across all machines and consistently outperform the baseline reported in Forum 2025. The results indicate that incorporating more multichannel information through difference signals is beneficial regardless of the decoder configuration.

Comparing the two models, no single architecture uniformly dominates across all settings. The dual-decoder model achieves the best performance for *Slider*, reaching an AUC of 98.66 for a latent size of 64, while also providing slight improvements for *Pump*. In contrast, the single-decoder model performs better for *Fan* and achieves the highest score for *Valve* at larger latent dimensions.

Regarding the latent space size, both models tend to perform best with smaller or moderate latent dimensions (16 or 64). Increasing the latent size to 128 does not consistently improve performance and, in some cases, leads to a degradation, particularly for *Pump*. This suggests that overly large latent spaces may reduce the model's ability to capture compact representations of normal behavior.

Finally, performance varies significantly across machines. *Slider* consistently achieves the highest AUC values across all configurations, indicating that its anomalies are easier to detect. In contrast, *Valve* remains the most challenging machine, with relatively lower AUC values, suggesting more subtle or variable anomaly patterns.

These results highlight that both architectural choices are viable, with their relative performance depending on the machine type and latent space configuration.

3.5.2. Effect of the Weighting Parameter α

Following with our previous work, in the case of the two-decoders model, we tested the impact of setting the α parameter into the performance. Tables 4 and 3 present the results obtained per machine with latent space length of 16 and 64, respectively. Overall, the performance is relatively robust to the choice of α. *Fan* and *Pump* show slightly better results with smaller α, whereas *Valve* and *Slider* perform marginally better with larger α, but these variations are minimal. This indicates that the specific value of α is not critical for

Table 1: Detection performance (AUC %) for different latent space sizes (K) using the single-decoder model.

Machine	$K = 16$	$K = 64$	$K = 128$	[12] ($K = 16$)	Mono [3] ($K = 16$)
Fan	**96.00**	95.30	94.53	94.03	86.98
Valve	59.50	59.25	**59.94**	58.24	55.13
Slider	97.80	98.40	**98.49**	96.00	92.24
Pump	**78.29**	73.32	72.65	75.79	69.10

Table 2: Detection performance (AUC %) for different latent space sizes (K) using the dual-decoder model.

Machine	$K = 16$	$K = 64$	$K = 128$	[12] ($K = 16$)	Mono [3] ($K = 16$)
Fan	95.11	**95.38**	94.38	94.03	86.98
Valve	**59.78**	59.25	57.31	58.24	55.13
Slider	97.17	**98.66**	98.57	96.00	92.24
Pump	**78.63**	69.93	68.94	75.79	92.24

achieving high performance, simplifying the practical tuning of the model.

Table 3: AUC (%) for different α values ($K = 64$).

Machine	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$
Fan	**95.95**	95.38	94.86
Valve	58.65	59.25	**59.36**
Slider	98.13	**98.66**	98.56
Pump	**77.88**	69.93	69.67

Table 4: AUC (%) for different α values ($K = 16$).

Machine	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$
Fan	**96.93**	95.11	95.52
Valve	59.34	**59.78**	58.36
Slider	97.18	97.17	**97.64**
Pump	78.36	**78.63**	76.79

3.5.3. Effect of the Number of Channels

We further investigate the impact of the number of input channels using the dual-decoder model. Tables 5 and 6 report the performance for different channel configurations, where the indices denote the selected microphones in the array described in Figure 3.

Overall, the results show that increasing the number of channels generally improves or maintains high AUC, with the best performance often achieved when all eight channels are used. Importantly, the results also indicate that the specific choice and spatial position of the microphones affect detection performance. For example, in the two-channel configurations, selecting microphones (0,4) versus (2,6) can lead to noticeable differences for some machines, such as *Fan* and *Pump*. Similarly, among four-channel configurations, the sets (0,2,4,6) and (1,3,5,7) yield different results depending on the machine.

Some machines, such as *Fan* and *Slider*, benefit noticeably from including more channels, whereas others, like *Valve* and *Pump*, show only marginal differences across configurations. This indicates that while adding more channels can help, the model remains relatively robust even with fewer channels, simplifying practical deployment when sensor availability is limited.

Table 5: AUC (%) for different channel configurations (latent size = 64, $\alpha = 1$).

Machine	2 (0,4)	2 (2,6)	4 (0,2,4,6)	4 (1,3,5,7)	8
Fan	92.12	88.26	95.04	93.84	**95.38**
Valve	**62.13**	60.00	58.99	58.52	59.25
Slider	97.38	96.74	98.09	97.66	**98.66**
Pump	63.61	69.81	68.65	68.40	**69.93**

Table 6: AUC (%) for different channel configurations ($K = 16$, $\alpha = 1$).

Machine	2 (0,4)	2 (2,6)	4 (0,2,4,6)	4 (1,3,5,7)	8
Fan	92.67	87.26	93.52	94.00	**95.11**
Valve	59.03	55.75	58.80	58.48	**59.78**
Slider	97.10	96.86	97.25	**97.32**	97.17
Pump	73.72	77.15	74.32	78.01	**78.63**

4. Conclusion

In this paper, we extended our previous work by proposing a self-supervised anomalous sound detection approach based on autoencoders with multichannel processing. The proposed method demonstrates improved performance compared to our earlier approach, showing consistent gains across different experimental settings. Overall, the results highlight the importance of exploiting multichannel and spatial information, as they enable the model to better capture complex acoustic patterns. These findings emphasize the relevance of such information for achieving robust and reliable anomalous sound detection in industrial applications, while also providing a practical direction for future system design.

5. Acknowledgments

Authors thank the Spanish Research Agency (AEI) and the European Regional Development Fund (ERDF) for partially funding this research with Grants PID2022-137048OB-C41, PREP2022-000357, funded by MCIN/AEI/10.13039/501100011033 and by the "EU Union NextGenerationEU/PRTR". Authors acknowledge the IFIC Artemisa resources funded by the EU ERDF and Comunitat Valenciana.

References

[1] J. A. Dar, K. K. Srivastava, and A. Mishra, "Lung anomaly detection from respiratory sound database (sound signals)," *Computers in Biology and Medicine*, vol. 164, p. 107 311, 2023, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2023.107311.

[2] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a blstm network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58 043–58 055, 2018. DOI: 10.1109/ACCESS.2018.2872931.

[3] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland: IEEE, Sep. 4, 2023, pp. 191–195, ISBN: 978-94-6459-360-0. DOI: 10.23919/EUSIPCO58844.2023.10289721. Accessed: Feb. 5, 2025.

[4] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, Jan. 2019, ISSN: 2329-9304. DOI: 10.1109/taslp.2018.2877258.

[5] K. Dohi et al., *Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques*, Nov. 22, 2022. DOI: 10.48550/arXiv.2206.05876. arXiv: 2206.05876[cs]. Accessed: Feb. 5, 2025.

[6] M. D. Chinnasamy, M. Sumbwanyambe, and T. S. Hlalele, "Acoustic anomaly detection of machinery using autoencoder based deep learning," in *2024 32nd Southern African Universities Power Engineering Conference (SAUPEC)*, 2024, pp. 1–6. DOI: 10.1109/SAUPEC60914.2024.10445053.

[7] F. Fang, C. Wang, L. Chen, X. Liao, and Z. Huang, "An industrial equipment anomaly detection model with adaptive thresholding based on graph attention networks-diffusion variational autoencoder," *International Journal of Dynamics and Control*, vol. 13, no. 8, p. 282, Jul. 24, 2025, ISSN: 2195-2698. DOI: 10.1007/s40435-025-01790-8.

[8] T.-T.-H. Le, A. A. Adiputra, J. Yun, and H. Kim, "Anomaly detection in industrial machine sounds using high-frequency features and gate recurrent unit networks," *IEEE Access*, vol. 13, pp. 77 165–77 186, 2025. DOI: 10.1109/ACCESS.2025.3565812.

[9] S. Sahu, K. Kumar, A. Majumdar, A. A. Kumar, and M. G. Chandra, "Dictionary learning augmented beamforming for industrial machine inspection with microphone array," *IEEE Sensors Letters*, vol. 8, no. 1, pp. 1–4, Jan. 2024. DOI: 10.1109/LSENS.2023.3344697.

[10] A. Nishikawa, K. Hattori, M. Tanaka, H. Muranami, and H. Nishi, "Anomalous sound detection, extraction, and localization for refrigerator units using a microphone array," in *Proceedings of the 48th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Oct. 2022, pp. 1–6. DOI: 10.1109/IECON49645.2022.9969098.

[11] C. Mok, S. Moon, E. S. Choi, H. G. Shim, and S. B. Kim, "Variable-guided attention for multichannel signal anomaly detection," *IEEE Access*, vol. 13, pp. 170 862–170 875, 2025. DOI: 10.1109/ACCESS.2025.3615243.

[12] C. Luzón-Álvarez, A. M. Torres-Aranda, F. J. Ferri, and M. Cobos, "Exploiting spatial information for anomaly detection in industrial machines using autoencoders," in *Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum / EuroNoise 2025*, Málaga, Spain: European Acoustics Association, Dec. 25, 2025, pp. 4351–4356. DOI: 10.61782/fa.2025.0381. Accessed: Mar. 30, 2026.

[13] H. Purohit et al., *MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection*, Sep. 20, 2019. DOI: 10.48550/arXiv.1909.09347. arXiv: 1909.09347[cs]. Accessed: Feb. 7, 2025.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.