Sensor applications

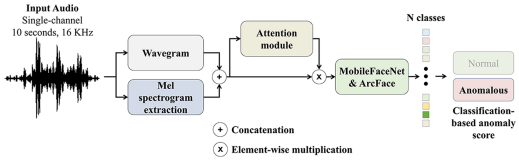
Low-Complexity Attention-Based Unsupervised Anomalous Sound Detection Exploiting Separable Convolutions and Angular Loss

Michael Neri* and Marco Carli**

Department of Industrial, Electronic, and Mechanical Engineering, Roma Tre University, 00146 Rome, Italy
**Graduate Student Member, IEEE*
***Senior Member, IEEE*

Manuscript received 9 July 2024; revised 13 September 2024; accepted 11 October 2024. Date of publication 14 October 2024; date of current version 30 October 2024.

Abstract—In this letter, a novel deep neural network, designed to enhance the efficiency and effectiveness of unsupervised sound anomaly detection, is presented. The proposed model exploits an attention module and separable convolutions to identify salient time–frequency patterns in audio data to discriminate between normal and anomalous sounds with reduced computational complexity. The approach is validated through extensive experiments using the Task 2 dataset of the DCASE 2020 challenge. Results demonstrate superior performance in terms of anomaly detection accuracy while having fewer parameters than state-of-the-art methods.



Index Terms—Sensor applications, attention, audio processing, deep learning, explainability, unsupervised anomaly detection, wavegram.

I. INTRODUCTION

In the context of unsupervised anomaly detection, an *anomaly* refers to data patterns that deviate from the expected *normal* behavior [1]. Likewise, anomalous sound detection (ASD) is the task of understanding whether a sound is *normal* or not (*anomalous*) [2]. ASD is applied in the field of machine condition monitoring [3], [4], medical diagnosis [5], safety and security in urban environments [6], and multimedia forensics [7], [8]. Generally, deep neural networks (DNNs) are employed for ASD due to their ability to identify subtle and unknown anomalous data patterns [9]. Unsupervised or semisupervised models are generally adopted in ASD problems because of the limited availability of anomalous sounds. State of the art (SOTA) unsupervised sound anomaly detection (USAD) approaches can be classified into two categories [10], [11]: *reconstruction based* and *classification based*. In the first scenario, models are based on the hypothesis that only nonanomalous samples, which have been analyzed during training, can be effectively retrieved after lossy compression, e.g., autoencoder (AE) [12]. In [13], a DNN has been designed to interpolate masked time bins of the log-Mel spectrogram. Similarly, in [14], normalizing flows have been used for estimating the probability density of normal data. However, these models suffer from generalization problems, e.g., an anomalous sample may be correctly reconstructed by an AE [15]. Classification-based approaches, instead, compute the anomaly score exploiting probability-based distances between prediction and ground truth, e.g., cross-entropy. The classification is carried out on metadata, which can be the identification number of a specific machine that produced the sound. The design rationale is that a model cannot classify successfully the metadata associated with a sound if it is anomalous [10], [11], [16]. The use of metadata as an auxiliary loss function allows the modeling of the probability distribution of normal data, namely, inlier modeling [17]. One such model, STgram-MFN [18], extracts temporal and spectral features to classify the IDs of machines using ArcFace [19]. Similarly, in [20], two novel angular losses, ArcMix and Noisy-Arcmix, have been designed to enhance the compactness of intraclass distribution during the classification of IDs. Differently, Guan et al. [11] involved contrastive learning in the pretraining to reduce distances between pairs of feature embeddings from the same machine IDs.

However, it is important to consider the computational complexity in the context of USAD. The response time of an anomaly detector is critical to limit the damage caused by an anomalous event [12]. Hence, this work also analyzes the computational complexity of SOTA approaches in terms of the amount of learnable parameters. Moreover, it is often challenging to interpret why these models flag certain audio segments as anomalies due to their closed box nature. To address this, for the first time in the literature, we employ an attention module [22] to provide explanations for the decisions made by the anomaly detection system. The attention mechanism highlights which parts of the input are most influential in the model's anomaly detection, thereby enhancing the interpretability of the model's outputs.

To summarize, the contributions of this work are as follows.

1) We define an attention module focused on identifying time–frequency anomalous pattern detected both in the log-Mel spectrogram and from the learned representation, i.e., Wavegram [23].
2) We use separable convolutions to reduce the computational complexity of the model, decreasing by approximately 13% of the number of learnable parameters concerning the top-tier approaches of the literature.
3) We statistically analyze the attention maps highlights the importance of high-frequency bins in the log-Mel spectrogram as the main cue for the identification of anomalous sounds in this scenario. Moreover, a comparison with SOTA approaches, in terms of performance and computational complexity, is carried out.

Corresponding author: Michael Neri (e-mail: michael.neri@uniroma3.it).
Associate Editor: S. Ostadabbas.
Data is available on-line at https://github.com/michaelneri/unsupervised-audio-anomaly-detection.
Digital Object Identifier 10.1109/LSENS.2024.3480450

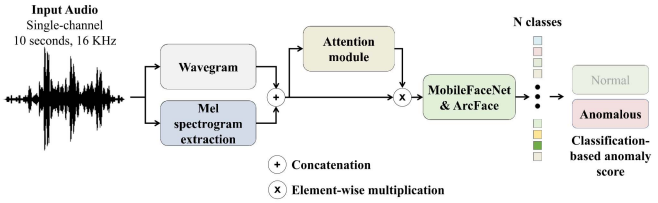


Fig. 1. Description of the proposed pipeline for USAD.

The rest of this letter is organized as follows. Section II details the proposed approach, providing insights on the feature extraction and anomaly score computation. Details of the dataset, metrics, and experimental results are explained in Section III. Finally, Section IV concludes this letter.

II. PROPOSED METHOD

In this section, the proposed unsupervised approach for audio anomaly detection is detailed. The goal is to determine whenever a single-channel audio signal with l samples $\mathbf{x} \in \mathbb{R}^{1\times l}$ is anomalous without using in training the binary anomaly label $y \in \mathbb{Z}^2$. To do so, we employ time–frequency representations as features, namely, log-Mel spectrogram and Wavegram, to jointly identify patterns in time and frequency since audio signals are generally nonstationary [23]. In conjunction with an attention module and angular loss, an efficient DNN is proposed for classification-based anomaly detection. The overall architecture is shown in Fig. 1.

A. Feature Extraction

Initially, a preprocessing stage is employed to extract the complex short-time Fourier transform (STFT) STFT{$\mathbf{x}$} from the audio signal $\mathbf{x}$. This transform is performed using a Hann window of length 64 ms with 50% overlap. The selection of the window function is critical since windowing in the time-domain results in a convolution in the frequency domain, disrupting the spectral characteristics of the audio signal. Hann window mitigates this problem, thanks to its characteristic of having the localization of spectral energy around the normalized frequency $w = 0$, minimizing spectral leakage [24]. Length and overlap of windows are consistent with those found in the literature for ASD. Next, a log-Mel spectrogram $X_{\text{Mel}} \in \mathbb{R}^{t\times f}$ is extracted using a Mel filterbank $\text{H}_{\text{Mel}}(\cdot)$ as $X_{\text{Mel}} = 20\log_{10}\text{H}_{\text{Mel}}(\text{STFT}\{\mathbf{x}\})$, where t and f denote the number of time and frequency bins, respectively.

In [22], Wavegram is introduced as a new learned time–frequency representation for audio tagging. In particular, Wavegram is designed to capture relevant time–frequency cues for the classification that may go unnoticed like hand-crafted log-Mel spectrograms due to its lossy representation [22]. Within the scope of USAD, several methods have been based on Wavegram by applying a 1-D convolution that acts as a learnable STFT [18]. Next, the features have been further processed by layer normalization and 1-D convolutions with small kernel sizes [2], [20]. To reduce the computational complexity, in this work, Wavegram consists only of a separable 1-D convolutional layer with f strided filters to mimic the windows' overlap in the STFT computation. Finally, the log-Mel spectrogram and the output of Wavegram $X_{\text{Wave}} \in \mathbb{R}^{t\times f}$ are concatenated along the channel dimension $X = [X_{\text{Mel}}, X_{\text{Wave}}] \in \mathbb{R}^{t\times f\times 2}$. An example of input acoustic features is depicted in Fig. 2.

B. Attention Module

The attention module is responsible for learning an attention map $H \in \mathbb{R}^{+t\times f\times 2}$ from the log-Mel spectrogram and the Wavegram. Its objective is to emphasize regions of features that are most informative for the classification task. This module has been extensively analyzed for evaluating the distance between a microphone and a speaker [21]. However, its application in ASD has not been investigated yet. In this work, it is denoted as the function $\text{f}_{\text{ATT}} : \mathbb{R}^{t\times f\times 2} \to \mathbb{R}^{+t\times f\times 2}$. It comprises two separable convolutional blocks, having 16 and 64 3×3 filters, respectively. Then, a 1×1 convolutional layer that acts as a linear projection to reduce the number of channels, with two filters, followed by a sigmoid activation function for mapping each pixel into a probability, is used to map the features to yield the $t \times f \times 2$ attention map. Finally, the weighted acoustic features $\tilde{X} \in \mathbb{R}^{t\times f\times 2}$ are obtained by elementwise multiplication ($\otimes$) between the two time–frequency representations and the attention map as $\tilde{X} = f_{\text{ATT}}(X) \otimes X$. Examples of attention maps are shown in Fig. 3.

C. Data Augmentation

To improve the robustness of the model, we synthetically augment the dataset using mixup [25] in each batch during the training, defined as $\mathbf{x}^{ij} = \lambda\mathbf{x}^i + (1-\lambda)\mathbf{x}^j$ and $\mathbf{y}^{ij} = \lambda\mathbf{y}^i + (1-\lambda)\mathbf{y}^j$, where $(\mathbf{x}, \mathbf{y})$ is the tuple describing the waveform $\mathbf{x}$ and the one-hot encoded metadata $\mathbf{y} = [y_1, y_2, \ldots, y_c]$ with c classes of a single audio recording under analysis, respectively. $i, j \in \{0, 1, \ldots, n-1\}$ are randomly selected indexes of training audio samples in the batch with size n, and $\lambda \sim \text{Beta}(\alpha, \alpha)$ is the mixup coefficient. This augmentation can be performed at different levels of the deep learning architecture, e.g., input level or at intermediate feature levels [25]. In this work, the augmentation procedure is applied to input signals before the preprocessing step, following [20].

D. Self-Supervised Anomaly Score

To distinguish between anomalous and normal sound, an anomaly score $\mathcal{A}_\theta$ is computed from the predicted metadata and the ground truth. As a classification-based approach, if a sound is misclassified, then it is anomalous since the model is trained to correctly classify normal sounds. We utilize ArcFace [19] as the classification layer

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y})_{\text{AF}} = -\mathbf{y}^T \frac{e^{s\cos\boldsymbol{\theta}+m\mathbf{y}}}{\sum_{i=1}^{c} e^{s\cos\theta_i+m\hat{y}_i}} \tag{1}$$

where the angular vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]$ is obtained for each class by computing $\theta_i = \arccos(\mathbf{w}_i^T\mathbf{h})$, which is the result of the mapping between the features obtained from the classifier $\mathbf{h} \in \mathbb{R}^{h\times 1}$ and learned ArcFace weights $\mathbf{w}_i \in \mathbb{R}^{h\times 1}$ for the ith class. The scalars $s \in \mathbb{R}^+$ and $m \in \mathbb{R}^+$ are the scale and margin coefficients for the ArcFace loss, respectively. As introduced in [20], the employed loss function for training the model is

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}, \mathbf{y}^{ij}) = \lambda\mathcal{L}(\boldsymbol{\theta}, \mathbf{y})_{\text{AF}} + (1-\lambda)\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}^{ij})_{\text{AF}}. \tag{2}$$

During the testing phase, as the augmentation is not performed, the anomaly score is computed as $\mathcal{A}_\theta(y, \hat{y}) = \mathcal{L}(\boldsymbol{\theta}, \mathbf{y})_{\text{AF}}$.

III. EXPERIMENTAL RESULTS

Task 2 development dataset of the DCASE 2020 challenge [3] is used to assess the performance of the proposed approach. It encompasses six machines (Fan, Pump, Slider, Valve, ToyCar, and ToyConveyor), and each machine is labeled with a unique identifier to differentiate audio recordings from various machines within the same category. A total of 41 machines with 10 s of audio signals are collected. To assess the performance of the proposed approach, we evaluate the area under the curve (AUC) and partial area under the curve (pAUC) metrics. The latter is the AUC over a low FPR in the range $[0, p]$ with $p = 0.1$, following [26]. Our approach is trained to classify the $c = 41$ labels derived from machine types and IDs [3], [4]. For the

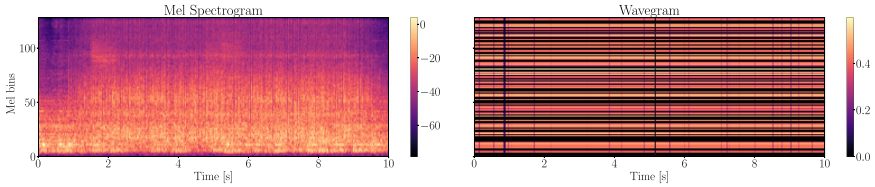


Fig. 2. Example of acoustic features X from an anomalous sound of Task 2 DCASE 2020 dataset.

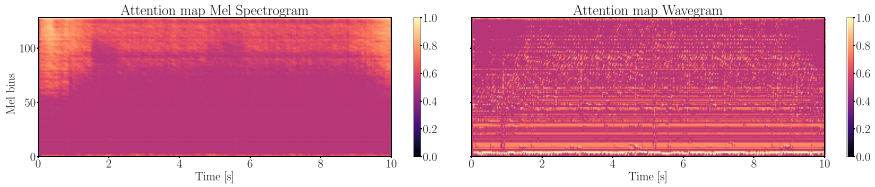


Fig. 3. Attention maps $H = f_{\text{ATT}}(X)$ obtained from the attention module with acoustic features in Fig. 2.

Table 1. Comparison With SOTA Methods

Methods	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor	
	AUC [%]	pAUC [%]	AUC [%]	pAUC [%]	AUC [%]	pAUC [%]	AUC [%]	pAUC [%]	AUC [%]	pAUC [%]	AUC [%]	pAUC [%]
IDNN [13]	67.71	52.90	73.76	61.07	86.45	67.58	84.09	64.94	78.69	69.22	71.07	59.70
MobileNetV2 [16]	80.19	74.40	82.53	76.50	95.27	85.22	88.65	87.98	87.66	85.92	69.71	56.43
Glow-Aff [14]	74.90	65.30	83.40	73.80	94.60	82.80	91.40	75.00	92.20	84.10	71.50	59.00
GMM + Arcface [10]	87.97	80.66	**95.63**	85.74	99.22	97.55	91.26	84.00	95.28	86.91	69.80	61.21
STgram-MFN [18]	94.04	88.97	91.94	81.75	**99.55**	**97.61**	99.64	98.44	94.44	87.68	74.57	63.60
SW-WaveNet [2]	97.53	91.54	87.27	82.68	98.96	94.58	99.01	97.26	95.49	90.20	81.20	68.20
Noisy-ArcMix [20]	**98.32**	**95.34**	95.44	**85.99**	99.53	97.50	99.95	99.74	96.76	90.11	77.90	67.15
Proposed approach	95.10	87.25	91.97	80.00	99.24	96.10	**99.99**	**99.96**	**96.99**	**90.30**	**84.59**	**73.55**

Bold and underline are used to highlight first and second-best results, respectively.

Table 2. Number of Parameters, Average AUC, and Average pAUC of SOTA Approaches and Proposed Method

Methods	Parameters	AUC [%]	pAUC [%]
IDNN [13]	**46 k**	76.96	62.57
MobileNetV2 [16]	1.1 M	84.34	77.74
Glow-Aff [14]	30 M	85.20	73.90
GMM + Arcface [10]	1 M	89.86	82.68
STgram-MFN [18]	1.1 M	92.36	86.34
SW-WaveNet [2]	27 M	93.25	87.41
Noisy-ArcMix [20]	1.1 M	**94.65**	**89.31**
Proposed approach	884 k	93.44	85.71

Table 3. Selection of Parameters of the Proposed Approach

Features	h	AUC [%]	pAUC [%]
Feature study			
$[X_{\text{Mel}}]$	128	92.26	84.55
$[X_{\text{Wav}}]$	128	63.48	54.12
Dimensionality study			
$[X_{\text{Mel}}, X_{\text{Wav}}]$	256	90.87	83.94
$[X_{\text{Mel}}, X_{\text{Wav}}]$	64	91.94	85.00
$[X_{\text{Mel}}, X_{\text{Wav}}]$	128	**93.43**	**85.71**

Table 4. Ablation Study

Methods	Parameters	AUC [%]	pAUC [%]
w/o separable convs., w/o f_{ATT}	1 M	90.50	83.62
w/o separable convs.	1 M	92.25	84.82
w/o f_{ATT}	882 k	91.72	84.52
Proposed approach	884 k	**93.43**	**85.71**

loss and mixup computation, parameters are set as $\alpha = 0.2$, $m = 0.7$, and $s = 40$, following the guidelines provided by their corresponding works [19], [25]. Log-Mel spectrogram and Wavegram output have $t = 313$ and $f = 128$ bins. The classifier is MobileFaceNet [27], which yields a feature vector with dimensionality $h = 128$. The network is optimized using AdamW with a learning rate of 0.0001, epochs of 300, and a batch size of 64. Hyperparameters of the training procedure have been assigned by means of a grid search optimization procedure.

A. Results

The performance of the proposed approach compared with those obtained with SOTA architectures are represented in Table 1. Overall, the proposed approach shows the best performance in three of the six equipment types (Valve, ToyCar, and ToyConveyor). Specifically, the approach achieves SOTA performance on ToyConveyor, which is the most difficult machine in this dataset. This is possible, thanks to the combination of Wavegram and log-Mel spectrogram, providing additional cues to the classifier. In the other classes, the performance is still competitive. Generally, Table 1 can be used as a reference for the selection of the approach that is most suitable to the specific use case. Regarding the computational complexity, Table 2 highlights the number of parameters and the performance of our approach compared with those of the SOTA. Our system offers a good tradeoff between model complexity and performance.

Table 3 tabulates the selection of parameters regarding the type of features and the dimensionality of the ArcFace layer. The use of Wavegram representation $[X_{\text{Wav}}]$ in conjunction with the log-Mel spectrogram can improve the performance of the proposed model by 1.17% in terms of AUC, albeit being ineffective using it alone. Moreover, the best performance is obtained by setting the dimensionality of the classification layer to $h = 128$.

To better explain which parts of the log-Mel spectrogram are relevant for the ID classification, Fig. 4 shows the average and standard deviation maps on the testing set of the proposed heatmap. Interestingly, the most important frequency bins for the identification of anomalies, i.e., ID misclassification, are contained in the range [1.7, 8] kHz of the log-Mel spectrogram. In addition, the range [0, 71] Hz is also relevant. The rest of the log-Mel spectrogram [0.071, 1.7] kHZ is assigned a value of 0.5 with zero variance, denoting this region as less important for the ID classification and, thus, less reliable for the identification of anomalies. To assess the impact of both separable convolutions and the attention module, an ablation study has been carried out. Table 4 demonstrates the effectiveness of using the attention map in combination with separable convolutions.

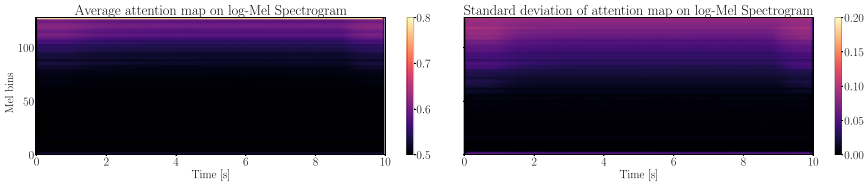


Fig. 4. Mean and standard deviation of the attention map of the log-Mel spectrogram on testing set.

IV. CONCLUSION

In this letter, a learning-based low-complexity approach is proposed to detect anomalous sound in a machine monitoring scenario. To this aim, a DNN is proposed. It exploits an attention module to highlight the most salient time–frequency patterns for identifying machine IDs. Then, an anomaly score is computed from the classification errors between predicted and ground truth metadata. Experimental results demonstrate the validity of the proposed low-complexity model. Although retraining the entire architecture is necessary to handle domain shifts in new environments or with different machines, the approach's low complexity enables efficient fine-tuning with new normal data. Future work will focus on the improvement of the attention module, coping with more complex tasks in the realm of sound anomaly detection, such as few and one-shot unsupervised anomaly detection.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[2] H. Chen, L. Ran, X. Sun, and C. Cai, "SW-WAVENET: Learning representation from spectrogram and wavegram using wavenet for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[3] K. Dohi et al., "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proc. DCASE*, 2022.

[4] 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," in *Proc. DCASE*, 2022.

[5] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2162–2171, Jun. 2021.

[6] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound event detection for human safety and security in noisy environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022.

[7] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, 2007, pp. 21–26.

[8] M. Neri, A. Ferrarotti, L. D. Luisa, A. Salimbeni, and M. Carli, "ParalMGC: Multiple audio representations for synthetic human speech attribution," in *Proc. 10th Eur. Workshop Vis. Inf. Process.*, 2022, pp. 1–6.

[9] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 608–622, 2024.

[10] J. Wu, F. Yang, and W. Hu, "Unsupervised anomalous sound detection for industrial monitoring based on ArcFace classifier and gaussian mixture model," *Appl. Acoust.*, vol. 203, 2023, Art. no. 109188.

[11] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[12] 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 Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, Jan. 2019.

[13] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 271–275.

[14] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 336–340.

[15] G. Bovenzi, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, and A. Pescapé, "Network anomaly detection methods in IoT environments via deep learning: A. fair comparison of performance and robustness," *Comput. Secur.*, vol. 128, 2023, Art. no. 103167.

[16] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. DCASE*, 2020.

[17] Y. Kawaguchi et al., "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE*, 2021, pp. 186–190.

[18] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 816–820.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4685–4694.

[20] S. Choi and J. Choi, "Noisy-ArcMix: Additive noisy angular margin loss combined with mixup for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 516–520.

[21] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, "Speaker distance estimation in enclosures from single-channel audio," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2242–2254, 2024.

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[23] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.

[24] M. Prabhu, K. M., *Window Functions and Their Applications in Signal Processing*, New York, NY, USA: Taylor & Francis, 2014.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[26] Y. Koizumi et al., "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2020.

[27] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Biometric Recognit.: 13th Chin. Conf.*, 2018, pp. 428–438.

Open Access funding provided by 'Università degli Studi Roma Tre' within the CRUI CARE Agreement