Unsupervised Anomalous Sound Detection Focused on Timbral-Related Features

Ryoya Ogura, Masashi Unoki
Japan Advanced Institute of Science and Technology, Japan
Email:{s2310029, unoki}@jaist.ac.jp

***Abstract*—Using acoustical features related to timbre makes anomalous sound detection (ASD) able to make discrimination logic easily understandable to humans. However, the previous system using these acoustical features was developed as a supervised system, which requires both normal and anomalous sounds for supervised training. To this end, unsupervised learning systems that do not require anomalous sounds for training are necessary for real-world applications. This paper proposes an unsupervised ASD system using timbral-related features. Two novel ideas are used in this system. The first one is taking an outlier exposure approach, i.e., conducting apparent supervised learning by generating pseudo-anomaly timbral-related features from external data. These features are neither too similar nor too different from normal sounds, making them suitable for training the discriminator. The second one is extracting features specifically from intervals where machine sounds occur by using timbral frame selection (TFS). From the results of ASD evaluations, it was found that the proposed system performs comparable to previous supervised learning systems for ASD and outperforms typical unsupervised ASD systems such as Gaussian mixture models. It was also found that TFS improves ASD performance for non-stationary machine sounds.**

***Index Terms*—anomalous sound detection, timbral-related features, sound quality metrics, pseudo-anomaly, timbral frame selection**

I. Introduction

Experienced factory inspectors can determine whether a machine is anomalous by listening to its sounds. However, due to the aging of experienced inspectors, there are issues such as a shortage of successors. Thus, it is necessary to develop anomalous sound detection (ASD) systems to identify machine sounds as normal or anomalous using computers [1]. ASD systems enable inspections that do not rely on experienced inspectors.

Machine abnormalities rarely occur, and it is challenging to collect anomalous sound recordings. Most research has focused on unsupervised ASD, which does not use anomalous sounds for model training [2–5]. In unsupervised ASD, spectrograms are often used as features. A standard system uses the autoencoder as the model. This system is an inlier modeling (IM) approach that uses only normal sounds for model training. Compared with IM, outlier exposure (OE) [6] that uses external data as pseudo-anomalous classes has shown higher performance. Currently, approaches using large-scale pre-trained models [7] and pseudo-labels [8] have been proposed, which are also based on OE.

We previously proposed a system for using timbral-related features, including multiple sound quality metrics (SQMs) rather than spectrograms [9]. Using a support vector machine (SVM) as a discriminator, this system achieves high performance. It is also used to analyze important timbres for anomaly detection, making the discrimination logic more interpretable than previous ASD systems. However, this system uses supervised learning, where both normal and anomalous sounds are used for training. For real-world applications, an unsupervised ASD system is required. Because timbral-related features have lower dimensionality compared to spectrograms, it is challenging to adapt previous unsupervised ASD systems that use spectrograms.

We propose an unsupervised ASD system that uses timbral-related features without requiring anomalous sounds for training. The system uses an OE approach for generating pseudo-anomalous timbral-related features using sounds from various machines other than the target machine type. Pseudo-anomalous timbral-related features are generated so that they are not too similar or different from normal sounds and are suitable for model training. A timbral frame selection (TFS) mechanism is also used to extract timbre-related features only when machine sounds occur, in order to improve performance.

II. Related Work

A. ASD based on auditory perception

Experienced inspectors excel at determining machine anomalies through auditory perception. Thus, ASD based on human auditory perception can be effective beyond just deep learning.

Temporal modulation features using gammatone auditory filterbanks have been proposed for ASD [10]. This approach emphasizes capturing temporal characteristics that are difficult to extract with a log-Mel spectrogram. SQMs have also been used for ASD. SQMs quantify human auditory sensations. Specifically, methods using SQMs, such as loudness and fluctuation strength, have been studied for bearing fault detection [11].

B. ASD focused on timbral-related features

Our previous system is based on auditory perception [9]. This system uses timbral-related features for ASD. The timbral-related features consist of SQMs and short-term features. For SQMs, this system uses five timbral attributes from the timbral models developed by the University of Surrey [12], i.e., boominess, brightness, depth, roughness, and sharpness, quantifying subjective impressions humans perceive from

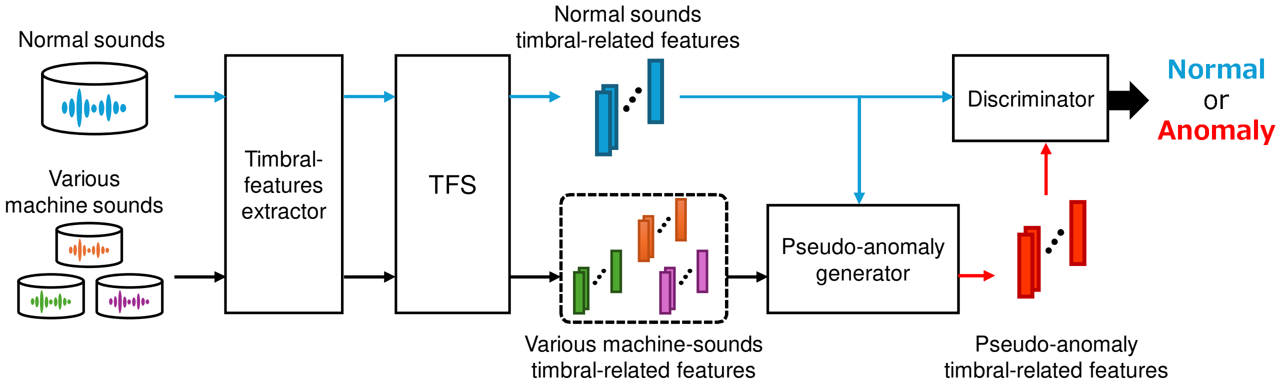
Normal sounds
Various machine sounds
Timbral-features extractor
TFS
Normal sounds timbral-related features
Various machine-sounds timbral-related features
Pseudo-anomaly generator
Pseudo-anomaly timbral-related features
Discriminator
Normal
or
Anomaly

Fig. 1: Training flow of proposed system.

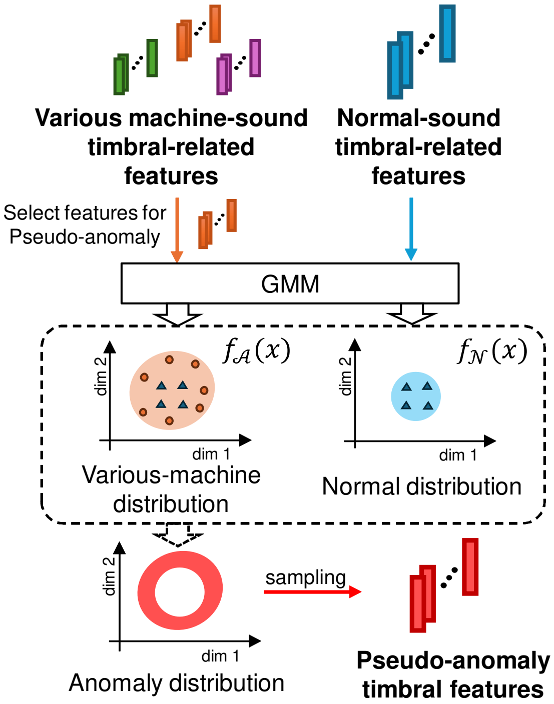
Various machine-sound timbral-related features
Normal-sound timbral-related features
Select features for Pseudo-anomaly
GMM
$f_{\mathcal{A}}(x)$
$f_{\mathcal{N}}(x)$
dim 2
dim 1
Various-machine distribution
Normal distribution
Anomaly distribution
sampling
Pseudo-anomaly timbral features

Fig. 2: Schematic illustration of pseudo-anomaly generator.

sounds. The short-term features consist of two types: amplified shimmer (AS) and amplified predominant frequency (APF), complementing the timbral-related features. AS captures fluctuations in sound amplitude, while APF captures variations in pitch. A timbral-related feature is a seven-dimensional vector.

This system achieved high performance with an F-measure of 0.920 on the MIMII dataset [13]. However, since this is a supervised ASD system, it would be challenging to use in the real world.

III. Proposed System

The training flow of the proposed unsupervised ASD system is shown in Fig. 1. In this section, we explain the pseudo-anomaly generator and timbral frame selection.

A. Pseudo-anomaly generator

The OE approach that uses external data as an anomalous class for training demonstrates high performance. However, there is a tendency for performance to decrease when the sounds of the anomalous class are too similar or too different from the normal sounds of the target machine [14]. It is considered more effective to generate and use data for training that is neither too similar nor too different from the normal sounds of the target machine based on the external data rather than using external data as anomalies.

The pseudo-anomaly generator generates pseudo-anomaly timbral-related features for training based on sound from machines other than the target machine. The flow of the pseudo-anomaly generator is shown in Fig. 2. Various types of normal machine sounds containing S class sounds are prepared in advance.

1) From the S classes, select the i classes whose distributional distance to the target machine is the smallest. The distance of the distribution is the L2 distance between the average timbral-related feature of all samples in each class is denoted as μ_s $(s = 1, \dots, S)$, and that of the target machine as μ_{target}. Define $\mathcal{N}$ as the set that includes only the timbral-related features of the target machine, $\mathcal{A}$ as the set that includes both the timbral-related features of the i selected classes and $\mathcal{N}$.
2) Estimate $f_{\mathcal{N}}(x)$ and $f_{\mathcal{A}}(x)$, the distributions of $\mathcal{N}$ and $\mathcal{A}$, using Gaussian mixture models (GMMs). The anomalous sounds of a target machine can be defined as machine sounds other than the normal sounds of the target machine [15]. Therefore, the distribution of anomalous sounds is approximated by subtracting the distribution of normal sounds, $f_{\mathcal{N}}(x)$, from that of various machine sounds, $f_{\mathcal{A}}(x)$.
3) The pseudo-anomaly is sampled from the distribution of anomalous sounds. Since direct estimation of the distribution obtained by subtracting $f_{\mathcal{A}}(x)$ from $f_{\mathcal{N}}(x)$ is difficult, a candidate pseudo-anomaly $\mathbf{Z}$ is generated from $f_{\mathcal{A}}(x)$. When $\log f_{\mathcal{N}}(\mathbf{Z}) < j \times \log f_{\mathcal{A}}(\mathbf{Z})$ is satisfied, $\mathbf{Z}$ is adopted as a pseudo-anomaly.

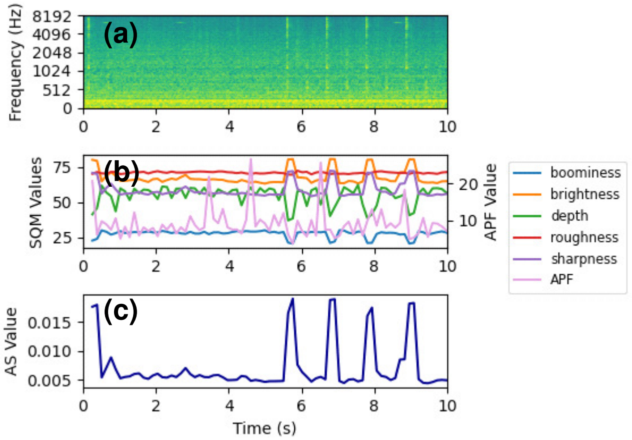

Fig. 3: Time-series data of timbral-related features of valve: (a) spectrogram, (b) SQMs and APF, (c) AS.

Parameter i in 1) is the number of machine classes included in $\mathcal{A}$. Increasing parameter i makes the distribution of $f_{\mathcal{A}}(x)$ larger, so the pseudo-anomalies generated in 3) become more diverse. Parameter j in 3) makes it more difficult to generate a pseudo-anomaly close to normal sounds. A pseudo-anomaly is generated from $f_{\mathcal{A}}(x)$, but if the log density of the generated pseudo-anomaly in $f_{\mathcal{N}}(x)$ is high, it is rejected. Increasing j makes it more likely to be rejected, generating a pseudo-anomaly far from normal sounds. Therefore, it appears that appropriate settings for parameters i and j are important in generating pseudo-anomalies that are neither too similar nor too different from normal sounds.

B. Timbral frame selection

Our previous system uses the average value of timbral-related features calculated from multiple frames of machine sounds. However, frames containing only environmental noise without machine sounds are also included in the averaging for non-stationary machine sounds, such as slider and valve in the MIMII dataset. This can significantly impact performance when the signal-to-noise ratio (SNR) decreases. Wang et al. reported that by selecting only frames containing machine sounds in spectrograms to input the discriminator [16], significant performance improvement can be achieved in non-stationary machine sounds. It is considered adequate to extract timbral-related features from frames containing machine sounds.

Figure 3 shows examples of the log-Mel spectrogram of a valve in the MIMII dataset and temporal changes in timbral-related features. The AS increases with machine sounds, and some SQMs change with the occurrence of machine sounds. Therefore, frames where AS becomes high represent the timbral-related features of the machine sound. SQMs that show changes correlated with AS are considered particularly important for discrimination.

The proposed system uses TFS to extract features from the frames where machine sounds occur for non-stationary sounds. The input to TFS is the time-series data of timbral-related features (SQMs and short-term features) extracted from a single sample. For the classes input to TFS, the correlation coefficient r_m $(m = 1, \ldots, M)$ between AS and the other M types of indicators (TRF$_m$) of timbral-related features is calculated for each training sample. The average value $\bar{r}_m$ of r_m for each indicator is then computed in advance for each class. For each indicator type of input data, frame selection is defined by the following equations:

$$L_k = \underset{n \in 1,2,\ldots,N}{\text{argtop-k}} \ \mathrm{AS}_n, \tag{1}$$

$$\mathrm{TRF}'_m = \begin{cases} \frac{1}{k} \sum\limits_{l \in L_k} \mathrm{TRF}_{m,l} & \text{if } |\bar{r}_m| > p, \\ \bar{\mathrm{TRF}}_m & \text{otherwise,} \end{cases} \tag{2}$$

where $\underset{t \in 1,2,\ldots,T}{\text{argtop-k}} \ X_t$ is a function that returns the set of top k indices t for which X_t is maximized, AS$_n$ is the AS of the n-th frame, TRF$_{m,l}$ is the value of the l-th frame of TRF$_m$, $\bar{\mathrm{TRF}}_m$ is the average of all frames of TRF$_m$, and TRF$'_m$ is the scalar value of TRF$_m$ after frame averaging. If $|\bar{r}_m|$ is above the threshold p, use the average of the frames with the same index as the k frames with the highest AS values. Otherwise, use the average over all frames. For AS, if there exists at least one m such that $|\bar{r}_m| > p$, the average of the highest k frames of AS is used. Otherwise, all frames of AS are averaged.

Through TFS, frame selection is executed only on indicators correlating with AS in non-stationary machine sounds. Stationary machine sounds contain machine sounds in all frames so that all frames will be averaged.

TABLE I: F-measure results on MIMII dataset (SNR 6 dB). SVM is our previous supervised ASD system, while others are unsupervised ASD systems.

System	Fan	Pump	Slider	Valve	Avg.
SVM [9]	0.985	0.959	0.937	0.798	0.920
GMM	0.970	0.916	0.882	0.777	0.886
TabPFN-ID	0.906	0.809	0.869	0.761	0.836
TabPFN-PA	0.970	0.943	0.872	0.804	0.897

IV. Evaluations

A. Database

To compare the proposed system with our previous system, we used the MIMII dataset [13] with an SNR of 6 dB. The MIMII dataset includes four types of industrial machines: fan, pump, slider, and valve. Each machine type contains four machine IDs.

We also used the DCASE 2020 Task 2 development dataset as the benchmark for unsupervised ASD. It contains recordings of six machine types: fan, pump, slider, valve, toy car, and toy conveyor, each of which includes three or four machine IDs. As in our previous study, we focused on industrial machines and excluded toy car and toy conveyor. Thus, we used a total of 16 classes, consisting of four machine types with four machine IDs each.

TABLE II: AUC [%] results on DCASE 2020 Task2 development dataset.

Feature	System	Fan	Pump	Slider	Valve	Avg.
log-Mel spectrogram	Autoencoder [17]	65.83	72.89	84.76	66.28	72.44
timbral-related features	GMM	83.58	79.63	75.88	76.55	78.91
	GMM w/ TFS	83.58	79.63	82.06	83.96	82.31
	TabPFN-ID	79.88	72.38	75.25	77.86	76.34
	TabPFN-PA	**86.24**	80.33	78.65	83.65	82.22
	TabPFN-PA w/ TFS	**86.24**	**82.67**	**86.00**	**86.45**	**85.34**

B. Implementation details

The discriminator in the proposed system is TabPFN [18], a deep learning model based on transformer [19] for tabular data. For experiments, we call the proposed system TabPFN-PA, which uses pseudo-anomaly (PA). The hyperparameter of TabPFN was set to default value. The discriminator was trained and evaluated for each ID. The evaluation of each machine type was calculated as the average evaluation of all machine IDs included in the machine type.

For anomaly generation, all classes except the target machine class are used, so we set the parameter $S = 15$. We also set $i = 2$ and $j = 1$ in the pseudo-anomaly generator, and $N = 77$, $k = 5$, and $p = 0.9$ for TFS.

C. Evaluation score

To evaluate the proposed system by comparing it with the previous system, we evaluated using the F-measure used for the previous system. The F-measure is the harmonic mean of precision and recall. Precision is the proportion of actual anomalous samples among the samples identified as anomalous by the system. Recall is the proportion of correctly identified anomalous samples out of all anomalous samples.

We also evaluated using the area under the receiver operating characteristic (ROC) curve (AUC) commonly used in unsupervised ASD.

D. Results

We compared the proposed system with three systems: our previous supervised system and two proposed unsupervised systems using timbral-related features. Our previous supervised system uses an SVM (hereafter, SVM). Among the two unsupervised systems, one uses a GMM trained only on normal sounds (hereafter, GMM), and the other uses TabPFN, which trains the model to distinguish between the target machine ID and other machine IDs (hereafter, TabPFN-ID), following a typical OE approach. TabPFN-ID replaces the classifier in the MobileNetV2 [20] baseline of DCASE 2021 Task 2 [14] with TabPFN and replaces the features with timbral-related features.

Table I shows a comparison of the F-measure between SVM and TabPFN-PA on the MIMII dataset. We did not apply TFS for this comparison. Although TabPFN-PA is an unsupervised ASD system, it performed comparable to SVM. TabPFN-PA also performed better than TabPFN-ID and GMM.

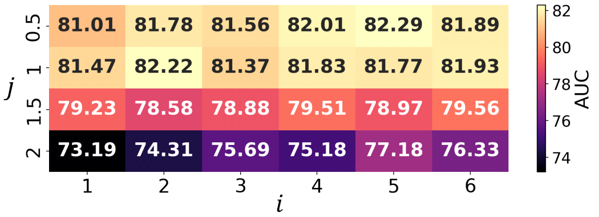

Fig. 4: Heatmap showing AUC [%] for each parameter of pseudo-anomaly generator.

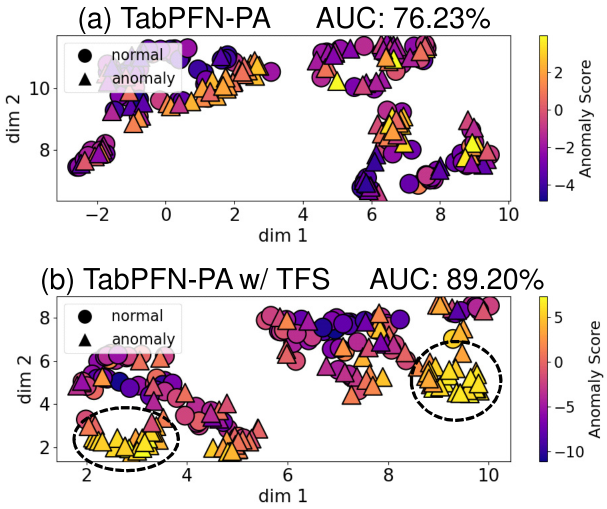

Fig. 5: UMAP visualization of distribution and anomaly scores for valve ID 04: (a) without TFS, (b) with TFS.

Table II shows a comparison of the AUC between TabPFN-PA and the other unsupervised ASD systems on the DCASE 2020 Task 2 development dataset. For reference, Table II also shows the AUC of the autoencoder using log-Mel spectrogram as features, which is the baseline of DCASE 2020 Task 2 [17]. Similar to the evaluation of F-measure, TabPFN-PA performed better than the other unsupervised systems. The application of TFS improved the performance of both GMM and TabPFN-PA for the slider and valve, whose sounds are non-stationary. These results indicate that learning pseudo-anomalies and TFS contribute to performance improvement. Furthermore, TabPFN-PA with TFS performed the best with an average AUC of 85.34% across all machine types.

E. Discussion

Figure 4 shows the change in AUC when i and j of the pseudo-anomaly generator in TabPFN-PA were changed on the DCASE 2020 Task 2 development dataset. Increasing i tended to improve performance slightly, and it is considered good to use multiple classes for generating a pseudo-anomaly. As j became larger, performance dropped significantly. This is thought to be because the pseudo-anomaly becomes too different from normal sounds as the parameter increases. It is also thought that appropriate settings of i and j have a significant effect on performance.

Figure 5 shows the distribution of timbral-related features by UMAP (Uniform Manifold Approximation and Projection) [21] and anomaly scores for valve ID 04 in the DCASE 2020 Task 2 development dataset before and after applying TFS. The application of TFS improved the degree of feature separation between normal and anomaly, thus performance. For non-stationary sounds, the features of the mechanical sound section are considered particularly effective for ASD.

V. Conclusion

We proposed an unsupervised ASD system using timbral-related features. A pseudo-anomaly generator in OE is used to handle external data in unsupervised learning. TFS is also used to extract timbral-related features from non-stationary machine sound. Evaluation of the proposed system showed that it performed comparable to our previous system. It also achieved superior performance compared with the unsupervised ASD systems using a GMM and that using the ID classification-based OE approach.

Acknowledgment

This work was supported by the JSPS Grant-in-Aid for Transformative Research Areas (A) (23H04344).

References

[1] 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 Workshop*, 2024, pp. 111–115.

[2] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. ICASSP*, 2020, pp. 271–275.

[3] T. Nishida, K. Dohi, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Anomalous sound detection based on machine activity detection," in *Proc. EUSIPCO*, 2022, pp. 269–273.

[4] K. Wilkinghoff, "Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection," in *Proc. ICASSP*, 2023, pp. 1–5.

[5] T. Fujimura, K. Imoto, and T. Toda, "Discriminative neighborhood smoothing for generative anomalous sound detection," in *Proc. EUSIPCO*, 2024, pp. 156–160.

[6] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. ICLR*, 2019.

[7] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, J. Liu, and Y. Qian, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *Proc. ICASSP*, 2024, pp. 1326–1330.

[8] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *Proc. ICASSP*, 2024, pp. 276–280.

[9] Y. Ota and M. Unoki, "Anomalous sound detection for industrial machines using acoustical features related to timbral metrics," *IEEE Access*, vol. 11, pp. 70 884–70 897, 2023.

[10] K. Li, Q.-H. Nguyen, Y. Ota, and M. Unoki, "Unsupervised anomalous sound detection for machine condition monitoring using temporal modulation features on gammatone auditory filterbank," in *Proc. DCASE Workshop*, 2022.

[11] T. Mian, A. Choudhary, and S. Fatima, "An efficient diagnosis approach for bearing faults using sound quality metrics," *Applied Acoustics*, vol. 195, p. 108839, 2022.

[12] A. Pearce, S. Safavi, T. Brookes, R. Mason, W. Wang, and M. Plumbley, "Deliverable D5.8: Release of timbral characterisation tools for semantically annotating non-musical content," Audio Commons, document D5.8, 2019.

[13] 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 *Proc. DCASE Workshop*, 2019, pp. 209–213.

[14] 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 *Proc. DCASE Workshop*, 2021, pp. 186–190.

[15] 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 TASLP*, vol. 27, no. 1, pp. 212–224, 2019.

[16] Y. Wang, Q. Zhang, W. Zhang, and Y. Zhang, "A lightweight framework for unsupervised anomalous sound detection based on selective learning of time-frequency domain features," *Applied Acoustics*, vol. 228, p. 110308, 2025.

[17] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2020, pp. 81–85.

[18] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," in *Proc. ICLR*, 2023.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, vol. 30, 2017.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, 2018, pp. 4510–4520.

[21] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.