Anomalous sound detection (ASD) is one of the most significant tasks of mechanical equipment monitoring and maintaining in complex industrial systems. In practice, it is vital to precisely identify abnormal status of the working mechanical system, which can further facilitate the failure troubleshooting. In this paper, we propose a multi-pattern adversarial learning one-class classification framework, which allows us to use both the generator and the discriminator of an adversarial model for efficient ASD. The core idea is learning to reconstruct the normal patterns of acoustic data through two different patterns of auto-encoding generators, which succeeds in extending the fundamental role of a discriminator from identifying real and fake data to distinguishing between regional and local pattern reconstructions. Furthermore, we present a global filter layer for long-term interactions in the frequency domain space, which directly learns from the original data without introducing any human priors. Extensive experiments performed on four real-world datasets from different industrial domains (three cavitation datasets provided by SAMSON AG, and one existing publicly) for anomaly detection show superior results, and outperform recent state-of-the-art ASD methods.

**Keywords** Anomaly Detection; Cavitation Detection; Acoustic Signals; Global Filter layer; Generative Adversarial Network and Multi-pattern Generator

1 Introduction

Anomaly detection has been extensively studied in different domains [1] (e.g., image, video and time series, etc.). Due to the rare occurrence of anomalous scenarios, the anomaly detection problem is often called one-class classification (OCC) in which only normal data is present and used to learn a novelty detection model. In this paper, we focus on anomalous sound detection (ASD) [2] for cavitation or other faults in complex industrial mechanical systems, which
has been an active research topic in industry during these years. ASD is the task of determining whether the acoustic signal emitted from a target machine is normal or abnormal, which is a fundamental technology in the fourth industrial revolution to monitor the health of machines by ‘listening’ to their acoustics [3].

Conventionally, for anomaly detection in industry, engineering experts manually establish static thresholds for abnormal status based on physical definitions of cavitation [4] or mechanical structure of the target machine [5]. However, this process requires huge human and economic costs as the scale and the complexity of industrial data grows significantly over the years. To tackle this problem, many unsupervised anomaly detection algorithms have been developed for vibration signals or time series [6], where anomalies are detected mainly for one specific state. However, for a complex real-world industrial system, the anomaly states are often complicated, diversified and interacted with each other (e.g., multiple states of cavitation, leakage, contamination, etc.) due to their different mechanical design configurations and different sensors. Therefore, specific single-state anomaly detection is no longer compatible with the current complex real-world industrial system.

Formally, ASD data consist of multiple measured signal events, each of which describes an anomalous state of a complex entity and each records the entire physical process from the beginning to the end of an anomalous state. Thus, there is global long-term dependency of each entity measurement event. Violating the dependency of each entity measurement event would lead to degradation of anomaly detection performance. To help the system operates better in anomalies monitoring for complex industrial mechanical equipment, the algorithm should consider the global long-term dependency of measurement events [7].

Recent methods for ASD can be roughly divided into two classes: encoder-decoder based reconstruction and adversarial training based reconstruction. Encoder-decoder-based methods [8] try to train a model to produce good quality reconstructions of input signals and detect abnormal data based on reconstruction loss. For adversarial training-based approaches [9], it can substantially improve the quality of reconstruction. At test time, the trained generator $G$ is then uncoupled from the discriminator $D$ to be used as a reconstruction model. As mentioned in [10], the reconstruction loss between normal and abnormal data is significantly different due to the adversarial training, which gives rise to good anomaly detection results. However, in some cases currently using only the reconstruction capability of a generator $G$ is useless because of the spillover of the reconstruction capability of the generator $G$, which thus largely degrades the performance of anomaly detection.

Based on the above considerations, for industrial acoustic signals anomaly detection, we propose to explicitly learn global long-term dependencies and multi-pattern high quality reconstructions to better capture the normal patterns of industrial acoustic signals. To address the global long-term dependence problem, a global filter layer is introduced and applied to the spectrum of the input features. And, the global filter layer is directly learned from the original sub-sequence acoustic signals without introducing extra human priors. Since the global filter layer is able to cover all frequencies, our method therefore capture the global long-term interactions. To address multi-pattern high quality reconstructions, we devise an adversarial learning approach with multiple generators and a single discriminator (see Figure 1). The two generators $G_{\text{local}}$ and $G_{\text{regional}}$ try to generate multi-pattern real-looking fake data, and they focus on the local and regional details of acoustic signals under the global long-term space, respectively. Furthermore, the capability of discriminator $D$ is extended from distinguishing between real and fake to differentiating between local and regional reconstructions, making it more suitable for anomaly detection task on industrial acoustic signals. In addition, we break the conventional idea of using only $G$ or $D$ for anomaly detection. In our method, the $G_{\text{local}} + G_{\text{regional}} + D$ are used simultaneously for anomaly detection. Finally, the above form our data-driven adversarial training anomaly
detection approach, namely Regional-Local Adversarial Learning One-Class Classifier Anomalous Sound Detection in Global Long-Term Space (GRLNet) for cavitation detection of acoustic signals in complex industrial systems.

The contributions of this paper are summarised as follows:

- To the best of our knowledge, our proposed GRLNet is the first adversarial learning one-class classifier for industrial acoustic signals anomaly event detection which employs \( \mathcal{D} \) along with \( \mathcal{G} \).
- We propose a global filter layer based on a global filter (1D FFT) for long-term interactions in the frequency domain space, which directly learns from the original data without introducing extra human priors.
- We extend the capability of \( \mathcal{D} \) from discriminating real signals against fake ones to differentiating between local and regional reconstructions by producing different patterns of reconstructed signals using \( \mathcal{G}^{local} \) and \( \mathcal{G}^{regional} \).
- Our proposed GRLNet outperforms state-of-the-art methods in the experiments conducted on four real-world acoustic datasets from different industrial domain (our own three cavitation datasets provided by SAMSON and one public dataset). Moreover, ablation studies further demonstrate the effectiveness of our proposed structure design for industrial acoustic anomaly detection. We publish our code for better reproducibility.

2 PRELIMINARES

2.1 Abnormal Cavitation Event Detection

Cavitation is defined as the phenomenon that bubbles form on the solid surface when the pressure at the contact point of the liquid and the solid surface is lower than its vapor pressure \([11][12]\). For working valves, the cavitation acoustic signals of four different flow states are obtained as continuous waveforms using acoustic sensors as shown in Figure 2. Each of the observed acoustic signals record the whole physical process from the beginning to the end of the event of corresponding cavitation state. For complex industrial systems, cavitation can be roughly classified into three types: cavitation choked flow, constant cavitation and incipient cavitation (see (a)-(c) in Figure 2). For incipient cavitation, only a few vapor bubbles are generated inside pipes or valves and bubbles implode with noise, which indicates the beginning of an abnormal system. In constant cavitation, more vapor bubbles are produced from interior of pipes or valves and accompanied by a certain level of noise, which means the system is completely abnormal. When both vapor bubbles concentration and noise inside pipes or valves reach maximum value which leads to a choked flow condition with cavitation. At this situation, the system is almost impossible to operate and most of abnormalities are actually easier to be detected. No matter severe or subtle, each type of cavitation can indicate a potential problem in the system. Therefore, precisely detecting such abnormal cavitation events is urgently requested for operators of industrial systems.

2.2 Acoustic Signals Augmentation

Formally, we suppose there is \( x \subseteq \mathbb{R}^{M \times N} \) with \( M \) measurements for each acoustic signal. Considering purposely steady flow status (i.e. it’s always the same fluid status class within the individual measurement duration with 3 sec or 25 sec time-length) in each recorded stream and fine resolution for the sensor. One can actually split each stream into several pieces which still can hold enough essential information for detection. And, with independent features per piece due to the intrinsic randomness of the noise emission, given the piece is not so short. Therefore, according to the above theory, we apply a sliding window with window size \( s_w \) and step size \( s_s \) to divide the acoustic signal sequence into a set of sub-sequences \( X = \{x_{i,j}, i = 1, 2, \ldots, N; j = 1, 2, \ldots, k\} \subseteq \mathbb{R}^{s_w \times N} \), where \( k = \frac{(M-s_w)}{s_s} \) is the number of sub-sequences (as shown in the top left of Figure 3) and \( N \) is stream.

2.3 Discrete Fourier Transform

The Discrete Fourier Transform (DFT) \([13]\) converts a signal from time domain space to frequency domain space, which is an essential part of the global filtering layer in GRLNet. Given a sequence of signal \( \{x_n\}_{n=0}^{N-1} \), the 1D DFT converts the signal sequence into the frequency domain is defined by the formula:

\[
X[k] = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi j n k}{N}} := \sum_{n=0}^{N-1} x_n W_N^{kn} \tag{1}
\]

where \( j \) is the imaginary unit and \( W_N = e^{-\frac{2\pi j}{N}} \). The formulation of 1D DFT of Equation 1 can be derived from the Fourier transform for continuous signals (see Appendix A for details). The 1D DFT generates a sequence \( X[k] \) as
Figure 2: Acoustic signals in (a)-(d) for cavitation choked flow, constant cavitation, incipient cavitation and non-cavitation, respectively.

3 Method

3.1 Architecture Overview

The structure of the generator and discriminator in GRLNet is similar to those proposed by Sabokrou et al [10]. The generator is a 1D convolutional autoencoder, which is coupled with a discriminator to learn one class of data in an unsupervised adversarial manner. GRLNet has two generators, $G_{local}$ and $G_{regional}$, and a discriminator $D$ (more details see Appendix B). The $G_{local}$ and $G_{regional}$ focus on local and regional reconstruction of the signal, respectively, which can capture different patterns of the real signal distribution at the same time. The training of the model is using min-max strategy to optimize the following overall objective function.

$$
\min_{\theta_g} \max_{\theta_d} \mathbb{V}(\theta_d, \theta_g) := \mathbb{E}_{X \sim p_t}[\log(1 - D(X; \theta_d))] + \mathbb{E}_{\tilde{X} \sim p_t + N_\sigma}[\log(D(G(\tilde{X}; \theta_g); \theta_d))]
$$

where $X$ denotes a set of true samples with distribution $p_t$; $\tilde{X}$ is the input sample $X$ with added noise $N_\sigma$; $\theta_g = (\theta_{g_{local}}, \theta_{g_{regional}})$ and $\theta_d$ represent learnable parameters of the $G$ and $D$, respectively. The above objective function is optimised in a block-wise manner, i.e. one of $\theta_g$ and $\theta_d$ is optimised while the other is fixed. Theoretically, the generator could produce real samples in Nash-equilibrium, which means $p_t = p_g$ ($p_g$ is distribution from the generator).

The overall architecture of our model is depicted in Figure 3. In our model, we take the non-overlapping acoustics signal patches of length $L$ obtained through the sliding window as input. The basic building blocks of GRLNet...
Figure 3: Our proposed GRLNet framework. In the training phase, the segmented acoustics signals are fed through the global filter layer (involving layer normalization and 1D FFT) to $G_{\text{regional}}$, $G_{\text{local}}$ and $D$ to obtain a reasonably trained state of $G_{\text{regional}}$, $G_{\text{local}}$ and $D$. Good quality samples correspond to real training of segmented acoustics signals, local reconstructed acoustic signals acquired by $G_{\text{local}}$ and regional reconstructed acoustic signals obtained from $G_{\text{regional}}$. During testing, the to be detected segmented acoustics signals are inferred through $G_{\text{regional}}$, $G_{\text{local}}$ and $D$ after passing through the global filter layer and the output of $D$ is considered as anomaly score.

3.2 Training

Given a training dataset $\mathcal{X}^{\text{train}} \subseteq \mathcal{R}^{M \times N}$ and a test dataset $\mathcal{X}^{\text{test}} \subseteq \mathcal{R}^{M \times N}$ with $N$ streams and $M$ measurements for each stream. Our task is to assign binary labels to the acoustic signals of test dataset. Note that the acoustic signals in the training dataset are normal.

In the training stage, the proposed GRLNet is similar to common practices in training adversarial one-class classifiers [15, 16]. $X$ is passed through a global filter layer to obtain global frequency features of acoustics signals. Then, $G_{\text{local}}$ attempts to regenerate real-looking fake data which focuses on the local details of acoustics signals and $G_{\text{regional}}$ tries to regenerate real-looking fake data which focuses on the regional details of acoustics signals, which are then fed into $D$ along with real data. Finally, $D$ learns to discriminate between real and fake data, and also learns to reconstruct acoustics signals with different patterns generated by $G_{\text{local}}$ and $G_{\text{regional}}$. In general, this stage minimises the following loss function:

$$L = \lambda (L_{G+D}) + (1 - \lambda) L_{R_{\text{total}}}$$

(3)

where $L_{G+D} = L_{G_{\text{local}}+D} + L_{G_{\text{regional}}+D}$ is the loss function of our joint training, and the overall objective defined in Equation 2

$L_{R_{\text{total}}} = L_{R_{\text{local}}} + L_{R_{\text{regional}}}$ is the total reconstruction loss with $L_{R_{\text{local}}} = \| X - G(\hat{X}) \|^2$ is the local reconstruction loss and $L_{R_{\text{regional}}} = \| X - G(\hat{X}) \|^2$ is the regional reconstruction loss; $\lambda$ is a weighing hyperparameter, which is usually set 0.5.

Global Filter Layer. We propose the global filter layer to extract global frequency features, which can mix subsequences representing different spatial locations. Given the sub-sequences $x \in X$, we first apply layer normalization [17] to normalize $x$:

$$\hat{x} = LN[x]$$

(4)
We conduct extensive experiments on four different datasets (three cavitation datasets are provided by SAMSON AG where, at $k$ is the comprehensive output factor to weight the output of $\hat{x}$. The final classification decision for an input acoustic signal is given as follows:

$$\text{OCC} = \begin{cases} \text{Normal signal} & \text{if } \beta \mathcal{D}(\mathcal{G}^{\text{regional}}(X)) + (1 - \beta)\mathcal{D}(X) < \tau \\ \text{Anomaly signal} & \text{otherwise} \end{cases}$$

where $\beta$ is comprehensive output factor to weight the output of $\mathcal{D}$ under different patterns and $\tau$ is a predefined threshold. The proposed GRLNet is summarized in Algorithm 1.

### 3.3 Testing

In the test stage, as shown on the right in Figure 3, $\mathcal{G}^{\text{regional}}$ and $\mathcal{D}$ are utilised for one-class classification (OCC) [15, 16]. The final classification decision for an input acoustic signal is given as follows:

$$\text{OCC} = \begin{cases} \text{Normal signal} & \text{if } \beta \mathcal{D}(\mathcal{G}^{\text{regional}}(X)) + (1 - \beta)\mathcal{D}(X) < \tau \\ \text{Anomaly signal} & \text{otherwise} \end{cases}$$

where $\beta$ is comprehensive output factor to weight the output of $\mathcal{D}$ under different patterns and $\tau$ is a predefined threshold. The proposed GRLNet is summarized in Algorithm 1.

### 4 Experiments

We conduct extensive experiments on four different datasets (three cavitation datasets are provided by SAMSON AG and a public dataset) to validate the effectiveness of the proposed GRLNet. And detailed analysis of the performance and its comparison with state-of-the-art methods [6, 20, 21, 22, 8] are also reported. In addition, we provide extensive discussions and ablation experiments to demonstrate the stability and importance of the proposed GRLNet.

**Evaluation criteria.** We calculate an anomaly score for each sample $X$, where samples with higher scores are more anomalous. Following previous researches [23, 24, 25], we select the AUC score of ROC curve area [26] as an evaluation metric to determine the degree of separability of samples. It measures the one-class classification performance of the model among all possible thresholds and distinguishes between normal and anomaly events. The higher ROC-AUC
score, the better the model. A ROC-AUC score which is closer to 1.0 indicates the model can perfectly differentiate between normal and anomaly events. In addition, we also use F1-score as another evaluation criterion. It represents the average of precision and recall for a specific threshold value. The threshold is determined through Youden index in ROC-AUC (more details see Appendix B).

### 4.1 Datasets

**Cavitation Datasets.** This dataset is provided by SAMSON AG and contains three sub-datasets, called cavitation2017, cavitation2018 and cavitation2018 with real noise, respectively. The cavitation acoustic signals are collected from different valves with different upstream pressures and different valve opening rates in a professional environment. The cavitation includes incipient cavitation, constant cavitation and cavitation choked flow. The non-cavitation contains turbulent flow and no flow. Cavitation2017 has a total of 356 acoustic signals and each acoustic signal with time duration of 3 s. Cavitation2018 and Cavitation2018-noise have 806 and 160 acoustic signals and each of them with time duration of 25 s, respectively. The sampling rate of all recording acoustic signals is 1562.5 kHz (more details see Appendix B).

**MIMII Dataset.** This dataset [27] comprises normal and abnormal acoustic signals from four industrial machine types: fan, pump, slider and valve. It records anomalous acoustic signals for different conditions, such as contamination, leakage, clogging, voltage change, rail damage, losses belt, no grease, etc. Each industrial machine type consists of four individual machines, named ID 00, ID 02, ID 04 and ID 06. Considering real industrial noise, each individual machine has three different levels of signal-to-noise ratio (-6 dB, 0 dB and 6 dB) data. The sampling rate for all acoustic signals is 16 kHz and each of them has a time duration of 10 s. The MIMII dataset is split into a training set and a test set. The
test set is consisted of all abnormal acoustic signals and the same number of normal acoustic signals. The other normal acoustic signals are regarded as the training set (more details see Appendix B).

4.2 Anomaly Detection Results

Anomaly detection is one of the significant applications of one-class classification learning algorithm. In unsupervised one-class anomaly detection problem, events belonging to known classes are regarded as normal values for the training of the model. Other events that do not belong to the known classes are treated as anomalous values for model determination.

![Diagram](image)

Figure 4: Example acoustic signals from different stages of our framework (Frequency (kHz) as the $x$-axis and normalized amplitude as the $y$-axis). From top to bottom: The output signals from global filter layer $X$, $\tilde{X}$ shows the signal with noise, local pattern reconstructed $G_{local}(\cdot)$ and regional pattern reconstructed $G_{regional}(\cdot)$. Blue and red colors represent normal and abnormal, respectively.

**Results on Cavitation datasets.** Figure 4 shows anomalies examples reconstructed using $G_{local}$ and $G_{regional}$. It indicates that $G_{local}$ and $G_{regional}$ can better reconstruct normal signals (non-cavitation) and cannot reconstruct abnormal signals (cavitation). As seen in Table 1, GRLNet shows superior results in terms of AUC compared to other state-of-the-art methods. And the proposed GRLNet retains superior and robust performance even with an increased percentage of anomalies at test time. The proposed GRLNet achieves an average of $0.9855$, $0.9705$ and $0.9999$ AUC in the three cavitation datasets, respectively.

Detailedly, (1) **Cavitation2017 dataset:** The AUC of GRLNet is above $0.975$ for each specific abnormal proportion, which outperforms over all baselines. (2) **Cavitation2018 dataset:** The AUC of GRLNet is above $0.96$ as the proportion of abnormalities increases, and it improves on average by $0.0357$ over the standard GAN. (3) **Cavitation2018-noise dataset:** Our method obtains perfect performance in all cases with different anomaly proportions. The reasons why our method gets superior results are as follows. First, although this dataset with real background noise compared to other cavitation datasets, the global filter layer can filter most of the noise. Second, this dataset is obtained with only one operating of valve stroke and upstream pressure (see Appendix B).
In addition, we also report $F_1$ score as an evaluation metric of our method on three real-world cavitation datasets. A comparison provided in Figure 5 shows that our approach performs robustly to detect anomalies even when the percentage of anomalies is increased.

| Dataset       | Cavitation2017 | Cavitation2018 | Cavitation2018-noise |
|---------------|----------------|----------------|-----------------------|
| Abnormal percentage | 10 20 30 40 50 | 10 20 30 40 50 | 10 20 30 40 50 |
| AE [8]        | 0.8981 0.8871 0.8862 0.8821 0.8816 | 0.8881 0.8876 0.8624 0.8549 0.8483 | 0.9378 0.9346 0.9282 0.9218 0.9178 |
| VAE [22]      | 0.9398 0.9387 0.9340 0.9290 0.9152 | 0.9383 0.9319 0.9299 0.9159 0.9158 | 0.9498 0.9474 0.9442 0.9430 0.9424 |
| GAN [20]      | 0.9449 0.9424 0.9382 0.9365 0.9262 | 0.9462 0.9419 0.9283 0.9299 0.9276 | 0.9578 0.9524 0.9510 0.9499 0.9499 |
| GRLNet        | 1.0000 0.9901 0.9829 0.9790 0.9756 | 0.9918 0.9716 0.9634 0.9633 0.9622 | 1.0000 0.9999 0.9999 0.9999 0.9999 |

Figure 5: $F_1$ score results on three real-world cavitation datasets. The window size is 466944 and step size is 466944.

Results on MIMII. Table 2 shows the experimental results of best AUCs. In general, GRLNet achieves higher AUCs compared to AE, VAE, CAE, CVAE and GAN for each industrial machine type consisting of four individual machines. We observe that most methods achieve high detection performance on fan, pump, slider and valve, but the GRLNet’s best-AUC still outperforms them by 0.0004 - 0.0188.

Detailedly, (1) **Fan:** For 0 dB SNR, GRLNet achieves excellent performance for all machines. In addition, GRLNet gets better AUCs than all baselines for ID 02, ID 04 and ID 06 at 6 dB SNR. And AE shows the higher performance at ID 00. For -6 dB SNR, AE and VAE achieve the better performance at ID 02 and ID 04, respectively. The AUCs of GRLNet are improved by 0.0188 and 0.0088 compared to other methods at ID 00 and ID 06, respectively. (2) **Pump:** The proposed GRLNet outperforms other methods, except for some certain cases. Specifically, the CVAE obtains the better performance for ID 00 and ID 06 at -6 dB SNR, and ID 00 at 6 dB SNR. Meanwhile, GAN has a higher AUC for ID00 at 0 dB SNR than GRLNet. (3) **Slider:** The AE exhibits better performance for ID 04 at both -6 and 6 dB SNRs. CAE shows higher AUCs at ID 04 (0 and 6 dB SNRs). The GAN also has a higher AUC at ID 00 (-6 dB SNR). In other cases, the proposed GRLNet obtains the best performance compared to other baselines, which is improved on average by 0.0332, 0.0484, 0.0415, 0.0203 and 0.0221 compared to AE, VAE, CAE, CAVE and GAN, respectively. (4) **Valve:** The performance of the GRLNet and all baselines in valves are significantly lower than other machines due to the instability of valve signals and complex anomalous conditions (more than two kinds of contamination). However, the GRLNet achieves better performance compared to other methods on over half of all machines with different SNRs.

From Table 3, it can be seen that the GRLNet gets the highest performance at all SNRs of the pump. In addition, the GRLNet outperforms other methods for 0dB and 6dB SNRs of the fan, -6dB and 6dB SNRs of the slider, and 0dB SNR for the valve. For a few SNR-specific machines, the AVE, CAE and CVAE outperform our method at the best AUC average.

**Takeaways:** The proposed GRLNet achieves the best results for four real-world datasets. This is motivated by two following reasons. Firstly, this ASD task is evaluated based on event anomaly detection of acoustic segments rather than on acoustic frames. Secondly, the global filter layer and the different pattern generators ($G^{\text{regional}}$ and $G^{\text{local}}$) make the GRLNet keep both the global spatial acoustic features and the different detailed (local and regional) acoustic features.
Table 2: Comparison with anomaly detection architectures on MIMII. We report the best AUC for each industrial machine type (Fan, Pump, Slider and Valve) consisting of four individual machines (ID 00, ID 02, ID 04 and ID 06) with three different levels of signal-to-noise ratio (-6 dB, 0 dB and 6 dB).

| Input SNR | -6 dB | 0 dB | 6 dB |
|-----------|-------|------|------|
| 00        | 02    | 04   | 06   | 00    | 02    | 04   | 06   | 00    | 02    | 04   | 06   |
| AE        | 0.5629 | 0.7182 | 0.5957 | 0.8723 | 0.6414 | 0.9266 | 0.8327 | 0.9932  | 0.8559 | 0.9924 | 0.9635 | 1.0000 |
| VAE 23    | 0.5676 | 0.7054 | 0.7054 | 0.8802 | 0.6771 | 0.9297 | 0.7942 | 0.9943  | 0.8285 | 0.9949 | 0.9671 | 1.0000 |
| CAE 21    | 0.5462 | 0.6341 | 0.5839 | 0.7452 | 0.5062 | 0.7762 | 0.7373 | 0.9230  | 0.6773 | 0.9322 | 0.8190 | 0.9912 |
| GAN 20    | 0.5417 | 0.6639 | 0.6125 | 0.8826 | 0.6538 | 0.9188 | 0.8315 | 0.9829  | 0.8147 | 0.9869 | 0.9615 | 0.9943 |

GRLNet 0.5864 | 0.6874 | 0.6321 | 0.8914 | 0.6857 | 0.9468 | 0.8375 | 0.9949  | 0.8359 | 0.9996 | 0.9781 | 1.0000 |

Table 3: Comparison of best AUC averages on MIMII.

| Machine | Input SNR | AE | VAE | CAE | CV AE | GAN | GRLNet |
|---------|-----------|----|-----|-----|-------|-----|--------|
|         | -6 dB     |    |     |     |       |     |        |
| Fan     | 0.5629    |    |     |     |       |     |        |
|         | 0.7182    |    |     |     |       |     |        |
|         | 0.5957    |    |     |     |       |     |        |
|         | 0.8723    |    |     |     |       |     |        |
| Pump    | 0.5676    |    |     |     |       |     |        |
|         | 0.7054    |    |     |     |       |     |        |
|         | 0.7054    |    |     |     |       |     |        |
|         | 0.8802    |    |     |     |       |     |        |
| Slider  | 0.5462    |    |     |     |       |     |        |
|         | 0.6341    |    |     |     |       |     |        |
|         | 0.5839    |    |     |     |       |     |        |
|         | 0.7452    |    |     |     |       |     |        |
| Valve   | 0.5417    |    |     |     |       |     |        |
|         | 0.6639    |    |     |     |       |     |        |
|         | 0.6125    |    |     |     |       |     |        |
|         | 0.8826    |    |     |     |       |     |        |
4.3 Ablation Analysis

**Detection strategy.** We report different detection strategies in Table 4. It can be seen from Table 4 that our detection strategy for GRLNet significantly outperforms the other detection strategies. Where, $D(G_{local}(X)) + D(G_{regional}(X))$ and $D(X)$ and are special cases of our detection strategy $\beta D(G_{local}(X)) + D(G_{regional}(X)) + (1 - \beta)D(X)$ when $\beta = 1.0$ and $\beta = 0$, respectively.

| Detection Strategies                  | Type | Results |
|---------------------------------------|------|---------|
| $D(X)$                                | ✓    | 0.6333 0.7317 |
| $G_{local}(X)$                        | ✓    | 0.5167 0.6667 |
| $G_{regional}(X)$                     | ✓    | 0.5258 0.6667 |
| $D(G_{local}(X))$                     | ✓    | 0.7859 0.8067 |
| $D(G_{regional}(X))$                  | ✓    | 0.8356 0.8487 |
| $\beta D(G_{local}(X)) + D(G_{regional}(X)) + (1 - \beta)D(X)$ | ✓ | 0.9570 0.9655 |

Figure 6: AUC and $F_1$ score results of various ablation experiments. (a)-(c) are all experimented in Cavitation2017 with anomaly percentage of 50%.

**Analysis of window size.** We compare the anomaly detection results of our method with different window sizes in cavitation2017 (anomaly percentage of 50%). It is clear from Figure 5a that the window size is sensitive to the detection results of GRLNet.

**Parameter sensitivity.** We show the impact of different values of the comprehensive output factor $\beta$ on the detection results of the proposed GRLNet in cavitation2017 (anomaly percentage of 50%), see 6b. When $\beta \in (0, 1)$, the GRLNet is not sensitive to $\beta$ and the detection strategy is $D(G_{local}(X)) + D(G_{regional}(X)) + D(X)$ in this case. In general, we set $\beta = 0.5$. Where, the detection strategy is $D(X)$ when $\beta = 0$; when $\beta = 1.0$, the detection strategy is $D(G_{local}(X)) + D(G_{regional}(X))$.

**Analysis of downsampling.** In practical applications, the resolution of sensors represent the quality of the data obtained. The ability to detect different levels of the data obtained from low-level sensors becomes very significant and challenging. Our method is evaluated in cavitation2017 (anomaly percentage of 50%) under the original frequency of samples ($F_s$ = 1562500 Hz), one-half, one-quarter, one-sixth, one-eighth of the original frequency of samples (781250 Hz, 390625 Hz, 260416 Hz and 195312 Hz) and that of the mobile phone can mostly accommodate (48000 Hz $\approx F_s/32$), respectively. As shown in Figure 6c, although the proposed GRLNet begins to suffer from a gradual reduction in AUC and $F_1$ score as the frequency of samples decreases, the AUC and $F_1$ score always remain above 0.96 when 1562500 Hz $\geq F_s \geq 195312$ Hz. And even for mobile phone setup with the frequency of samples equals to 48000 Hz, the proposed GRLNet achieves an AUC and $F_1$ score of 0.9611 and 0.9655, respectively.

**Individual abnormal state detection capability.** We compare the detection capability of GRLNet in individual cavitation states in Cavitation2017 (anomaly percentage of 50%), i.e., the anomalies all consist of a individual cavitation state. It can be seen from Figure 6d that both continuous and choked cavitation can be detected well. However, the incipient cavitation is not well detected since it is the critical state between cavitation and non-cavitation.
5 Related Work

**Anomaly Detection.** Anomaly detection is often regarded as a novelty detection problem in which is trained based on a known normal class and ultimately unknown outliers as abnormal. Data-based regeneration works by using generative networks to learn features in an unsupervised manner. Zhou et al. in [28] proposed to use deep autoencoder to learn the reconstruction of normal images. Lonescu et al. in [29] to use convolutional auto-encoder based on object detection to learn motion and appearance representations. Su et al. in [25] proposed a multivariate time series anomaly detection based on stochastic recurrent neural networks. Xu et al. in [30] used a one-class SVM using features from stacked auto-encoders. Ravanbakhsh et al. [31] used generators as a reconstructor to detect anomalous events, assuming that the generator is unable to reconstruct inputs that do not match the normal training data. Pathak et al. [32] proposed adversarial training to improve the quality of regeneration. However, they also discarded the discriminator at the end of training.

**Fourier transform in network.** Fourier transform has been an important tool for digital signal and image processing for decades [13, 33]. With the breakthroughs of CNNs in signal and vision, there are various works that start to incorporate Fourier transform into deep learning methods. Some of these studies employ discrete Fourier transform to convert images to the frequency domain and leverage frequency information to improve performance on certain tasks [34, 35]. There are also some studies that use the convolution theorem of the Fourier transform to accelerate CNNs [36].

6 Conclusions

ASD is an essential task for complex industrial system monitoring. In this paper, we propose GRLNet, a novel unsupervised anomaly detection method, which tackles global long-term interactions using a global filter layer and captures the normal patterns of acoustic signals employing multi-pattern generators. Furthermore, we extend the fundamental role of a discriminator from identifying real and fake data to distinguishing between regional and local reconstructions. Moreover, we use both generators and discriminators for acoustic anomaly detection, which breaks the tradition of using only generator and discriminator for acoustic anomaly detection. GRLNet outperforms the SOTA methods on four real-world dataset. The ablation experiments clearly show that GRLNet successfully meets the anomaly detection requirements of complex industrial systems in the real world, which provides insights in applying the method on other industrial applications.

Acknowledgements

This research is supported by Xidian-FIAS International Joint Research Center (Y. S.), by the AI grant at FIAS through SAMSON AG (J. F., K. Z.), by the BMBF funding through the ErUM-Data project (K. Z.), by SAMSON AG (D. V., T. S., A. W.), by the Walter Greiner-Gesellschaft zur Förderung der physikalischen Grundlagenforschung e.V. through the Judah M. Eisenberg Lau-reatus Chair at Goethe Universität Frankfurt am Main (H. S.), by the NVIDIA GPU grant through NVIDIA Corporation (K. Z.).

References

[1] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407, 2019.
[2] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Yuta Kawachi, and Noboru Harada. Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(1):212–224, 2018.
[3] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiaki Nakamura, Yuki Nkaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, et al. Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. arXiv preprint arXiv:2006.05822, 2020.
[4] Jean-Pierre Franc and Jean-Marie Michel. Fundamentals of cavitation, volume 76. Springer science & Business media, 2006.
[5] Yu Wei, Yuqing Li, Minqiang Xu, and Wenhu Huang. A review of early fault diagnosis approaches and their applications in rotating machinery. Entropy, 21(4):409, 2019.
Running Title for Header

[6] Minh-Hieu Nguyen, Duy-Quang Nguyen, Dinh-Quoc Nguyen, Cong-Nguyen Pham, Dai Bui, and Huy-Dung Han. Deep convolutional variational autoencoder for anomalous sound detection. In 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), pages 313–318. IEEE, 2021.

[7] Yoshua Bengio, Paolo Frasconi, and Patrice Simard. The problem of learning long-term dependencies in recurrent networks. In IEEE international conference on neural networks, pages 1183–1188. IEEE, 1993.

[8] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In Proceedings of ICML workshop on unsupervised and transfer learning, pages 37–49. JMLR Workshop and Conference Proceedings, 2012.

[9] Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training. arXiv preprint arXiv:2105.14240, 2021.

[10] Mohammad Sabokrou, Mohammad Khalooei, Mahmoud Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3379–3388, 2018.

[11] A Masjedian Jazi and H Rahimzadeh. Detecting cavitation in globe valves by two methods: Characteristic diagrams and acoustic analysis. Applied Acoustics, 70(11-12):1440–1445, 2009.

[12] Marion Bonnier and Olivier Eiff. Experimental investigation of the collapse of a turbulent wake in a stably stratified fluid. Physics of fluids, 14(2):791–801, 2002.

[13] Vijay Madisetti. The digital signal processing handbook. CRC press, 1997.

[14] S Lennart Johnsson and Robert L Krawitz. Cooley-tukey fft on the connection machine. Parallel Computing, 18(11):1201–1221, 1992.

[15] Zhiwei Wang, Zhengzhang Chen, Jingchao Ni, Hui Liu, Haifeng Chen, and Jiliang Tang. Multi-scale one-class recurrent neural networks for discrete event sequence anomaly detection. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 3726–3734, 2021.

[16] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1):1–39, 2012.

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[18] Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8513–8521, 2018.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[20] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. arXiv preprint arXiv:1802.06222, 2018.

[21] Robert Müller, Fabian Ritz, Steffen Illium, and Claudia Linnhoff-Popien. Acoustic anomaly detection for machine sounds based on image transfer learning. arXiv preprint arXiv:2006.03429, 2020.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[23] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 3395–3404, 2020.

[24] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pages 387–395, 2018.

[25] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2828–2837, 2019.

[26] Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data mining and knowledge discovery, 30(4):891–927, 2016.

[27] Harsh Purohit, Ryotanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. arXiv preprint arXiv:1909.09347, 2019.
[28] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 665–674, 2017.

[29] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric autoencoders and dummy anomalies for abnormal event detection in video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7842–7851, 2019.

[30] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 156:117–127, 2017.

[31] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In 2017 IEEE International Conference on Image Processing (ICIP), pages 1577–1581. IEEE, 2017.

[32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2536–2544, 2016.

[33] Ioannis Pitas. Digital image processing algorithms and applications. John Wiley & Sons, 2000.

[34] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. Advances in Neural Information Processing Systems, 34, 2021.

[35] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4085–4095, 2020.

[36] Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David Wei, and Tao Wan. Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8705–8714, 2020.
A Discrete Fourier transform

A.1 From continuous Fourier transform to discrete Fourier transform

Discrete Fourier transforms (DFT) can be derived in many ways. Since the Fourier transform (FT) is originally designed for continuous signals. Therefore, we will derive the DFT from the one-dimensional continuous Fourier transform (CFT). Specifically, given a continuous time signal \( f(t) \), 1D CFT is given by:

\[
F(u) = \int_{-\infty}^{+\infty} f(t) e^{-2j\pi u t} dt
\]

(A.1)

The continuous function \( f(t) \) is sampled once at a time interval \( \Delta T \) from a certain time (noted as moment 0) and a total of \( N \) times. Then, we get a discretised sequence:

\[
f(\Delta T), f(2\Delta T), \ldots, f(N\Delta T)
\]

(A.2)

and the discrete sequence is denoted as:

\[
\hat{f}(0), \hat{f}(1), \ldots, \hat{f}(N-1)
\]

(A.3)

where for the other unsampled points, i.e., \( \forall t \notin \{\Delta T, 2\Delta T, \ldots, N\Delta T\} \), the value of the function for these points are 0, i.e. \( f(t) = 0 \). Applying equations \( A.1 \) and \( A.2 \), we have

\[
F(u) = \int_{-\infty}^{+\infty} f(t) e^{-2j\pi u t} dt
\]

\[
= \int_{-\infty}^{0} f(t) e^{-2j\pi u t} dt + \int_{0}^{+\infty} f(t) e^{-2j\pi u t} dt
\]

\[
= \int_{-\infty}^{0} f(t) e^{-2j\pi u t} dt + \int_{0}^{N\Delta T} f(t) e^{-2j\pi u t} dt + \int_{N\Delta T}^{+\infty} f(t) e^{-2j\pi u t} dt
\]

\[
= \int_{0}^{N\Delta T} f(t) e^{-2j\pi u t} dt
\]

(A.4)

Next, we divide the interval \( [0, N\Delta T] \) equally into \( N \) parts, the \( i \)-th interval \( k_i = [(i-1) \Delta T, i \Delta T] \), \( i = 1, 2, \ldots, N \) and the length of each interval is \( \Delta T \). Let the value of the function taking the right endpoint on each interval \( k_i \) be \( f(i\Delta T) e^{-2j\pi u (i\Delta T)} \). Furthermore, we can convert \( F'(u) \) to the Riemann sum form of \( F''(u) \) as follows

\[
F''(u) = \sum_{i=1}^{N} \frac{f(i\Delta T) e^{-2j\pi u (i\Delta T)}}{\Delta T} \cdot \Delta T
\]

\[
= \sum_{n=0}^{N-1} f((n+1)\Delta T) e^{-2j\pi u ((n+1)\Delta T)} \cdot \Delta T
\]

\[
= \sum_{n=0}^{N-1} \hat{f}(n) e^{-2j\pi u ((n+1)\Delta T)} \cdot \Delta T
\]

(A.5)

If we let the length between the start and end times of sampling be 1 time unit, i.e. \( N\Delta T = 1 \), then we have:

\[
F''(u) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{f}(n)e^{-2j\pi u \frac{n+1}{N}}, u = 0, 1, \ldots, N - 1
\]

(A.6)

which is exactly the formulation of the DFT.

A.2 The convolutional theorem

The convolution theorem is the main property of the Fourier transform. Specifically, the FT for the circular convolution of two discrete sequences is equivalent to the dot product of these two sequences in the frequency domain. Given two discrete sequences \( x[n] \) (signals) and \( h[n] \) (filters), both of length \( N \), the circular convolution can be defined as:

\[
w[n] = (h * x)_n = \sum_{m=0}^{N-1} h[m] x[(n - m)_{mod N}]
\]

(A.7)
where $\text{mod}$ denotes modulo operation and $\ast$ is convolution symbol. Consider the DFT of $w[n]$, we have:

$$
W[k] = \sum_{n=0}^{N-1} \sum_{m=0}^{N-2} h[m] x[(n - m) \mod N] e^{-j(2\pi/N)kn}
$$

$$
= \sum_{m=0}^{N-1} h[m] e^{-j(2\pi/N)kn} \sum_{n=0}^{N-1} x[(n - m) \mod N] e^{-j(2\pi/N)k(n-m)}
$$

$$
= H[k] \left[ \sum_{n=m}^{N-1} x[n-m] e^{-j(2\pi/N)(n-m)} + \sum_{n=m}^{N-1} x[n-m+N] e^{-j(2\pi/N)(n-m+N)} \right]
$$

$$
= H[k] \left[ \sum_{n=0}^{N-m-1} x[n] e^{-j(2\pi/N)kn} + \sum_{n=N-m}^{N-1} x[n] e^{-j(2\pi/N)kn} \right]
$$

$$
= H[k] \sum_{n=0}^{N-1} x[n] e^{-j(2\pi/N)kn}
$$

where $H[k]X[k]$ is the multiplication of the two sequences in the frequency domain.

### A.3 Property of conjugate symmetric

The conjugate symmetry property is one of the properties of the DFT. Given a signal $x[n]$, we have:

$$
X[N - k] = \sum_{n=0}^{N-1} x[n] e^{-j(2\pi/N)(N-k)n} = \sum_{n=0}^{N-1} x[n] e^{j(2\pi/N)kn} = X^*[k]
$$

(A.9)

In our GRLNet, we use this property to reduce learnable parameters and redundant computations.

### B Experiments

#### B.1 Evaluation criteria

As discussed in evaluation criteria, we use dynamic thresholds to calculate the performance of anomaly detection for the proposed GRLNet. Therefore, given a specific threshold, we can calculate the TP (True Positives), FP (False Positives), TN (True Negatives), and FN (False Negatives). Furthermore, we have:

$$
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
$$

(A.10)

$$
\text{Precision} = \frac{TP}{TP + FP}
$$

(A.11)

$$
\text{Recall} = \frac{TP}{TP + FN}
$$

(A.12)

$$
F1\text{- score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}}
$$

(A.13)

where the best F1-score can be obtained from the optimal global threshold. Under all possible thresholds, we can obtain a precision-recall curve (with precision as the $y$-axis and recall as the $x$-axis) and $AP = \sum_{n} ((R_n - P_{n-1})/P_n)$, where $P_n$ and $R_n$ are precision and recall at the $n$-th threshold. Similarly, we can also get $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$. The ROC curve with FPR as the $x$-axis and TPR as the $y$-axis. And the AUC score is the area under the ROC curve.

#### B.2 Details of GRLNet

The GRLNet consists of two generators $G^{\text{local}}$ and $G^{\text{regional}}$, a discriminator $D$ and a global filtering layer. $G^{\text{local}}$ and $G^{\text{regional}}$ are autoencoder structures composed of 1D convolutions and their decoder parts are both constructed by 1D transposed convolutions. The sizes of the 1D convolutional kernels of $G^{\text{local}}$ and $G^{\text{regional}}$ are $3 \times 3$ and $7 \times 7$, respectively. The discriminator $D$ is composed of a 1D convolution and a fully connected layer. A 1D batch normalization layer and ReLU activation function are added following each of convolutional layers of $G^{\text{local}}$, $G^{\text{regional}}$ and $D$. During all of these epochs, we train on Adam with learning rate set to $10^{-3}$ and $10^{-3}$ for generators ($G^{\text{local}}$ and $G^{\text{regional}}$).
Figure A1: Schematic view of the test rack at SAMSON AG (Figure provided by SAMSON AG).

MIMII. In our research, the dataset is split into a training set and a test set. The test set is consisted of all abnormal acoustic signals and the same number of normal acoustic signals. The other normal acoustic signals are regarded as the training set (see Table A4).

B.4 Ablation analysis

Analysis of different intensity noise. We compare the anomaly detection capability of GRLNet with different intensity noise in cavitation2017 (anomaly percentage of 50%), see Table A5.
Table A3: Details of the training and test sets. (·) denotes the number after the sliding window (window size is 466944).

| Dataset       | Train | Test |
|---------------|-------|------|
|               |       |      |
|               | Non-cavitation | Cavitation (%) |
|               | 10% | 20% | 30% | 40% | 50% |
| Cavitation2017 | 121 (1210) | 30 (300) | 3 (30) | 6 (60) | 12 (120) | 20 (200) | 30 (300) |
| Cavitation2018 | 158 (13114) | 40 (3320) | 4 (332) | 10 (830) | 17 (1411) | 26 (2158) | 40 (3320) |
| Cavitation2018-noise | 32 (2656) | 8 (664) | 1 (83) | 2 (166) | 3 (249) | 5 (415) | 8 (664) |

Table A4: Details of the training and test sets on MIMII.

| Machine type / model ID | Training set | Test set |
|--------------------------|--------------|----------|
| Valve                    | 00 872 238   | 02 588 240 |
|                          | 04 880 240   | 06 872 240 |
| Pump                     | 00 863 286   | 02 894 222 |
|                          | 04 602 200   | 06 934 204 |
| Fan                      | 00 604 814   | 02 657 718 |
|                          | 04 685 696   | 06 654 722 |
| Slide rail               | 00 712 712   | 02 801 534 |
|                          | 04 356 256   | 06 445 178 |

Table A5: AUC and \(F_1\) score results for different intensity noise.

| Results | Different intensity noise (dB) |
|---------|--------------------------------|
|         | -6   | -4   | -2   | 0    | 2    | 4    | 6    |
| AUC     | 0.9753 | 0.9747 | 0.9756 | 0.9750 | 0.9756 | 0.9657 | 0.9748 |
| \(F_1\) score | 0.9831 | 0.9699 | 0.9831 | 0.9633 | 0.9831 | 0.9616 | 0.9699 |