

CLF-AIAD: A Contrastive Learning Framework for Acoustic Industrial Anomaly Detection

Zhaoyi Liu¹, Yuanbo Hou², Haoyu Tang³, Álvaro López-Chilet⁴, Sam Michiels¹, Dick Botteldooren², Jon Ander Gómez⁴, and Danny Hughes¹

¹ imec-Distrinet, Computer Science, KU Leuven, Belgium

² WAVES Research Group, Ghent University, Belgium

³ Technical Platform Department, United-Imaging Microelectronics, China

⁴ PRHLT Research Center, Universitat Politècnica de València, Spain.

`zhaoyi.liu@kuleuven.be`

Abstract. Acoustic Industrial Anomaly Detection (AIAD) has received a great deal of attention as a technique to discover faults or malicious activity, allowing for preventive measures to be more effectively targeted. The essence of AIAD is to learn the compact distribution of normal acoustic data and detect outliers as anomalies during testing. However, recent AIAD work does not capture the dependencies and dynamics of Acoustic Industrial Data (AID). To address this issue, we propose a novel Contrastive Learning Framework (CLF) for AIAD, known as CLF-AIAD. Our method introduces a multi-grained contrastive learning-based framework to extract robust normal AID representations. Specifically, we first employ a projection layer and a novel context-based contrast method to learn robust temporal vectors. Building upon this, we then introduce a sample-wise contrasting-based module to capture local invariant characteristics, improving the discriminative capabilities of the model. Finally, a transformation classifier is introduced to bolster the performance of the primary task under a self-supervised learning framework. Extensive experiments on two typical industrial datasets, MIMII and ToyADMOS, demonstrate that our proposed CLF-AIAD effectively detects various real-world defects and improves upon the state-of-the-art in unsupervised industrial anomaly detection.

Keywords: Acoustic industrial anomaly detection · Contrastive learning · Anomaly detection · Unsupervised learning

1 Introduction

Faults or failures in industrial machines significantly degrade industrial production efficiency and quality. The Industrial Internet of Things (IIoT) tackles this problem by enabling the detection of anomalies in near real time using wireless sensors and actuators. The exponential growth of the IIoT has led to the generation of vast volumes of Acoustic Industrial Data (AID) from industrial machines that frequently handle critical operations [19,34]. Consequently, Acoustic Industrial Anomaly Detection (AIAD) has attracted substantial attention and

holds significant importance as it may serve as an indicator of faults, failures or malicious activities, thereby enabling targeted intervention to avert potential issues [6, 15, 17].

AIAD tasks can be categorized into supervised and unsupervised approaches. On the one hand, supervised AIAD emphasizes detecting known abnormal noises [29], which represent rare acoustic events. On the other hand, unsupervised AIAD detects unseen anomalous noises that have not been previously observed [3, 23]. Due to the nature of anomalous behaviours in industrial equipment, collecting an extensive set of anomalous sounds is impossible. This naturally leads us to focus on unsupervised AIAD.

In unsupervised AIAD, models are trained solely on normal AID, and during inference, anomalies are identified as those significantly deviating from the normal AID. For this purpose, the acoustic feature acquisition and latent representation of normal acoustic data are essential to distinguish the normal and unknown sounds. A range of autoencoder-based (AE-based) models have been proposed for learning better acoustic representations [6, 12, 25, 27]. Typically, the AE-based methods perform a frame-to-frame reconstruction hypothesizing that the error will be higher on abnormal data, allowing to use it as an anomaly detector [9]. However, temporal dependencies in AID are complex and cannot comprehensively characterize the temporal pattern of AID. In addition, the normal AID may change dynamically over time. Thus, existing approaches need to be improved to construct a precise profile for normal AID and hence acquire a robust representation, that delivers good performance in real-world industrial applications.

In this paper, we propose a novel Contrastive Learning Framework for AIAD, namely CLF-AIAD, which aims to capture the dependencies and dynamic features of AIAD. CLF-AIAD presents a multi-grained contrastive learning-based framework, combining contextual-wise and sample-wise contrasting information to construct a comprehensive normal profile in AID. More precisely, we first employ a projection layer combined with a novel acoustic contextual contrasting method. This allows us to effectively learn the temporal feature vectors that capture essential position information within the AID data. Furthermore, we investigate Data Augmentation (DA) in the time and frequency domain. After that, we introduce a sample-level contrasting mechanism to effectively capture the local invariant characteristics of latent representations. Finally, a transformation classifier is introduced as an auxiliary task to improve the performance of the primary task under an unsupervised learning framework. Extensive experiments are conducted to validate the effectiveness of the proposed method. Moreover, our approach achieves state-of-the-art anomaly detection performance in MIMII [24] and ToyAdmos [18] real-record AIAD datasets.

Contributions: The key scientific contributions of this paper are

1. We propose a novel Contrastive Learning Framework for AIAD, called CLF-AIAD. Contrastive learning methods are employed to improve the construction of a robust profile for normal AID, thereby enhancing the detectability of abnormal AID.

2. Multi-grained contrasting approaches are proposed to discover robust AID representations, encompassing both the novel contextual and contrasting acoustic modules.
3. We contribute in-depth comparative experiments to evaluate the performance of CLT-AIAD on two real-world datasets. Our experimental results demonstrate that CLT-AIAD outperforms state-of-the-art baselines.

The remainder of this paper is structured as follows: Section 2 provides an overview of relevant work. Section 3 describes the proposed CLF-AIAD technique and key implementation details. Experimental setup, results, and analysis appear in Section 4. Finally, Section 5 draws conclusions.

2 Related Work

Existing unsupervised Anomaly Detection (AD) methods can be grouped into two categories.

Reconstruction-based methods assume that a generalization gap with anomalies exists in the reconstruction model trained on normal acoustic data, preventing it from reconstructing anomalies. Autoencoders (AE) [15, 33], Variational Autoencoders (VAE) [16], and Generative Adversarial Networks (GAN) [14] are typical models belonging to this category. [9] designed a sample-level anomaly score using an AE, calculated by aggregating pixel-wise errors. [35] proposed a masked multi-scale reconstruction method, which enhances the model’s ability to infer causality among patches in normal samples and improves performance on different domain shifts.

Representation-based methods involve transforming input data into a feature domain to obtain representative feature embeddings; then, classification or clustering tasks are performed based on these learned representations. Since the embeddings include representative input data characteristics, it is much easier to determine anomalies using normal data. Gaussian Mixture Models (GMMs) based methods [20], k-means [28], and One-Class Support Vector Machines [1] are the common approaches within this category. In [21] combined two classification models as feature extractors and local representations to detect anomalies. PaDiM [4] learns the parameters of multivariate Gaussian distribution from different CNN layers.

Recently, self-supervised learning techniques that do not require hand-crafted labels have been applied to detect anomalous data [8, 13]. **Contrastive learning-based methods** mainly aim to learn transformation-invariant data representations in a self-supervised setting. These methods seek to minimize the distance between similar samples in the feature space while maximizing the distance between different samples within the same space. SimCLR [2] emphasized the importance of data augmentation in contrastive learning and proposed a learnable nonlinear transformation module to enhance the quality of contrastive representation vectors. He et al. [10] proposed the MoCo to address the challenge of obtaining negative samples by constructing a dynamic dictionary to facilitate the retrieval of sample pairs for contrastive learning. For time series data, Eldele et al. [7] proposed a self-supervised framework with multiple contrasting

methods to learn informative representations of time series. Yue et al. [32] introduced a hierarchical contrasting method to capture multi-grained contextual information in time series.

In this paper, we extend these ideas to develop a contrastive learning framework for representation learning and anomaly detection in AID.

3 Proposed Method

In this section, we first present the standard problem description of unsupervised AIAD. Then, we give an overview of the proposed CLF-AIAD and explain its main modules. Finally, we describe in detail the proposed model and AIAD.

3.1 Problem Description

Unsupervised AIAD is an identification problem of determining whether the acoustic signals emitted from a target are normal or anomalous.

Let us define $\mathbf{x}_\tau \in \mathbb{C}^T$ is the input spectrogram calculated from the observed acoustic signal. Here $\tau = \{1, 2, \dots, T\}$ denote the time frames. To detect anomalies, a set of low-dimensional representations is extracted as:

$$\mathbf{z} = \mathcal{F}_{dr}(\mathbf{x}_\tau) \quad (1)$$

where $\mathcal{F}_{dr}$ is a dimensional reduction model usually utilizing a Deep Neural Network (DNN). To determine the spectro-temporal characteristics of observed acoustic feature, the input vector $\mathbf{x}_\tau$ is obtained by concatenating several frames of observation while accounting for previous and future frames, as $\mathbf{x}_\tau = (X_{\omega, \tau-p}, X_{\omega, \tau-p+1}, \dots, X_{\omega, \tau+f})^\top$, where $\top$ denotes transposition, and p and f are the context window size of previous and future indexes, respectively. Next, an anomaly score $\mathcal{A}_\theta(\mathbf{z})$ is calculated using a normal model as:

$$\mathcal{A}_\theta(\mathbf{z}) = -\ln p(\mathbf{z}|\theta, s=0) \quad (2)$$

Where s indicates the state of observed data, $s=0$ is normal, and $s \neq 0$ is abnormal. $p(\mathbf{z}|\theta, s=0)$ is a normal model generally constructed with a Probability Density Function (PDF) of normal acoustic data.

Finally, $\mathbf{x}_\tau$ can be identified as an abnormal acoustic signal when a score of $\mathcal{A}_\theta(\mathbf{z})$ surpasses the pre-defined threshold value ϕ :

$$\mathcal{H}(\mathbf{z}, \phi) = \begin{cases} 0(Normal) & \text{if } \mathcal{A}_\theta(\mathbf{z}) < \phi, \\ 1(Anomaly) & \text{if } \mathcal{A}_\theta(\mathbf{z}) \geq \phi. \end{cases} \quad (3)$$

3.2 Proposed Contrastive Learning Framework

Following the procedure of unsupervised AIAD described in Section 3.1, we note that the hidden representation $\mathbf{z}$ of acoustic data is critical to obtaining good performance to determine anomaly detection. Therefore, we propose a Contrastive Learning Framework for AIAD (CLF-AIAD), which utilizes multi-grained contrasting methods to capture AID's temporal dependency and dynamic variabilities. An overview of our anomaly detection framework is depicted in Fig. 1.

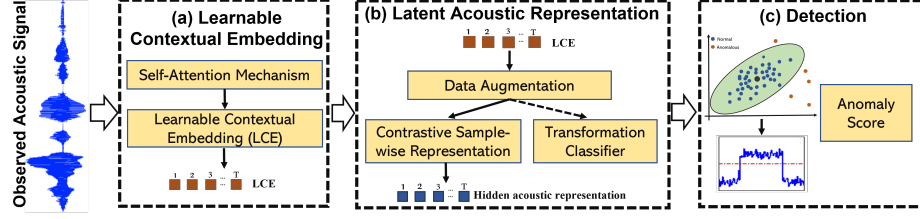


Fig. 1. Overview of proposed CLF-AIAD. (a) Learnable Contextual Embedding (LCE): A novel LCE is applied to learn the robust temporal dependency of AID. (b) Latent acoustic representation: ① LCE undergoes transformations into multiple perspectives through the utilization of both time-domain and frequency-domain time-series-specific augmentations; ② An acoustic sample contrasting method is proposed to learn invariant local hidden acoustic representation; ③ A transformation classifier is used as an auxiliary task to classify the applied augmentations in step ①. (c) Detection: An anomaly score is calculated based on the Mahalanobis distance between the hidden acoustic representations of the input sample and the normal training acoustic instances.

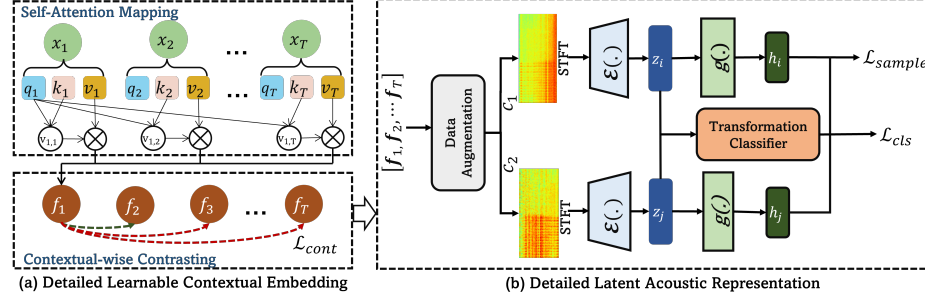


Fig. 2. Detailed Neural Network of CLF-AIAD.

Learnable Contextual Embedding While simple transformations are commonly employed in contrastive learning to generate diverse views of acoustic data, direct transformation of the original AID may not always be appropriate, especially when certain patterns or trends exist, such as upward trends in machine sounds. To address this issue, in the CLF-AIAD, we first employ a self-attention mechanism [31] as the projection layer to capture the temporal relationships and dependencies of the AID. A mask is applied to assign zero attention scores to future frames to prevent future data information leakage. For an acoustic input $\mathbf{x} = [x_1, x_2, \dots, x_T]$, linear mappings are applied to obtain the query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$. The output LCE ($\mathbf{f}_{1:T}$) obtained after applying the self-attention mechanism can be expressed as below.

$$\mathbf{f} = \text{Softmax}((\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k}) \odot \mathbf{Mask})\mathbf{V} \quad (4)$$

where symbols $\odot$ and $^\top$ denote element-wise and the transpose operations of the matrix, respectively. The $\mathbf{f} = [f_1, f_2, \dots, f_T]$, $f \in \mathbb{R}^T$ denotes the sequence of the LCE. The $\frac{1}{\sqrt{d_k}}$ is the scale factor. The matrix $\mathbf{Mask}$ is a lower triangular matrix, indicating that the data at frame τ only focuses on current and historical information. That is, the embedding f_τ only attends to the information of $\mathbf{x}_{1:\tau}$.

Utilizing only the self-attention mechanism overlooks position information, and direct incorporation of position information into the data can lead to interference [30]. To address this, we propose the contextual-wise contrasting for AID, depicted in (a) of Fig. 2, to learn position information within the AID’s embedding space and capture its temporal relationship.

The time frame τ in LCE shows higher similarity to neighbouring frames $[\tau - 1, \tau + 1]$ and lower similarity to distant frames. Contextual-wise contrasting minimizes the distance between neighbouring frames while maximizing the distance between different frames in the embedding space. This is expressed as below:

$$\ell_{cont} = -\log \frac{\exp(\text{sim}(\mathbf{f}_\tau, \mathbf{f}_{\tau-1}/t)) + \exp(\text{sim}(\mathbf{f}_\tau, \mathbf{f}_{\tau+1}/t))}{\sum_{k=1}^T \mathbb{1}_{[k \neq \tau]} \exp(\text{sim}(\mathbf{f}_\tau, \mathbf{f}_k)/t)}, \quad (5)$$

$$\mathcal{L}_{cont} = \frac{1}{N} \sum_{i=1}^N \sum_{\tau=1}^T \ell_{cont}. \quad (6)$$

where ℓ_{cont} indicates the loss function at frame τ . The function $\mathbb{1}_{[k \neq \tau]} \in \{0, 1\}$ is an indicator function that equals 1 if $k \neq \tau$, and 0 otherwise. t represents the temperature parameter. The $\mathcal{L}_{cont}$ shows the total contextual-wise contrastive loss of each batch. Besides, $\text{sim}(\cdot)$ denotes cosine similarity among embeddings and is equal to $\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$.

Latent Acoustic Representation To leverage the characteristics of the AID fully, we investigate data augmentation transformations for the learnable time-series LCE in both the time and frequency domains. We explore the following transformations to create augmented versions of a given sample:

- (a) **Time Shifting:** The time-series embeddings are shifted forwards or backwards, with the degree of shift randomly chosen from the range of zero to half of the embeddings’ length.
- (b) **Time-domain Gaussian Noise Injection:** In the time domain, we introduce Gaussian noise $N(0, \Sigma_\tau)$ to the embeddings to generate similar samples, where Σ_τ is a diagonal matrix with elements $\sigma_\tau = 0.5$ as the deviation.
- (c) **Frequency-domain Gaussian Noise Injection:** In the frequency domain, we perform a short-time Fourier transform (STFT) on the embeddings and add the Gaussian noise $N(0, \Sigma_f)$ for the STFT. The $\Sigma_f = \text{diag}(\sigma_f)$ and $\sigma_f = 0.5$ control the deviation of the noise.
- (d) **Time Masking:** This transformation randomly selects a segment of the LCE time axis and sets it equal to zero or the mean of the whole embedding. The size of the masked portion of the embedding is randomly chosen to be less than 8% of the LCE’s length.
- (e) **Frequency Masking:** It applies random masking to the STFT spectrum of the LCE, randomly setting to zeros a segment of frequencies from the embedding. The length of the masked segment is randomly chosen to be less than 5% of the number of frequency bands in the LCE.

In these data augmentations, we define $\mathcal{C}$ as the family of possible augmentation operators. We stochastically apply two operators, $c_1 \sim \mathcal{C}$ and $c_2 \sim \mathcal{C}$, to generate two correlated views from each sample in the batch: $\tilde{\mathbf{f}}_1^{(m)} = c_1(\mathbf{f}^{(m)})$ and $\tilde{\mathbf{f}}_2^{(m)} = c_2(\mathbf{f}^{(m)})$.

After applying the transformations, we extract the two-dimensional STFT spectrogram of the embeddings. As shown in Fig.2 (b), sample-wise contrasting is used to learn the latent variables' local invariant information. The augmented data is then fed into the siamese encoders $\mathcal{E}(\cdot)$, which are DNNs that map the input data to lower-dimensional latent acoustic representations $\mathbf{z}^{(m)} = \mathcal{E}(\tilde{\mathbf{f}}^{(m)})$. The Siamese encoders consist of two neural networks with shared weights.

Following the SimCLR paper [2], we introduce a projection head $g(\cdot)$ to map the latent representation $\mathbf{z}$ to a subspace for calculating the sample-wise contrastive loss. This projection head is implemented as a one-hidden-layer Multi-Layer Perceptron (MLP) with ReLU activation. Note that the projection head is not used during inference.

The contrastive loss is computed on the output of the projection head, denoted as $\mathbf{h}^{(m)} = g(\mathbf{z}^{(m)})$. The loss function promotes the proximity of latent representations for positive pairs $(\mathbf{h}_i^{(m)}, \mathbf{h}_i^{(m)+})$ while encouraging separation from negative pairs. We treat all other samples in the batch as negative pairs. In particular, we used Normalized Temperature-scaled Cross-Entropy loss (NT-Xent) as our contrastive loss function [26]. The contrastive loss for the positive pair instance is calculated as follows:

$$\ell_{(i,i+)} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/t)}{\sum_{j=1}^{2N} \mathbb{1}_{[j \neq i]} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/t)}, \quad (7)$$

$$\mathcal{L}_{\text{sample}} = \frac{1}{2N} \sum_{i=1}^N \ell_{(i,i+)} + \ell_{(i+,i)} \quad (8)$$

In self-supervised learning, incorporating an auxiliary task can significantly enhance the performance of the model on the main task [5]. One commonly used auxiliary task involves training a simple classifier on the latent representation to predict the type of transformation applied to the data. In our CLF-AIAD, we utilize a linear classifier, denoted as $p_{cls}(\mathcal{C} = c_i | \tilde{\mathbf{f}})$, which operates on the latent representation $\mathbf{z}$ to predict the applied transformation. The loss function employed for this classifier is the Multi-class Cross-Entropy Loss ($\mathcal{L}_{cls}$).

The final loss of the proposed CLF-AIAD is a weighted sum of the contextual-wise and sample-wise contrastive loss and the loss of the transformation classifier:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{sample}} + \lambda_2 \mathcal{L}_{\text{cls}}, \quad (9)$$

where λ_1 and λ_2 are balancing hyperparameters.

Detection Finally, we require an anomaly scoring function that maps the latent embedding to a scalar to quantify the degree of the AID's abnormality. In this study, we utilize the Mahalanobis distance [22], which is a metric for measuring the distance between a point and a distribution. During the test stage, we compute the distance between the latent embedding of the query sample and the representations of normal training instances. The anomaly score $S_{\mathbf{x}}$ is calculated

as follows for a given input $\mathbf{x}$:

$$S_{\mathbf{x}} = (\mathbf{z}_{\mathbf{x}} - \mu)^T \Sigma^{-1} (\mathbf{z}_{\mathbf{x}} - \mu). \quad (10)$$

Where Σ means their covariance matrix, μ denotes the mean vector of normal training AID, and $\mathbf{z}_{\mathbf{x}}$ denotes the latent representation of the acoustic clip x .

4 Experiments

4.1 Datasets

To evaluate the proposed method, we conducted experiments using two public datasets of machine-operating acoustic data, MIMII [24] and ToyADMOS [18].

MIMII is designed to support the automation of facility maintenance tasks for the machine learning and signal processing communities. It consists of normal and anomalous acoustic data recordings from the four most common industrial machines. The anomalous acoustic data simulate real-life scenarios such as contamination, leakage, rotating unbalance, and rail damage. The acoustic data includes background factory noise to provide a realistic environment. Each recording in the corpus lasts 10 seconds with a sampling rate of 16 kHz.

ToyADMOS is a dataset created for anomaly detection in machine operating sounds (ADMOS). It includes recordings of anomalous operating sounds from miniature machines (toys) deliberately damaged for this purpose. The dataset comprises three sub-datasets: machine-condition inspection, fault diagnosis of machines with fixed tasks, and fault diagnosis of machines with moving tasks. The recordings were made using four microphones at a sampling rate of 48 kHz.

In our study, we focus on the valves, pumps, fans, and sliders sub-datasets from MIMII as well as toy car and toy conveyor sub-datasets from ToyADMOS.

4.2 Implementation

The STFT spectrogram is obtained using a Blackman window with a length of 1024 and 50% overlap, resulting in 513 frequency bands for the STFT. The self-attention mechanism from [31] is adopted as the contextual projection layer. The ResNet-18 architecture from [11] is used as the encoder $\mathcal{E}(\cdot)$, with the 256 units in the linear output layer. The projection head consists of an MLP with a hidden layer of 256 units, and an output layer of 128 units for the representation vector. The Adam optimizer is employed with an initial learning rate of 0.01. A batch size of 128 is used, and the network is trained for 600 epochs. The temperature parameter t is empirically set to 0.08, and throughout all our experiments, we set $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$ in Eq.(9).

4.3 Results and Analysis

We evaluate our proposed model using the area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC) scores [6, 15] for each machine type.

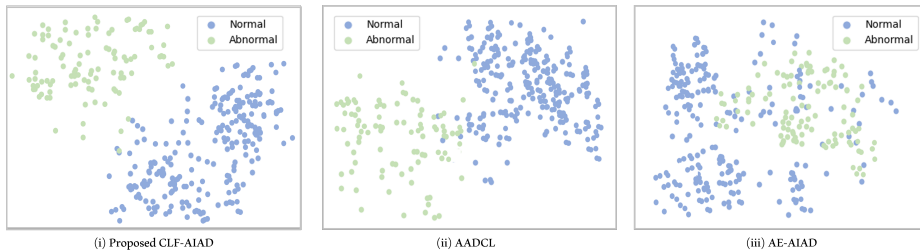


Fig. 3. The latent representations t-SNE visualization of proposed CLF-AIAD, AADCL and AE-AIAD for the machine type Fan.

Comparison Results We compared our CLF-AIAD with recent unsupervised AIAD baseline and state-of-the-art (SOTA) methods on DCASE2021/2022 Task2, including AE-AIAD [15], MobileNetV2 [15], IDNN [27], FREAK [21], AADCL [13]. Here, AE-AIAD acts as a baseline, IDNN represents the recent SOTA in reconstruction-based methods, MobileNetV2 and FREAK lead in representation-based classification methods, and AADCL stands as the SOTA self-supervised method utilizing contrastive learning.

As shown in Table. 1, our proposed method achieves the best performance in terms of average AUC and average pAUC across six machine types. The proposed CLF-AIAD provides improvements of 1.7% and 3.6% in terms of average AUC and average pAUC, respectively, compared to the current SOTA contrastive learning-based method AADCL.

In order to provide a visual depiction of the efficacy of the learned audio feature representations in anomaly detection, we present in Fig 3 the t-distributed Stochastic Neighbor Embedding (t-SNE) cluster visualization of the latent acoustic representations obtained from the proposed CLF-AIAD, AADCL, and baseline AE-AIAD methods. Through this visualization, we can discern that our method exhibits better discriminatory capabilities, effectively distinguishing anomalies in the audio data compared to the other approaches.

Ablation Study We evaluated the performance of CLF-AIAD along with two variants: CLF-cont-AIAD and CLF-samp-AIAD. Specifically, CLF-cont-AIAD represents the variant that retains the contextual-wise contrastive loss while

Table 1. Performance comparison in terms of AUC (%) and pAUC (%) on the test data of MIMII and ToyADMOS datasets.

Methods	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor		Average	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
AE-AIAD [15]	63.24	50.37	61.92	56.74	66.74	50.83	53.41	51.52	63.3	55.21	64.16	56.07	62.29	53.46
IDNN [27]	67.71	52.9	73.71	61.05	86.41	67.85	81.99	64.97	78.33	69.42	71.03	59.7	75.27	62.65
MobileNetV2 [15]	80.19	74.4	82.53	76.5	95.27	85.22	88.65	87.98	87.66	85.92	69.71	56.43	83.46	77.74
FREAK [21]	62.2	53.4	62.4	54.7	66.4	55.1	56.5	50.3	66	54	65.4	52.6	63.01	53.35
AADCL [13]	85.27	79.93	87.75	70.85	95.74	89.62	89.62	87.03	88.79	85.95	71.26	57.14	86.41	78.42
Proposed CLF-AIAD	87.11	81.27	90.79	83.29	96.57	90.23	90.89	89.54	92.2	87.16	72.08	60.6	88.11	82.02

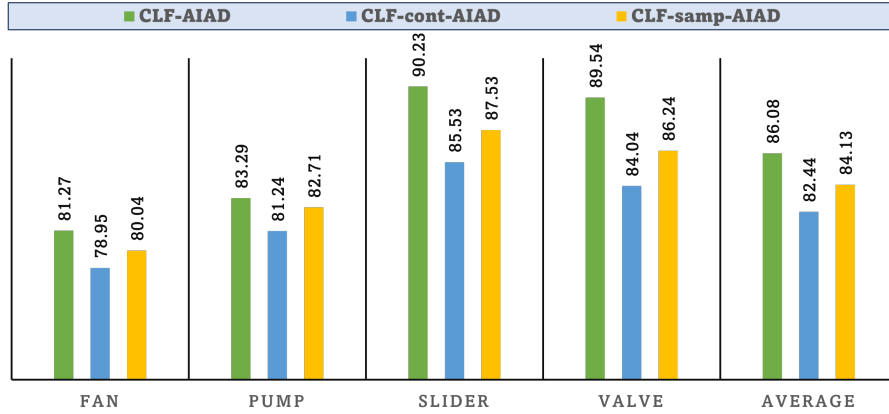


Fig. 4. The pAUC-best (%) results of proposed CLF-AIAD and its variants of CLF-cont-AIAD and CLF-samp-AIAD on four different machine types’ sub-datasets

eliminating the sample-wise contrastive loss. However, CLF-samp-AIAD retains the sample-wise contrastive loss while discarding the contextual contrastive loss.

As demonstrated in Fig. 4, CLF-AIAD performs best on the four machine types’ sub-datasets. CLF-samp-AIAD outperforms CLF-cont-AIAD due to the inherent periodic nature of machine operation, which leads to similarities between distant time frames. The contextual contrastive loss in CLF-cont-AIAD focuses on fine-grained contextual information by maximizing the similarity between neighboring frames and minimizing the similarity between distant frames. However, the generation of periodic acoustic data by machines can lead to similarities between not only neighboring but also more distant frames. And the instance contrastive loss in CLF-samp-AIAD optimizes similarity among the same samples and minimizes similarity between different samples. This allows CLF-samp-AIAD to capture coarse-grained contextual information and local invariance within AID, effectively incorporating the periodic information present in AID.

In summary, using the multi-grained contrasting method in the CLF-AIAD facilitates the acquisition of multi-granularity temporal-dependent and local invariant representations, allowing the model to generate a normal profile accurately and improving the discriminative ability of aberrant AID.

5 Conclusion

In this paper, we propose a novel multi-grained Contrastive Learning-based Framework for AIAD (CLF-AIAD), which can capture dependencies and dynamics in normal AID. CLF-AIAD first incorporates a self-attention layer and a context-based contrast method to learn robust temporal embeddings. Additionally, a sample-wise contrasting method is employed to capture local invariant characteristics, enhancing the discriminative capabilities of the model. We also introduce a transformation classifier within a self-supervised learning framework to improve performance on the primary task. Extensive experiments on the

MIMII and ToyADMOS datasets establish the proposed CLF-AIAD as a new effective method for unsupervised anomaly detection that substantially surpasses the state-of-the-art baseline model by 25.82% in mean AUC score. As a future extension, we will enhance our CLF-AIAD by incorporating the inherent periodic nature of machine operation.

6 Acknowledgements

Research Fund KU Leuven in the context of the ReSOS project (C3/20/014) and by Ford Motor Company in the context of the Ford-KU Leuven Research Alliance project Automated S&R (KUL0134).

References

1. Bigoni, C., Hesthaven, J.S.: Simulation-based anomaly detection and damage localization: an application to structural health monitoring. *Computer Methods in Applied Mechanics and Engineering* **363**, 112896 (2020)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. pp. 1597–1607. PMLR (2020)
3. Cui, Y., Liu, Z., Lian, S.: A survey on unsupervised industrial anomaly detection algorithms. *arXiv preprint: 2204.11161* (2022)
4. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: *Pattern Recognition*. pp. 475–489 (2021)
5. Dohi, K., Endo, T., Purohit, H., Tanabe, R., Kawaguchi, Y.: Flow-based self-supervised density estimation for anomalous sound detection. In: *Proceedings of ICASSP*. pp. 336–340 (2021)
6. Dohi, K., Imoto, K., et al.: Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. *arXiv preprint: 2206.05876* (2022)
7. Eldele, E., Ragab, M., Chen, Z., Wu, M., et al.: Time-series representation learning via temporal and contextual contrasting. pp. 2352–2359 (08 2021)
8. Eltoumy, K., Gomaa, M., Liang, X.: Unsupervised learning methods for data-driven vibration-based structural health monitoring: A review. *Sensors* **23**(6), 3290 (2023)
9. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., et al.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: *Proceedings of ICCV*. pp. 1705–1714 (2019)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of CVPR*. pp. 9729–9738 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of CVPR*. pp. 770–778 (2016)
12. Hojjati, H., Armanfard, N.: Dasvdd: Deep autoencoding support vector data descriptor for anomaly detection. *arXiv preprint: 2106.05410* (2021)
13. Hojjati, H., Armanfard, N.: Self-supervised acoustic anomaly detection via contrastive learning. In: *Proceedings of ICASSP*. pp. 3253–3257 (2022)
14. Jiang, A., Zhang, W.Q., et al.: Unsupervised anomaly detection and localization of machine audio: A gan-based approach. In: *Proceedings of ICASSP*. pp. 1–5 (2023)

15. Kawaguchi, Y., Imoto, K., et al.: Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions. arXiv preprint: 2106.04492 (2021)
16. Kim, M.S., Yun, J.P., Lee, S., Park, P.: Unsupervised anomaly detection of lm guide using variational autoencoder. In: Proceedings of ATEE. pp. 1–5 (2019)
17. Koizumi, Y., Saito, S., Uematsu, H., Harada, N.: Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma. In: Proceedings of EUSIPCO. pp. 698–702 (2017)
18. Koizumi, Y., Saito, S., Uematsu, H., Harada, N., Imoto, K.: Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In: Proceedings of WASPAA. pp. 313–317 (2019)
19. Liu, Y., Garg, S., Nie, J., Zhang, Y., et al.: Deep anomaly detection for time-series data in industrial iot: A communication-efficient on-device federated learning approach. IEEE Internet of Things Journal **8**(8), 6348–6358 (2020)
20. Liu, Z., Tang, H., Michiels, S., Joosen, W., Hughes, D.: Unsupervised acoustic anomaly detection systems based on gaussian mixture density neural network. In: Proceedings of EUSIPCO. pp. 259–263 (2022)
21. Lu, H., Fu, Y., Qin, H., Huang, S., et al.: Anomalous sounds detection using auto-encoder and classification methods. Tech. Rep., DCASE2021 Challenge (2021)
22. McLachlan, G.J.: Mahalanobis distance. Resonance **4**(6), 20–26 (1999)
23. Patcha, A.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks **51**(12), 3448–3470 (2007)
24. Purohit, H., Tanabe, R., et al.: Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. arXiv preprint: 1909.09347 (2019)
25. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **109**(5), 756–795 (2021)
26. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
27. Suefusa, K., Nishida, T., Purohit, H., et al.: Anomalous sound detection based on interpolation deep neural network. In: Proceedings of ICASSP. pp. 271–275 (2020)
28. Taheri, H., Koester, L.W., et al.: In situ additive manufacturing process monitoring with an acoustic technique: clustering performance evaluation using k-means algorithm. Journal of Manufacturing Science and Engineering **141**(4) (2019)
29. Trapp, M., Chen, F.: Automotive buzz, squeak and rattle: mechanisms, analysis, evaluation and prevention. Elsevier (2011)
30. Tuli, S., Casale, G., et al.: Tranad: Deep transformer networks for anomaly detection in multivariate time series data. arXiv preprint: 2201.07284 (2022)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al.: Attention is all you need. Advances in neural information processing systems **30** (2017)
32. Yue, Z., Wang, Y., Duan, J., Yang, T., et al.: Ts2vec: Towards universal representation of time series. In: Proceedings of AAAI. vol. 36, pp. 8980–8987 (2022)
33. Yun, H., Kim, H., Jeong, Y.H., Jun, M.B.: Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor. Journal of Intelligent Manufacturing **34**(3), 1427–1444 (2023)
34. Zhang, W., Yang, D., Peng, H., Wu, W., et al.: Deep reinforcement learning based resource management for dnn inference in industrial iot. IEEE Transactions on Vehicular Technology **70**(8), 7605–7618 (2021)
35. Zhang, Z., Zhao, Z., Zhang, X., et al.: Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. arXiv preprint: 2304.02216 (2023)