


## Article

# Unsupervised Anomaly Detection Method for Electrical Equipment Based on Audio Latent Representation and Parallel Attention Mechanism

Wei Zhou <sup>1</sup>, Shaoping Zhou <sup>1</sup>, Yikun Cao <sup>1</sup>, Junkang Yang <sup>2</sup>  and Hongqing Liu <sup>2,\*</sup><sup>1</sup> Power China Chengdu Engineering Corporation Limited, Chengdu 610072, China<sup>2</sup> School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; junkangyang@outlook.com

\* Correspondence: hongqingliu@cqupt.edu.cn

## Abstract

The stable operation of electrical equipment is critical for industrial safety, yet traditional anomaly detection methods often suffer from limitations, such as high resource demands, dependency on expert knowledge, and lack of real-world capabilities. To address these challenges, this article proposes an unsupervised anomaly detection method for electrical equipment, utilizing audio latent representation and a parallel attention mechanism. The framework employs an autoencoder to extract low-dimensional features from audio signals and introduces a phase-aware parallel attention block to dynamically weight these features for an improved anomaly sensitivity. With adversarial training and a dual-encoding mechanism, the proposed method demonstrates robust performance in complex scenarios. Using public datasets (MIMII and ToyADMOS) and our collected real-world wind turbine data, it achieves high AUC scores, surpassing the best baselines, which demonstrates our framework design is suitable for industrial applications.

**Keywords:** anomaly detection; mechanical fault diagnosis; unsupervised learning; audio signal analysis; electrical safety



Academic Editor: Douglas O'Shaughnessy

Received: 25 June 2025

Revised: 28 July 2025

Accepted: 28 July 2025

Published: 30 July 2025

**Citation:** Zhou, W.; Zhou, S.; Cao, Y.; Yang, J.; Liu, H. Unsupervised Anomaly Detection Method for Electrical Equipment Based on Audio Latent Representation and Parallel Attention Mechanism. *Appl. Sci.* **2025**, *15*, 8474. <https://doi.org/10.3390/app15158474>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the rapid developments of industry and smart manufacturing, the stable operation of electrical equipment is crucial for ensuring the safety and reliability of power supply. However, during long-term operation, electrical equipment may experience various abnormal conditions due to aging, overload, environmental factors, and other reasons such as partial discharge, mechanical loosening, and insulation damage. If these abnormalities are undetected and not addressed in a timely manner, they may lead to equipment failures or even serious safety incidents [1]. Therefore, developing efficient and accurate anomaly detection methods for electrical equipment is of significant practical importance.

Traditional anomaly detection methods for electrical equipment primarily rely on periodic inspections, online monitoring systems, and rule-based fault diagnosis techniques. However, these methods often have certain limitations. For instance, periodic inspections require substantial human labor and material resources and cannot achieve real-time monitoring [2]. On the other hand, online monitoring systems, while capable of providing continuous monitoring data, typically depend on specific sensors and involve complex data processing. The rule-based anomaly detection techniques [3] rely on expert experience, making it difficult to handle complex abnormal situations.

To address these issues, many researchers have also explored various types of signals for anomaly detection, including vibration signals, force signals, audio signals and others. Vibration signals [4,5] have been widely used for condition monitoring and fault diagnosis in rotating machinery, such as motors, turbines, and bearings. These signals provide rich information about the mechanical health of equipment, enabling the detection of imbalances, misalignments, and wear. By using this signal, techniques such as time-domain analysis, frequency-domain analysis, and time–frequency analysis have been employed to extract meaningful features from vibration data. Similarly, force signals, which measure the interaction forces between components, are particularly useful in detecting anomalies in mechanical systems subjected to dynamic loads [6]. For instance, force signals can reveal issues such as excessive friction, impact forces, or structural deformations. Using force signals, signal processing methods, including wavelet transforms [7] and empirical mode decomposition [8], have been applied to analyze these signals effectively.

In the field of anomaly detection for electrical equipment, audio signals, as an important monitoring tool, offer advantages such as non-invasiveness, real-time capabilities, and rich information. Electrical equipment produces different sound characteristics during normal operation and abnormal states. By analyzing these audio signals, the abnormal states of equipment can be effectively identified. However, audio signals are typically characterized by high dimensionality, nonlinearity, and time-varying properties [9], making it challenging to achieve effective anomaly detection using traditional signal processing methods. Therefore, how to extract effective feature representations from audio signals has become a key issue in this field.

Recently, deep learning technologies have made significant progress in audio signal processing. Specifically, representation learning methods based on autoencoders [10] can extract low-dimensional latent representations from high-dimensional data in an unsupervised manner, effectively reducing data complexity while retaining important feature information. Additionally, the attention mechanism [11], as an emerging deep learning technology, can automatically learn the important parts of data, enhancing the model's representation capability and generalization performance. Integrating autoencoders with attention mechanisms is expected to achieve better results in anomaly detection for electrical equipment.

In this work, we propose an unsupervised anomaly detection method based on audio latent representation and a parallel attention mechanism, aiming to improve the accuracy and robustness of anomaly detection for electrical equipment. Specifically, the method uses an autoencoder to extract features from the audio signals of electrical equipment, obtaining low-dimensional latent representations, with a parallel attention mechanism applied to weigh the latent representations, highlighting features related to anomalies. Finally, an anomaly scoring function is designed based on reconstruction error or feature distance to achieve anomaly detection for electrical equipment. Experimental validations show that the proposed method shows excellent performance in multiple anomaly detection tasks for electrical equipment, effectively identifying different types of abnormal conditions and demonstrating high practical value.

The main contributions of this article are as follows:

- An unsupervised anomaly detection method based on adversarial training is proposed, which can effectively extract key features from audio signals of electrical equipment;
- A novel attention block is proposed in this article, which improves the model's sensitivity to anomalous features through multi-channel attention weighting;
- In addition to testing on common public datasets, we also use real-world experiments to verify the effectiveness of proposed model.

The structure of this paper is organized as follows: Section 2 introduces the relevant research background and existing methods. Section 3 details the proposed unsupervised anomaly detection method. Section 4 presents the experimental design and result analysis. Section 5 summarizes the work and outlines future research directions.

## 2. Related Work

In recent years, audio analysis based mechanical anomaly detection become a research hotspot in the field of mechanical health diagnosis, primarily due to its high detection accuracy, strong generalization capability, non-intrusive measurement, and low cost.

Traditional methods mainly achieve the goal of detection by analyzing the statistical features of audio signals and identifying abnormal characteristics. These methods typically rely on signal processing techniques. For instance, time–frequency filters combined with inverse short-time Fourier transform (ISTFT) [12] can be used to detect bearing damages through dynamic resampling techniques and envelope spectrum analysis. Although these traditional methods perform well in many cases, they often require expert knowledge [3] and have weak robustness. In real-world working environments, various random factors such as noises can interfere with audio signals, leading to a lower accuracy in anomaly diagnosis.

Supervised learning-based methods require labeled data with known fault categories to guide the training of detection models, primarily used to determine whether the overall state of mechanical equipment is healthy and to identify the fault status of specific components, modules, or parts. Depending on the results of model, supervised learning-based methods can be further divided into binary classification and multi-class classification. Binary classification models learn audio features of labeled signal samples from healthy or abnormal equipment and combines with machine learning classification algorithms to assess the status of equipment. Some of the studies [13–15] extracted statistical features from audio signals, including mean, variance, entropy, and peak factor, and used them as inputs for a support vector machine (SVM) [16] classifier to identify the state of industrial equipment; this approach can achieve high accuracy on specific data. Currently, deep learning methods have gained significant attention due to their unique advantages. For instance, the methods proposed in [17,18] extract the mel frequency cepstrum coefficient (MFCC) from audio clips and train a deep convolution neural network (CNN) model, reaching an accuracy over 90% in on-site experiments. On the other hand, multi-class classification model is typically used to identify various abnormal categories and different degrees of faults in the overall equipment or its individual components. For example, a biological population anomaly recognition system [19] based on the Internet of Things (IoT) and MFCC was proposed, in which the K-nearest neighbor (KNN) algorithm is used to perform multi-classification of biological sounds. It can accurately identify whether the organisms are infected by viruses or contaminated by pesticides, with a classification accuracy of 98.8%. Deep learning methods have also excelled in multi-class anomaly detection, such as the end-to-end CNN model constructed by Peng et al. [20,21], which accepts SPWVD-MFCC [20] as inputs and significantly improves classification accuracy in low signal-to-noise ratio (SNR) environments.

Unsupervised learning-based methods do not require labeled data for model training and typically quantify the distance between input signals and normal audio signals, using thresholds to determine the “normal or abnormal” state of targets, making them suitable for practical application scenarios because collecting data can be difficult sometimes. A common approach involves the use of autoencoders and their variants, such as variational autoencoders (VAE) and sparse autoencoders, which learn feature representations by reconstructing normal audio signals and identify anomalies during the testing phase by calculating reconstruction errors [22]. The underlying assumption is that normal samples

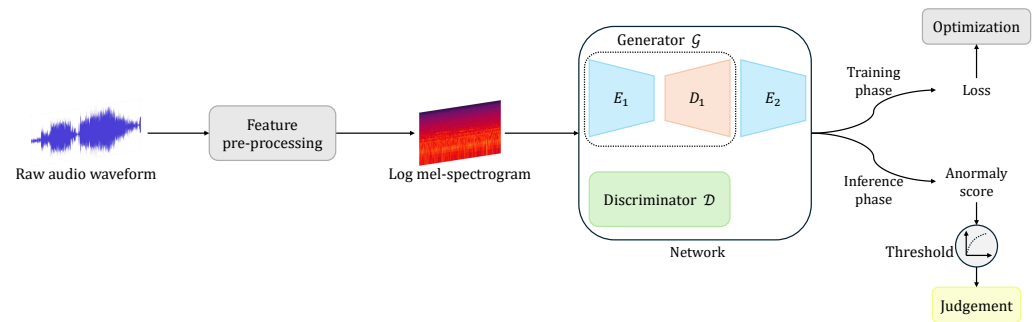
can be well-reconstructed by the model, while anomalous samples, due to their different distribution, result in significantly higher reconstruction errors. Additionally, generative adversarial networks (GANs) have been widely applied in anomaly detection tasks [23,24], where the generator learns the distribution of normal data, and the discriminator is used to distinguish between normal and anomalous samples. By training the generator to produce samples similar to normal data, the model can detect anomalies during testing based on the discriminator's output or the generator's reconstruction error. In recent years, methods based on contrastive learning have also gained considerable attention, learning robust feature representations by maximizing the similarity between normal samples and minimizing the similarity between anomalous samples [25,26]. This approach leverages data augmentation techniques to generate positive sample pairs and optimizes the model through a contrastive loss function, enabling the learning of more discriminative features in an unsupervised manner. Furthermore, time-series modeling methods, such as long short-term memory transformers [11] and selective state space model [27], have been widely used to capture temporal dependencies in audio signals, thereby improving the accuracy of anomaly detection [28,29]. These models effectively handle long-term dependencies in audio data and detect anomalies by predicting future frames or reconstructing input sequences. In practical applications, data augmentation techniques and multimodal fusion methods have also been introduced to enhance the generalization capability of models. For example, combining spectral features (such as mel spectrograms) and visual features of targets can more comprehensively capture anomalous patterns [30]. Additionally, some studies have explored clustering-based methods, where normal samples are clustered into the same category, and samples far from the cluster centers are labeled as anomalies [31]. Despite the impressive performance of these methods on various public datasets, several challenges remain, such as robustness to complex environmental noise, efficiency in on-site experiments, and generalization to unknown anomaly types. Moreover, with the proliferation of edge computing and IoT devices, achieving efficient anomaly detection on resource-constrained devices is also an important research direction.

To summarize, traditional methods, supervised learning-based methods, and unsupervised learning-based methods each have their own advantages and disadvantages. Traditional methods rely on expert knowledge and exhibit weak robustness but offer unique advantages in certain scenarios. Supervised learning-based methods require labeled data, provide high diagnostic accuracy, and are suitable for diagnosing known fault types. Unsupervised learning-based methods do not require labeled data and are suitable for practical application scenarios but still need improvement in detecting specific anomaly types.

### 3. Method

As illustrated in Figure 1, the pipeline of proposed anomaly detection system in this article is as follows: The system first performs feature pre-processing on the original audio waveform, converting it into a logarithmic mel spectrogram, and then processes it through a network consisting of a generator containing an encoder  $E_1$  and a decoder  $D_1$ , a discriminator, and an extra encoder  $E_2$ . During the training phase, the network learns audio representations by optimizing the loss function. During the inference phase, the system calculates the anomaly score based on the output of network and makes an abnormal judgment on the audio by using a threshold.

In this section, we will describe the details of these above-mentioned parts of our system.



**Figure 1.** The framework of the proposed anomaly detection system.

### 3.1. Pre-Processing

The pre-processing of the proposed system is mainly carried out to extract log-mel spectrum features from the input audio waveform through signal processing method, which involves time–frequency transformation and auditory feature simulation. First, the raw audio signal  $x(n)$  is divided into short frames (20 ms duration) with an overlapping of 50%, and each frame is multiplied by a Hamming window to reduce spectral leakage; this window function is

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right). \quad (1)$$

After that, short-time Fourier transform (STFT) is applied to all frames to obtain the frequency representation of the  $x(n)$ , given by

$$X(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{-j2\pi kn/N}, \quad (2)$$

followed by calculating the power spectrum

$$P(k) = |X(k)|^2. \quad (3)$$

In addition, a mel filterbank is designed to mimic human auditory perception, where linear frequency is converted to mel scale via a function, as follows:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \quad (4)$$

where  $m$  and  $f$  are the frequency in linear and mel scales, respectively. A group of triangular filters are uniformly spaced on the mel scale. Each filter's response  $H_m(k)$  forms a triangular shape in the frequency domain. The power spectrum is projected onto these filters to compute the energy in each mel band. That is,

$$E(m) = \sum_k P(k)H_m(k). \quad (5)$$

Finally, the logarithmic operation  $\log(E(m) + \epsilon)$  (where  $\epsilon$  is a small constant to prevent numerical instability) is applied to compress the dynamic range and align with human perception of sound intensity. The time–frequency matrix is the log-mel spectrogram, which provides a perceptually relevant representation of audio signals.

### 3.2. Parallel Attention

#### 3.2.1. General Attention

The main responsibility of the attention mechanism is to calculate the importance weights of each element in the input sequence, thereby performing a weighted summation of the input information to highlight key parts and suppress irrelevant parts, given an input sequence  $X = [x_1, x_2, \dots, x_n]$ , where each  $x_i$  is a vector. The goal of attention mechanism is to generate a context vector  $c$  that captures the key information in the input sequence. First, we need to compute the attention score  $a_i$  for each input element  $x_i$ , which is typically calculated using a function  $f$ , such as dot product, additive models, or scaled dot product. Taking the scaled dot product as an example, the formula is  $a_i = \frac{\exp(q \cdot k_i / \sqrt{d_k})}{\sum_{j=1}^n \exp(q \cdot k_j / \sqrt{d_k})}$ , where  $q$  is the query vector,  $k_i$  is the key vector, and  $d_k$  is the dimensionality of the key vector. This score represents the similarity between the current query and each key, and after softmax normalization, it yields the attention weight  $a_i$ .

Next, these weights are used to perform a weighted summation of the value vectors  $v_i$  in the input sequence to generate the context vector  $c$ , i.e.,  $c = \sum_{i=1}^n a_i v_i$ . This process can be understood as filtering the input information, where elements with larger weights contribute more to the final result. In practical applications, the query vector  $q$ , key vector  $k_i$ , and value vector  $v_i$  are typically obtained by applying linear transformations to the input sequence, such as  $q = W_q x$ ,  $k_i = W_k x_i$ , and  $v_i = W_v x_i$ , where  $W_q$ ,  $W_k$ , and  $W_v$  are learnable parameter matrices. In this way, the attention mechanism can dynamically adjust the weights of each input element, thereby demonstrating greater flexibility and adaptability when processing sequential data.

It is worth noting that the attention mechanism can be applied not only to integrate information within a single sequence (i.e., self-attention mechanism) but also to handle relationships between two sequences (i.e., cross-attention mechanism). In the self-attention mechanism, the query, key, and value all come from the same sequence; in the cross-attention mechanism, the query comes from one sequence, and the key and value come from another sequence. This mechanism has been widely used in tasks such as machine translation and text generation, significantly improving model performance by capturing dependencies between sequences.

#### 3.2.2. Proposed Parallel Attention

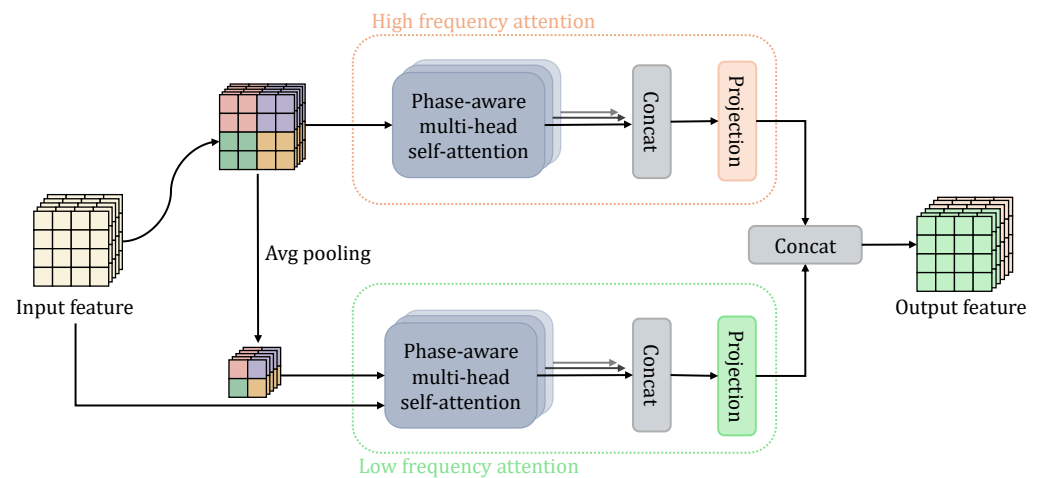
The working environment of electrical equipment usually contains various noises, such as mechanical noise, electromagnetic interference, etc. These noises may mask the abnormal audio signals of the equipment. Generally, the audio features of electrical equipment usually contain high-frequency information, such as partial discharge, and low-frequency information, such as mechanical vibration. Previous work lacks effective use of the phase features of audio signals and does not specifically consider different frequency components. Therefore, these models often show poor generalization and robustness in anomaly detection tasks, and they cannot achieve the best results in real-world application scenarios.

In addition, most previous studies have adopted a cascade network structure, which has large computational overhead and model parameters. The anomaly detection task usually has high real-time requirements. Such models are not conducive to use in resource-constrained embedded devices or real-time systems.

To address these issues, we proposed a new attention block in this section; its structure is depicted in Figure 2. The proposed parallel attention module primarily consists of two parallel pathways, each dedicated to processing high-frequency and low-frequency information, respectively. The input features are initially divided into two parts, with one part directly entering the high-frequency attention pathway and the other undergoing an



average pooling operation to reduce spatial resolution, thereby preserving low-frequency information and entering the low-frequency attention pathway.



**Figure 2.** The architecture of proposed parallel attention block.

In high-frequency attention, the input features are processed through a phase-aware multi-head self-attention mechanism. This phase-aware multi-head self-attention mechanism is similar to the standard multi-head self-attention mechanism but adjusts the attention weights based on the phase information of the input features. The processing in the low-frequency attention pathway is similar to that in the high-frequency attention pathway, with the difference being that the input features have reduced dimensions after average pooling, resulting in lower computational overhead. Similarly, low-frequency information is processed through the phase-aware multi-head self-attention mechanism to obtain the output of the low-frequency attention pathway. Finally, the outputs of the high-frequency and low-frequency attention pathways are concatenated to form a feature tensor. This concatenated feature is then further fused and transformed through a fully connected layer or convolutional layer to produce the final output feature.

The design of this parallel attention module effectively captures information at different frequencies and enhances sensitivity to phase information through the phase-aware mechanism, thereby improving the model's performance. Specifically, the advantage of the parallel attention module lies in its ability to simultaneously process information at different frequencies, capturing details and high-frequency features through the high-frequency pathway and capturing overall structure and semantic information through the low-frequency pathway. This parallel processing approach not only increases the efficiency of the model but also enhances its ability to express features at different frequencies. Moreover, the phase-aware mechanism allows the model to better utilize the phase information of the input features, which is crucial for audio classification tasks, as phase information often contains rich semantic and structural information.

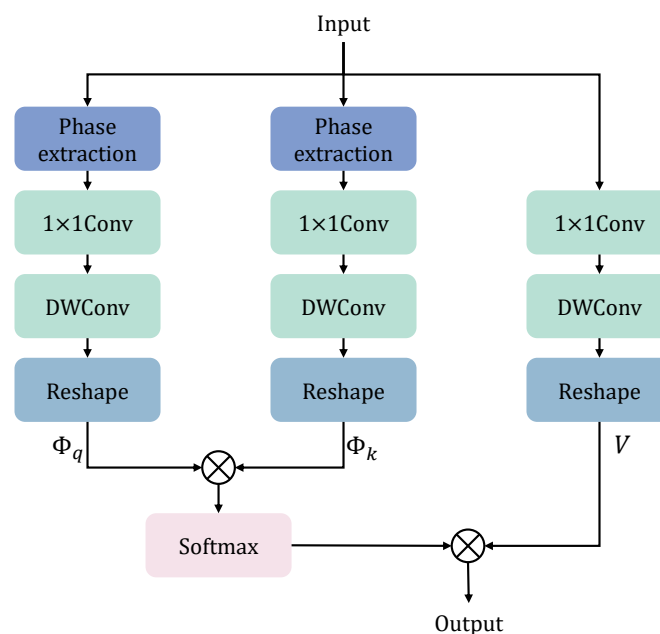
The proposed phase-aware multi-head self-attention module in Figure 2 is an innovative attention mechanism designed to enhance the model's ability to understand and process input features by incorporating phase information, inspired by [32]. The structure of this module is depicted in Figure 3.

First, the input features are fed into two parallel branches, each containing a phase extraction module to extract the phase information of the input features. Given an input  $X_{in}$ , this operation is

$$M_X \angle \phi_X = FFT2d(X_{in}), \quad (6)$$

$$X_{out} = IFFT2d(|M| \angle \phi_X), \quad (7)$$

where  $M_X$  and  $\angle\phi_X$  are amplitude and phase information of input,  $M$  is a stable value, and we set  $M = 1$  in this article. Through phase extraction, the model can better capture the phase characteristics in the input features, thereby increasing sensitivity to different frequency information. Next, the phase information in each branch undergoes a  $1 \times 1$  convolution operation to linearly transform and reduce the dimensionality of the phase features. The role of the  $1 \times 1$  convolution here is mainly to adjust the number of feature channels to match the subsequent depth-wise convolution (DWConv), which decomposes standard convolution into depthwise convolution and pointwise convolution, reducing the computational load and the number of parameters while maintaining the model's expressive power. Then, the features, after depthwise convolution, are reshaped into an appropriate form for subsequent attention calculations. The reshape operation here converts the two-dimensional feature map into a one-dimensional feature vector to match the query (Q), key (K), and value (V) vectors in the attention mechanism. After obtaining Q and K, they are fed into a dot product operation to calculate the similarity between them. The result of the dot product is passed through a softmax function for normalization to obtain the attention weight matrix. The attention weight matrix represents the correlation between different positions, where positions with higher weights have a greater influence on the current position. Finally, the attention weight matrix is dot-multiplied with the value vector V to obtain the final output features.



**Figure 3.** The architecture of proposed phase-aware multi-head self-attention.

Specifically, for the input feature  $X$ , the phase extraction module extracts the phase information  $\Phi_q$  and  $\Phi_k$ . Then,  $\Phi_q$  and  $\Phi_k$  are processed through  $1 \times 1$  convolution and depthwise convolution to obtain Q and K. Meanwhile, the input feature  $X$  is also processed to obtain V. The dot product result of Q and K is normalized by softmax and then multiplied with V to obtain the final output feature. The formulas are

$$Q = \text{DWConv}(1 \times 1\text{Conv}(\Phi_q)), \quad (8)$$

$$K = \text{DWConv}(1 \times 1\text{Conv}(\Phi_k)), \quad (9)$$

$$V = \text{DWConv}(1 \times 1\text{Conv}(X)), \quad (10)$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \quad (11)$$



$$\text{Output} = AV, \quad (12)$$

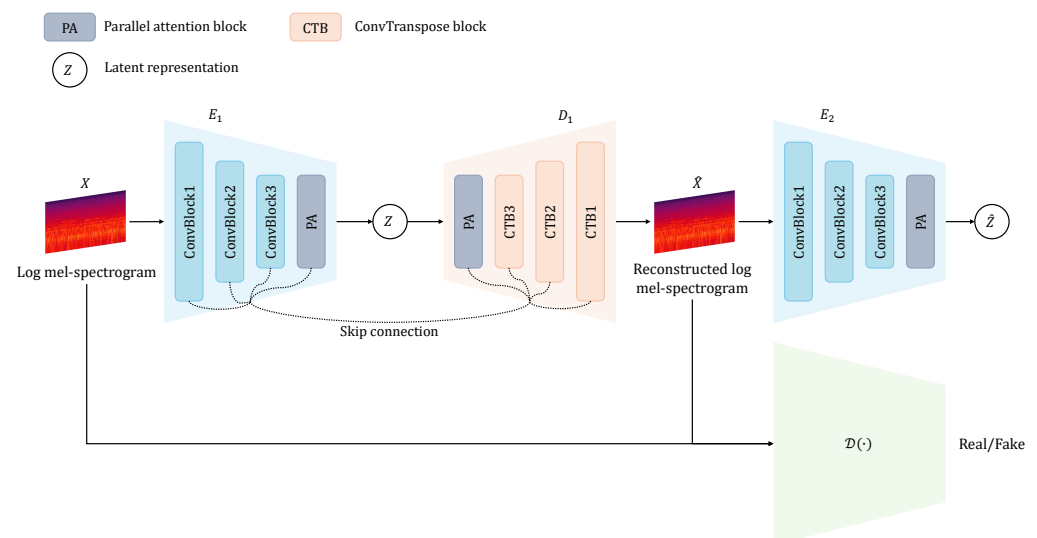
where  $d_k$  is the dimension of Q and K, used to scale the dot product result to prevent numerical overflow and gradient vanishing. In this way, the phase-aware multi-head self-attention module can effectively capture the phase information in the input features and incorporate it into the attention calculation, thereby enhancing the model's sensitivity and processing capability for different frequency information.

Additionally, this module also features a multi-head mechanism, where the input features are divided into multiple heads, and each head independently performs the aforementioned phase-aware attention calculation. The outputs of all heads are then concatenated and projected through a fully connected layer to obtain the final output features. The multi-head mechanism enhances the model's expressive power and its ability to model different feature subspaces.

In summary, the phase-aware multi-head self-attention module significantly improves the model's ability to understand and process input features by introducing phase information and a multi-head mechanism, making it suitable for various tasks that require capturing complex feature relationships.

### 3.3. Network Architecture

In this section, a generative adversarial network is designed to process log-mel spectrograms, primarily composed of a generator, a discriminator and an extra encoder. As Figure 4 shows, the generator's structure can be further divided into an encoder  $E_1$  and a decoder  $D_1$ , whose objective is to map the input log-mel spectrogram  $X$  to a latent representation  $Z$  and then decode it back into a reconstructed log-mel spectrogram  $\hat{X}$ . The discriminator  $\mathcal{D}(\cdot)$  is used to distinguish between real and fake log-mel spectrograms.



**Figure 4.** The architecture of network.

The encoder  $E_1$  of the generator consists of three convolution blocks (ConvBlocks) and a parallel attention block (PA). The input log-mel spectrogram  $X$  first undergoes progressive feature extraction and downsampling via ConvBlock1, ConvBlock2, and ConvBlock3. Each convolutional block includes a convolution layer, a batch normalization layer, and a LeakyReLU activation layer to learn the spatial features of the data. Subsequently, the data enters the parallel attention block (PA), which enhances the model's focus on important features through an attention mechanism, improving the robustness and distinctiveness of the feature representation. After processing by the encoder, the features are compressed into a latent representation  $Z$ , whose dimensionality is usually much lower than that of

the original input data but retains its key feature information. The decoder  $D_1$  conducts an inverse structure compared to the encoder, consisting of a parallel attention block (PA) and three transposed convolutional blocks (CTB3, CTB2, CTB1). The latent representation  $Z$  first passes through the parallel attention block (PA) to further enhance feature expressiveness. The data then sequentially undergoes CTB3, CTB2, and CTB1, where these transposed convolutional blocks gradually restore the spatial resolution of the data through upsampling while learning feature details via convolutional layers. Finally, the decoder outputs the reconstructed log-mel spectrogram  $\hat{X}$ , whose dimensions and structure should be the same as input  $X$ . Additionally, there is an extra encoder  $E_2$  in the network, which encodes the reconstructed log-mel spectrogram  $\hat{X}$  to obtain another estimated latent representation  $\hat{Z}$ . The structure of this encoder  $E_2$  is similar to that of  $E_1$ , also consisting of multiple convolutional blocks and a parallel attention block. The discriminator  $\mathcal{D}(\cdot)$  has a structure similar to the DCGAN's [24]. Its objective is to classify the  $X$  and the  $\hat{X}$  as real or fake, respectively.

Formally, the encoder  $E_1$  of the generator can be expressed by

$$Z = E_1(X), \quad (13)$$

where  $X$  is the input log-mel spectrogram and  $Z$  is the latent representation.

The decoder  $D_1$  process is

$$\hat{X} = D_1(Z), \quad (14)$$

where  $\hat{X}$  is the reconstructed log-mel spectrogram.

Similarly, encoder  $E_2$  encodes the reconstructed log-mel spectrogram  $\hat{X}$  to produce the following:

$$\hat{Z} = E_2(\hat{X}), \quad (15)$$

where  $\hat{Z}$  is the latent representation of the reconstructed data. The output of the discriminator  $\mathcal{D}(\cdot)$  can be represented as 1 or 0. Table 1 illustrates the details of our network, where we only list the encoder due to the limited space, as the decoder's configuration is the inverse version of the encoder.

**Table 1.** Details of network configuration, where “B” indicates batch size.

Module	Layer	Input Size	Output Size
ConvBlock1	Conv2d (1, 128, 5, 2, 2)	[B, 1, 128, 312]	[B, 128, 64, 156]
	LeakyReLU		
	BatchNorm2d (128)		
ConvBlock2	Conv2d (128, 256, 5, 2, 2)	[B, 128, 64, 156]	[B, 256, 32, 78]
	LeakyReLU		
	BatchNorm2d (256)		
ConvBlock3	Conv2d (256, 256, 5, 2, 2)	[B, 256, 32, 78]	[B, 256, 16, 39]
	LeakyReLU		
	BatchNorm2d (128)		
PA	-	[B, 256, 16, 39]	[B, 256, 16, 39]

The overall training objective of the network is to optimize the loss functions of the generator and the discriminator. The generator aims to minimize the reconstruction error and the discriminator's misjudgment probability for generated data, while the discriminator aims to maximize the correct classification probability for real and fake data. Additionally, encoder  $E_2$ 's aim is to generate the correct latent representation from reconstructed log-mel spectrogram. By training the whole network on the normal audio data, it can learn to

generate the correct features of normal data; however, when it comes to the abnormal one, the network's prediction of  $Z$  and  $\hat{Z}$  will be biased, which can be used to compute an anomaly score to judge the state of equipment. The details will be presented in Section 3.4.

### 3.4. Loss Function and Judgment Method

The proposed model is able to effectively capture the distribution of normal data through three jointly optimized loss functions. To that end, the first part is adversarial loss  $L_{adv}$ , which employs feature matching to ensure that the distribution of reconstructed spectrogram closely matches the original one by computing the  $L_2$  distance between them in the discriminator's latent space. Its definition is

$$L_{adv} = \mathbb{E}_{X \sim p} \|\mathcal{D}(X) - \mathcal{D}(\hat{X})\|_2. \quad (16)$$

Second, the spectrogram loss  $L_s$  computes  $L_1$  distance between the reconstructed and original log-mel spectrograms, given by

$$L_s = \mathbb{E}_{x \sim p} \|X - \hat{X}\|_1. \quad (17)$$

Lastly, the latent representation loss  $L_l$  ensures latent space consistency through a dual-encoding mechanism in normal data, which can be expressed as

$$L_l = \mathbb{E}_{x \sim p_X} \|G_E(x) - E(G(x))\|_2. \quad (18)$$

The total generator loss is a weighted sum of these three components, given by

$$L_G = w_1 L_{adv} + w_2 L_s + w_3 L_l. \quad (19)$$

In our experiments, the weights are set as  $w_1 = 1$ ,  $w_2 = 50$ , and  $w_3 = 1$ , following the previous work [10].

During the inference phase, the model computes an anomaly score to detect abnormal samples via forward propagation. The anomaly score is defined as the  $L_1$  distance between two latent representations:

$$\mathcal{A}(X) = \|Z - \hat{Z}\|_1. \quad (20)$$

The core to optimize the model to detect anomalous samples despite being trained only on normal samples lies in its explicit learning for the distribution of normal data through adversarial training and a dual-encoding mechanism. The generator is optimized to reconstruct the normal samples with a low error while maintaining consistency in the latent space. In contrast, anomalous samples, which deviate from the learned distribution, result in significant distortions in the reconstructed feature map and show notable encoding discrepancies in the latent space. By monitoring the reconstruction error in the latent space, the model detects anomalies when this metric simultaneously exceeds a predefined threshold, which essentially leverages the "misfit" of the model to unknown distributions as an anomaly signal, analogous to how humans identify irregularities based on experience.

To evaluate the model's overall anomaly detection capability on the whole test set, we follow the methodology established in prior research [10]. For each test sample  $X_i$  in the dataset  $\mathcal{P}$ , an anomaly score  $s_i = \mathcal{A}(X_i)$  is computed, generating a score set  $S = \{s_i : \mathcal{A}(X_i) | X_i \in \mathcal{P}\}$ . These scores are then normalized to a  $[0, 1]$  interval using min-max scaling:

$$s'_i = \frac{\mathcal{A}(X_i) - \min(S)}{\max(S) - \min(S)}. \quad (21)$$

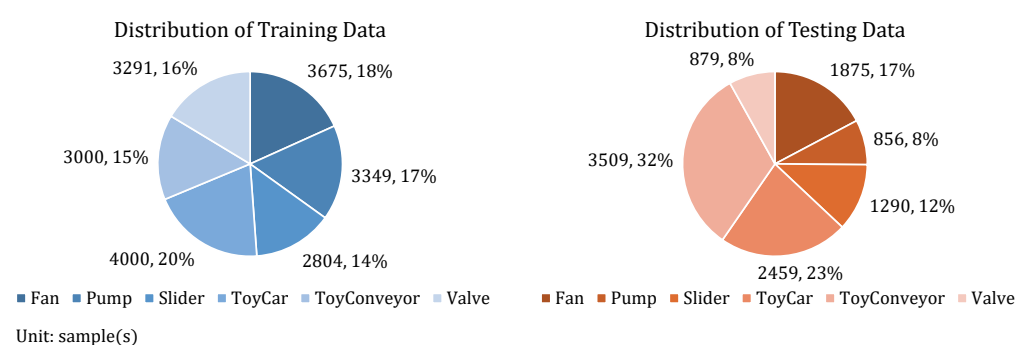
This rescaling ensures probabilistic interpretability. A predefined threshold  $t$  is subsequently applied to classify samples as anomalous or normal.

## 4. Experiments and Results

### 4.1. Data

We first train and evaluate the model on two public datasets. The MIMII dataset [33] focuses on abnormal sound detection of industrial machines in real factory environments, including four types of equipment, valves, pumps, fans, and sliders, each with seven different models of machines. The dataset contains 26,092 normal state sounds and 6065 abnormal state sounds, with a sampling rate of 16 kHz, recorded using an eight-channel circular microphone array, and the microphone is 50 cm away from the machine (10 cm for valves). Abnormal types include actual faults such as valve contamination, water pump leakage, fan imbalance, and rail damage. The dataset is specially designed with noise mixed versions with different SNRs (6 dB, 0 dB, −6 dB), and the reality is enhanced by superimposing real factory background noise. The data are stored in segments that are 10 s long. The second one, ToyADMOS dataset [34], is designed for machine operation sound anomaly detection (ADMOS), and it includes three sub-datasets, toy cars, toy conveyor belts, and toy trains, corresponding to product quality inspection, fixed machinery fault diagnosis, and mobile machinery fault diagnosis tasks, respectively. The dataset collects abnormal sounds by artificially damaging micro-machine parts. Each sub-dataset contains more than 180 h of normal operation sounds and more than 4000 abnormal samples, with a sampling rate of 48 kHz and recorded synchronously with four microphones. Specifically, the toy car subset contains four different model combinations (motor + bearing), each model records 1350 11 s normal samples (66 h in total) and 250 abnormal samples, and the abnormal types include 53 fault modes such as shaft bending and gear deformation; the toy conveyor belt subset covers three models, each model has 1800 10 s normal samples (60 h) and 355 abnormal samples, involving 60 faults such as belt tension abnormality; and the toy train subset contains four models, each model has 1350 11 s normal samples (66 h) and 270 abnormal samples, covering 54 faults such as track damage. The pumps, fans, valves, and sliders in the dataset have distinct acoustic characteristics due to their designs. To mitigate the risk of conflating sources of variation, we trained dedicated models for each equipment type. This prevents cross-equipment interference and ensures that latent representations capture device-specific anomalies.

In our experiments, all audio samples are resampled to 16 kHz and normalized to 10 s, and we only use the normal audio samples to ensure the unsupervised training setup. In addition, we select six types of equipment in these datasets in our experiment and split them into training data and testing data based on a previous study [35], whose specific data distributions are shown in Figure 5.



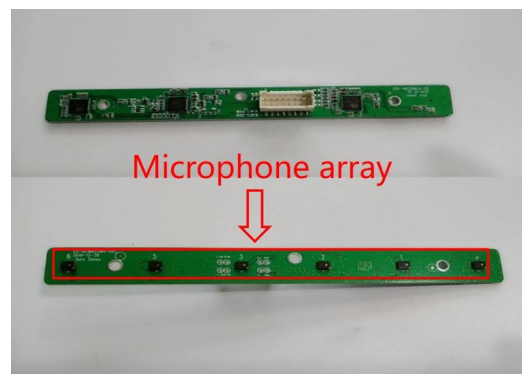
**Figure 5.** The distributions of training and testing data.

Furthermore, this study also used the audio recorded from the wind turbine equipment in real world to test and compare the proposed model with the baseline models. This part of the data was collected in our previous work [36] using three recording devices, each of which is described below.

The SS-ALIM6S2M1-002 voice algorithm motherboard is a six-microphone array motherboard product developed by SureSense Technology, Chongqing, China. The product is connected to AC108 externally, and the output format is a six-channel PCM signal with a sampling rate of 16 kHz and a depth of 16 bits. The device is shown in Figure 6.

The Zoom H1n recording device (Zoom, San Jose, CA, USA ) uses a stereo X/Y microphone configuration, supports 24 bit/96 kHz audio signals, has a low-cut filter, can reduce unwanted low-frequency noise such as popping sound and wind noise, improves the recording quality, and can use a compressor to obtain a maximum sound pressure of up to 120 dB SPL to ensure that the recorded audio is not distorted.

The SOGO AI recorder E1 (Beijing Sogou Network Technology Co., Ltd., Beijing, China) has a total of eight microphones, six of which are omnidirectional microphones and the remaining two are Harman 10 mm directional microphones. In addition, while improving the recording quality, the device also uses the clairVoice eight-microphone array algorithm and pureVoiceAI noise reduction algorithm for noise reduction.

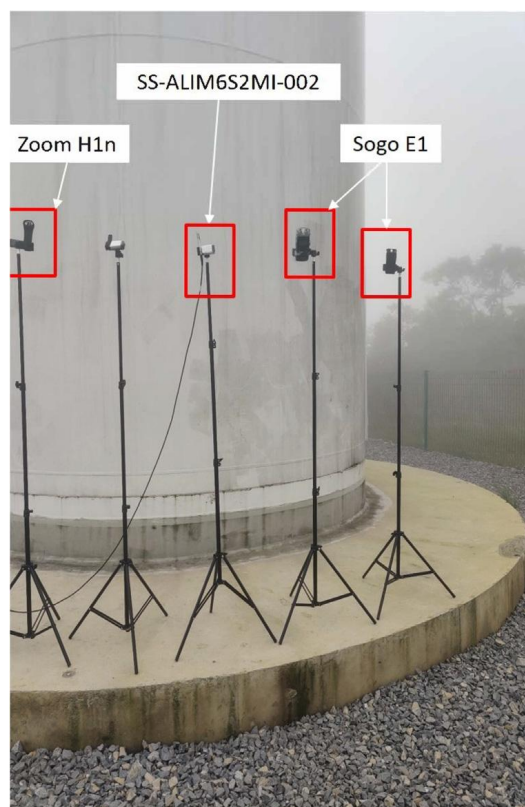


**Figure 6.** SS-ALIM6S2M1-002 voice algorithm motherboard [36].

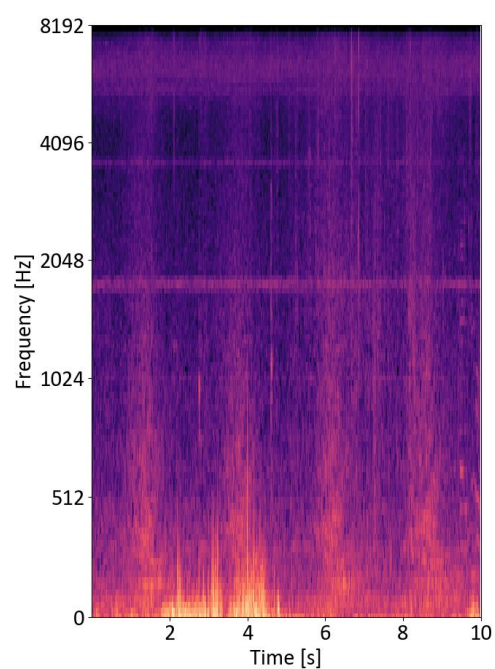
This part of the dataset is collected from the acoustic signals of a wind turbine at a height of 140 m and a blade diameter of 56 m under the conditions of 3–6 m/s wind speed and 976 hPa air pressure. The acquisition equipment recorded a total of 60 h of audio at a height of 2 m above the ground (Figure 7), with normal and fault conditions accounting for 50% each. Through 10 s segment processing and manual verification, a well-labeled acoustic sample library was finally formed. The log-mel spectrograms of a sample in these data are presented in Figure 8. As mentioned before, we only use normal data to train the model, and the final evaluation is based on 10% of the normal and abnormal data, respectively.

#### 4.2. Implementation Settings

In the feature extraction stage, we use 128 mel filters to convert the raw audio data into a log-mel spectrogram, with a window size of 1024 and a hop size of 512. After extraction, the feature tensor enters the network encoder and decoder, whose parameters are shown in Table 1. The implementation software is PyTorch 2.7.1. We use the Adam [37] optimizer to train the model for 300 epochs, with a batch size of 16 and a learning rate of 0.002. The entire training process is completed on an NVIDIA RTX 3090 GPU (Nvidia, Santa Clara, CA, USA). With this batch and device settings, the total VRAM cost is about 4000 MB and total training time is 10 h.



**Figure 7.** Data recording scene [36].



**Figure 8.** Log-mel spectrogram of the collected data.

#### 4.3. Evaluation Metrics

To evaluate the performance of proposed model, two evaluation metrics were employed: the area under the curve (AUC) of the receiver operating characteristic (ROC) and the partial-AUC (pAUC). In audio anomaly detection, AUC is a key metric for evaluating the overall performance of a model. It measures the model's ability to distinguish between normal and anomalous samples at different thresholds by calculating the area under the



ROC curve. Mathematically, with the true positive rate (TPR) on the y-axis and false positive rate (FPR) on the x-axis, AUC is defined as

$$\text{AUC} = \frac{\sum_{i=1}^{N_-} \sum_{j=1}^{N_+} \mathcal{H}(\mathcal{A}(x_j^+) - \mathcal{A}(x_i^-))}{N_- N_+}, \quad (22)$$

where  $x_j^+$  and  $x_i^-$  are normal and anomalous test samples, respectively,  $N_-$  and  $N_+$  are the numbers of normal and anomalous test samples,  $\mathcal{A}(\cdot)$  is an anomaly score function, and  $\mathcal{H}(x) = 1$  when  $x > 0$  and 0 otherwise. AUC covers the full range from 0 to 1. A higher AUC (closer to 1) indicates better model performance.

On the other hand, pAUC (partial AUC) focuses on a more practical low-FPR region (e.g.,  $0 \leq \text{FPR} \leq \alpha$ ), with the formula being

$$\text{pAUC} = \frac{\sum_{i=1}^{\lfloor \alpha N_- \rfloor} \sum_{j=1}^{N_+} \mathcal{H}(\mathcal{A}(x_j^+) - \mathcal{A}(x_i^-))}{\lfloor \alpha N_- \rfloor N_+}, \quad (23)$$

where  $\alpha$  is set to 0.1 in this study. By restricting the evaluation range, pAUC more accurately reflects performance in critical operational scenarios, especially in anomaly detection where class imbalance is severe. While optimizing AUC alone may lead to overfitting in high-FPR regions, pAUC ensures improved detection capability at low false positive rates.

#### 4.4. Baselines

In order to fairly evaluate the performance of the proposed method, we selected the following models as baselines, which are all unsupervised methods.

**AE [35]:** The baseline system in DCASE 2020 workshop [35], which is based on an autoencoder.

**EGBAD [38]:** An efficient anomaly detection model based on GAN. This model avoids the problem of expensive optimization steps required for traditional GAN during testing by training the generator, discriminator and encoder simultaneously, thereby significantly improving detection efficiency and achieving better performance.

**IDNN [39]:** An abnormal sound detection method based on an interpolation deep neural network. Different from an autoencoder, it removes the continuous spectrogram of the center frame through input features and predicts the interpolation result of the missing frame as output, thus avoiding the problem of the unstable prediction of edge frames.

**ANP [40]:** An unsupervised abnormal sound detection model based on the attentive neural process (ANP), which reconstructs the latent representation by dynamically selecting context and target areas of the spectrogram without predefined masks. The model uses attention mechanism and iterative optimization strategy to focus on the complex areas as the basis for abnormal scoring.

**PAE [41]:** A transformer-based unsupervised abnormal sound detection model. PAE captures inter-frame context information through the self-attention mechanism and combines it with the mask prediction strategy to improve the modeling ability of long sequences. It outperforms traditional methods on the DCASE2020 Task 2 dataset.

It is worth mentioning that all models keep the consistency of training data and test data when comparing, and the settings during training are the same with those mentioned in their papers.

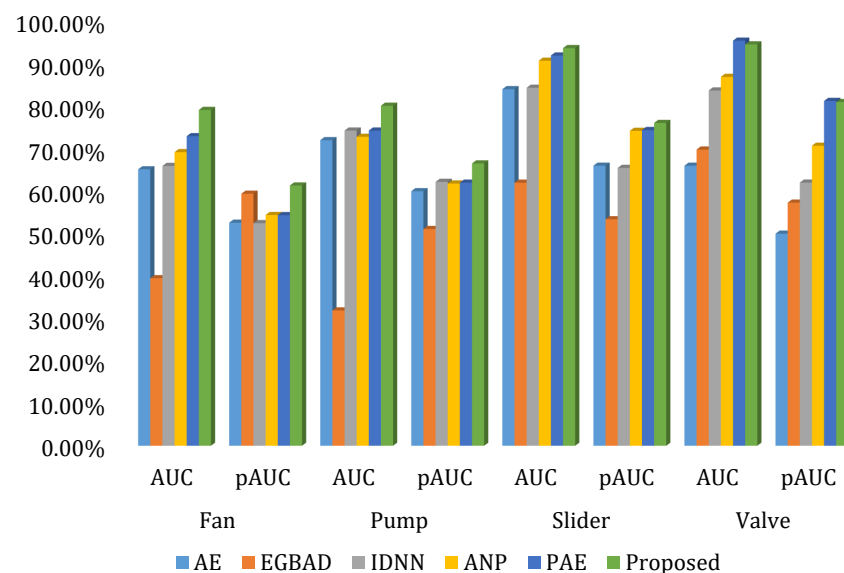
#### 4.5. Results

The evaluation results of our method and baseline models on MIMII are presented in Table 2 and Figure 9. As can be seen from the table and bar chart, there are significant differences in the performance of various methods on MIMII test data. Our method achieved

the highest AUC and pAUC values across all devices except for valve, demonstrating its superior performance in anomaly detection tasks. For example, on the pump device, our method attained an AUC of 80.11% and a pAUC of 66.53%, significantly outperforming other methods by over 4%. In contrast, the EGBAD method performed the worst, particularly on the Fan device, where its AUC was only 39.50% and pAUC was 59.40%, indicating substantial limitations in handling complex anomaly detection tasks. Other methods, such as AE, IDNN, ANP, and PAE, showed intermediate performance but performed notably well on certain devices. For instance, PAE achieved an AUC of 95.41% on the valve device, which is higher than that of our method. Additionally, the AUC values for the slider device were generally higher, suggesting that its anomaly detection task is relatively easier, while the pAUC values for the valve device exhibited greater fluctuations, indicating that detecting anomalies in certain regions of this device is more challenging. In summary, the proposed method demonstrates significant superiority in both AUC and pAUC metrics, proving its effectiveness in industrial equipment anomaly detection. In comparison, other methods' performance is suboptimal and may require further optimization to adapt to complex anomaly detection tasks.

**Table 2.** Evaluation results on MIMII test data, where the bold font indicates the best performance.

Method	Fan		Pump		Slider		Valve	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
DCASE [35]	65.15%	52.59%	72.00%	60.00%	84.00%	66.00%	66.00%	50.00%
EGBAD [38]	39.50%	59.40%	31.90%	51.10%	62.00%	53.40%	69.80%	57.30%
IDNN [39]	65.94%	52.48%	74.26%	62.20%	84.34%	65.48%	83.70%	62.02%
ANP [40]	69.20%	54.40%	72.80%	61.80%	90.70%	74.20%	86.90%	70.70%
PAE [41]	72.94%	54.37%	74.27%	62.01%	91.92%	74.39%	95.41%	81.24%
Proposed	<b>79.12%</b>	<b>61.34%</b>	<b>80.11%</b>	<b>66.53%</b>	<b>92.67%</b>	<b>75.06%</b>	<b>92.15%</b>	<b>79.98%</b>



**Figure 9.** Evaluation results on MIMII test data.

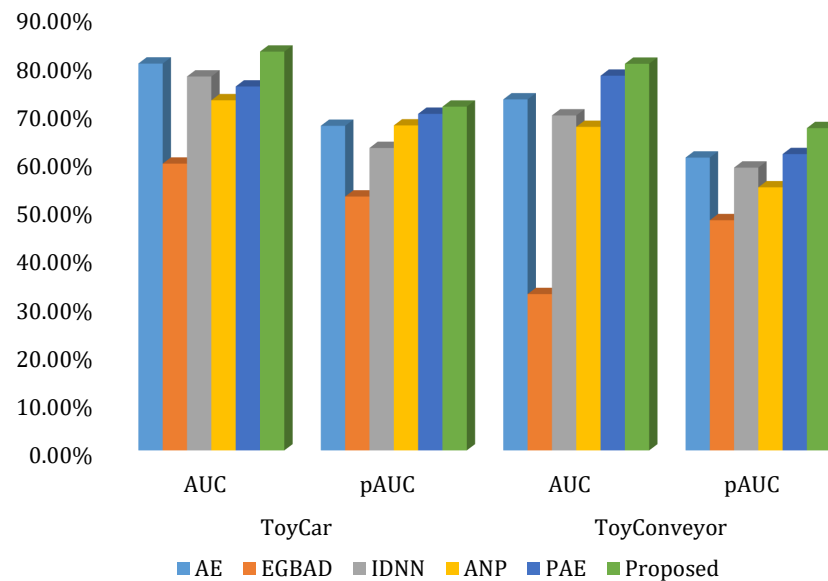
We also evaluate the performance of our model on the ToyADMOS dataset. According to the experimental results in Table 3 and Figure 10, our method achieved the best performance in both AUC and pAUC metrics on ToyCar, reaching 82.56% and 71.18%, respectively. Specifically, PAE achieved an AUC of 75.35% and a pAUC of 69.70%, slightly lower than our method. AE attained an AUC of 80.09%, close to our method, but its pAUC was 67.22%, which is the consequence of its simple pattern. IDNN achieved an AUC of

77.42% and a pAUC of 62.64%, while ANP reached an AUC of 72.50% and a pAUC of 67.30%. In contrast, EGBAD had the poorest performance among all methods, with an AUC of 59.40% and a pAUC of 52.60%, which demonstrates it can hardly handle this kind of task. On the ToyConveyor dataset, our method still outperformed all others in both AUC and pAUC, achieving 80.05% and 66.75%, respectively, showing a strong generalization ability. PAE demonstrated stable performance with an AUC of 77.58% and a pAUC of 61.37%. AE attained an AUC of 72.68% and a pAUC of 60.65%, while IDNN achieved an AUC of 69.36% and a pAUC of 58.58%. ANP reached an AUC of 67.00% and a pAUC of 54.50%. Notably, EGBAD performed significantly worse on these datasets, with an AUC of only 32.40% and a pAUC of 47.70%, indicating substantial limitations in its performance on the ToyConveyor dataset. Overall, our method surpassed all other methods in both AUC and pAUC across the two datasets, demonstrating stronger fault detection capabilities. Particularly on the ToyConveyor dataset, EGBAD's AUC was only 32.40%, far lower than other methods, while our method achieved 80.05%, showcasing a significant performance advantage. Additionally, PAE exhibited stable performance on both datasets but was still slightly inferior to our method. AE and IDNN performed well on the ToyCar dataset but showed worse performance on the ToyConveyor dataset, suggesting potential adaptability issues across different datasets. In contrast, our method excelled on both datasets, demonstrating strong generalization and robustness. These results indicate that our method is more effective in detecting faults on the ToyADMOS dataset, which is more focused on smaller faults and more exact diagnoses, showing the high practical value of our model for real-world applications.

**Table 3.** Evaluation results on ToyADMOS test data, where the bold font indicates the best performance.

Method	ToyCar		ToyConveyor	
	AUC	pAUC	AUC	pAUC
AE [35]	80.09%	67.22%	72.68%	60.65%
EGBAD [38]	59.40%	52.60%	32.40%	47.70%
IDNN [39]	77.42%	62.64%	69.36%	58.58%
ANP [40]	72.50%	67.30%	67.00%	54.50%
PAE [41]	75.35%	69.70%	77.58%	61.37%
Proposed	<b>82.56%</b>	<b>71.18%</b>	<b>80.05%</b>	<b>66.75%</b>

The lack of on-site experiments to evaluate the model's performance in real-world application has always been a drawback of the current studies. To address this issue, we collected real-world data to conduct the test; the results are presented in Table 4. It can be seen that our method performed substantially better in both AUC and pAUC, achieving an AUC of 80.67% and a pAUC of 70.33%. Compared to the second-best PAE method, our approach outperformed it by 8.5 percentage in AUC and 7.06 percentage in pAUC. This highlights the significant superiority of our method in the detection of fault features in wind turbine blades. Through detailed comparison, it is evident that our method exhibits exceptional performance in real-world wind turbine blade fault detection, particularly in robustness under complex environments and adaptability to small-sample data, surpassing other methods. This provides reliable technical support for real-time fault monitoring and maintenance of wind turbine blades.

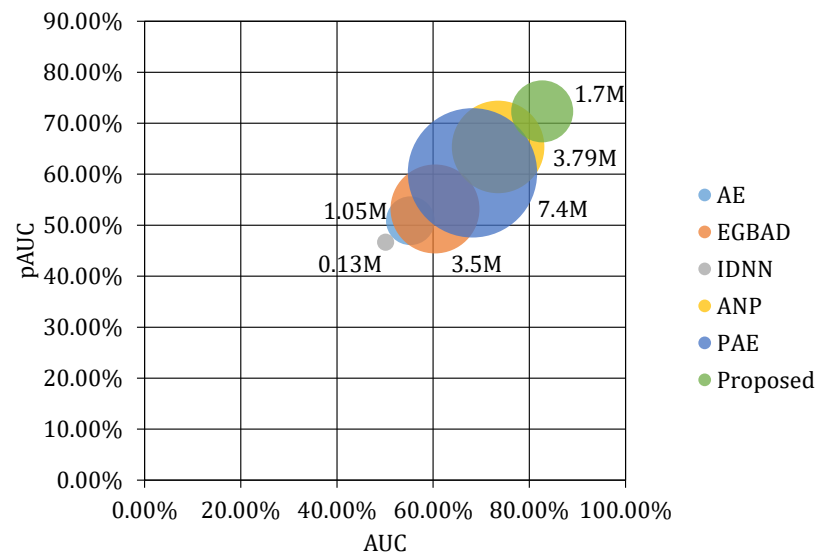


**Figure 10.** Evaluation results on ToyADMOS test data.

**Table 4.** Evaluation results on our collected wind blade test data, where the bold font indicates the best performance.

Method	AUC	pAUC
AE [35]	65.26%	60.88%
EGBAD [38]	70.38%	63.19%
IDNN [39]	69.99%	62.26%
ANP [40]	71.50%	63.35%
PAE [41]	72.17%	63.27%
Proposed	<b>80.67%</b>	<b>70.33%</b>

Figure 11 illustrates the relationship between the performance and the number of parameters of different models. The horizontal axis represents AUC, the vertical axis represents pAUC, and the size of the bubbles corresponds to the number of model parameters. Both AUC and pAUC are key metrics for evaluating model performance. Higher values indicate stronger classification capability. The bubble size reflects model complexity, with smaller values indicating a more lightweight model. From the plot, it is evident that our method (green bubble, 1.7 M parameters) excels in both AUC and pAUC, positioned in the high-performance region at the top right. This indicates that it achieves superior performance while maintaining a moderate parameter count, demonstrating strong efficiency. In contrast, other models show certain shortcomings in either performance or efficiency. AE (blue bubble, 1.05 M parameters) has fewer parameters but lower AUC and pAUC compared to our method, indicating a weaker performance. EGBAD (orange bubble, 3.5 M parameters) has a larger parameter but only marginal performance improvement, with both AUC and pAUC still trailing behind our model. IDNN (gray bubble, 0.13 M parameters) has the fewest parameters but the worst performance, with extremely low AUC and pAUC, making it unsuitable for practical applications. ANP (yellow bubble, 3.79 M parameters) and PAE (light blue bubble, 7.4 M parameters) show some improvements in performance, but their parameters increase significantly, leading to much higher model complexity, particularly for PAE, with a parameter count as high as 7.4 M, which far exceeds our method; moreover, the performance gain is not substantial, indicating low parameter efficiency. In contrast, our method achieves the best balance between performance and efficiency.

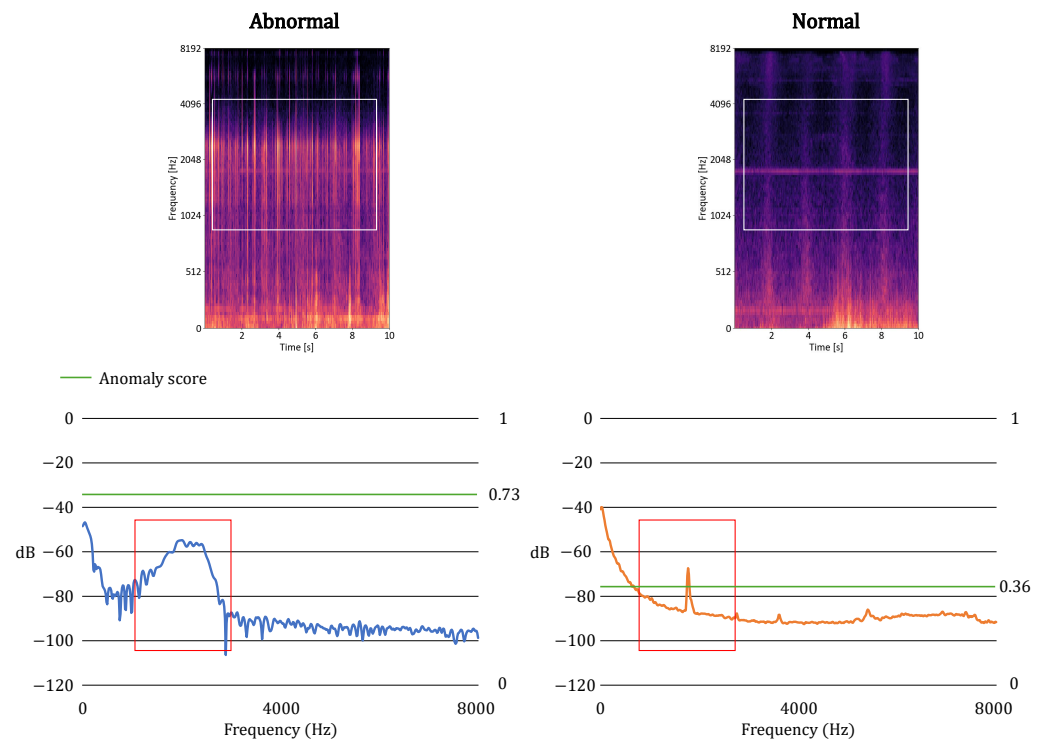


**Figure 11.** Scatter plot based on accuracy and number of parameters.

In addition, we also studied the interpretability of the model for a single sample. It is worth mentioning that the anomaly score is calculated based on Equations (13)–(15) and (20). As shown in Figure 12, the biggest difference between the audio of a faulty wind turbine blade and the audio of normal wind turbine equipment is the frequency at 2000–4000 Hz. The energy of the former in this part is significantly stronger than that of the latter, as shown in the white box in the upper half of Figure 12 and the red box in the lower half. This part of the energy indicates possible abnormal noise or damage of the equipment when it is in working condition. During the inference phase, we found that our model can capture such features. Therefore, the anomaly score of the former is 0.73 after inference, which is significantly higher than the latter's 0.36. This shows that the model has learned the domain knowledge in this area through a large amount of data combined with the phase characteristics of the latent space and can make accurate judgments on real data.

In order to study the impact of different hyperparameter settings and model structures on performance, we conducted ablation experiment on the collected wind turbine blade audio data. Similarly, it was mainly measured by two metrics, AUC and pAUC. As shown in Table 5, the experimental results show that the choice of model structure and hyperparameters has a significant impact on performance. First, when the PA module is removed (w/o PA), the AUC is 76.26% and the pAUC is 67.79%. This is due to the lack of balance of phase information in the latent space caused by the absence of the PA module. In addition, when the discriminator is removed (w/o  $\mathcal{D}$ ) and only general loss is used as the training target, the AUC is only 61.27% and the pAUC is 55.51%, which shows that adversarial training is crucial to improving the performance of unsupervised tasks.

On the other hand, we analyze the impact of hyperparameter combinations. We tested six different weight combinations of loss functions in the experiment, and the optimal combination  $w_1 = 1, w_2 = 50, w_3 = 1$  significantly outperformed other settings in both AUC and pAUC. Compared with the second-best combination  $w_1 = 50, w_2 = 1, w_3 = 1$ , the AUC is improved by 3.29% and the pAUC is improved by 2.44%. In addition, compared with the setting of removing the PA module (w/o PA), the AUC of this combination is improved by 4.41% and the pAUC is improved by 2.54%; compared with the setting of removing the discriminator  $\mathcal{D}$  (w/o  $\mathcal{D}$ ), the AUC is improved by 19.40% and the pAUC is improved by 14.82%. These comparisons can verify the positive effect of the settings we adopted in the overall architecture.



**Figure 12.** Interpretability analysis of the proposed model. The red rectangle highlights the anomalous region detected by our method. The blue curve represents the frequency amplitude of anomaly audio, while the orange curve denotes the frequency amplitude of normal audio.

**Table 5.** Results of ablation study experiments.

	AUC	pAUC
w/o PA	76.26%	67.79%
w/o $\mathcal{D}$	61.27%	55.51%
$w_1 = 50, w_2 = 1, w_3 = 1$	77.38%	67.89%
$w_1 = 1, w_2 = 50, w_3 = 1$	<b>80.67%</b>	<b>70.33%</b>
$w_1 = 1, w_2 = 1, w_3 = 50$	71.72%	61.33%
$w_1 = 1, w_2 = 50, w_3 = 50$	75.68%	67.02%
$w_1 = 50, w_2 = 1, w_3 = 50$	73.93%	65.34%
$w_1 = 50, w_2 = 50, w_3 = 1$	71.25%	61.50%

#### 4.6. Limitations and Prospects

**Audio duration.** Our experiments used fixed 10 s segments following the DCASE protocol. In real-world applications, we recommend to perform padding to audio samples that are less than 10 s and cutting to those that are more than 10 s. While this simplifies training, it may impact performance. Future work will explore adaptive segmentation.

**Low-SNR case and volume change.** Our MIMII-DUE dataset has considered natural factory noise ( $SNR \in [5, 15]$  dB) and volume change, but extreme noise scenarios (e.g.,  $SNR = -5$  dB) were not considered in this research. The system may degrade in such very low-SNR scenarios.

**Prospects.** For multi-channel audio, extending the encoder to process spatial dimensions via cross-channel attention can be a solution. Additionally, multi-sensor fusion is feasible through early or late integration strategies. Early fusion directly merges the original data and learns cross-modal features through a unified encoder, while late fusion independently processes the data of each sensor and then fuses the decision, which is more resistant to interference but may lose fine-grained associations.



## 5. Conclusions

This study introduces a novel unsupervised anomaly detection architecture for electrical equipment, centered on a parallel attention mechanism that enables multi-scale feature distillation. This core innovation, which processes time–frequency features simultaneously through dedicated attention pathways, directly enhances robustness and real-world performance. Experimental validation on benchmark datasets (MIMII, ToyADMOS) and domain-specific trials (wind turbine blades) confirms state-of-the-art accuracy, substantiating the architecture’s efficacy. The lightweight design (1.7 M parameters) further ensures edge deployability, while the dual-encoding strategy improves feature discrimination and model interpretability. Future work will extend this attention-driven framework to multi-modal data fusion and few-shot learning, advancing smart manufacturing applications.

**Author Contributions:** Methodology, W.Z., S.Z. and J.Y.; Software, Y.C. and J.Y.; Writing—original draft, J.Y.; Writing—review & editing, J.Y. and H.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Tran, T.; Lundgren, J. Drill Fault Diagnosis Based on the Scalogram and Mel Spectrogram of Sound Signals Using Artificial Intelligence. *IEEE Access* **2020**, *8*, 203655–203666. [\[CrossRef\]](#)
2. Huynh, K.; Barros, A.; Bérenguer, C.; Castro, I. A periodic inspection and replacement policy for systems subject to competing failure modes due to degradation and traumatic events. *Reliab. Eng. Syst. Saf.* **2011**, *96*, 497–508. [\[CrossRef\]](#)
3. Guo, Y.; Wang, J.; Chen, H.; Li, G.; Huang, R.; Yuan, Y.; Ahmad, T.; Sun, S. An expert rule-based fault diagnosis strategy for variable refrigerant flow air conditioning systems. *Appl. Therm. Eng.* **2019**, *149*, 1223–1235. [\[CrossRef\]](#)
4. Song, Q.; Jiang, X.; Du, G.; Liu, J.; Zhu, Z. Smart multichannel mode extraction for enhanced bearing fault diagnosis. *Mech. Syst. Signal Process.* **2023**, *189*, 110107. [\[CrossRef\]](#)
5. Vos, K.; Peng, Z.; Jenkins, C.; Shahriar, M.R.; Borghesani, P.; Wang, W. Vibration-based anomaly detection using LSTM/SVM approaches. *Mech. Syst. Signal Process.* **2022**, *169*, 108752. [\[CrossRef\]](#)
6. Dozsa, T.; Rado, J.; Volk, J.; Kisari, A.; Soumelidis, A.; Kovács, P. Road Abnormality Detection Using Piezoresistive Force Sensors and Adaptive Signal Models. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 9509211. [\[CrossRef\]](#)
7. Liu, H.; Zhang, X.; Dong, H.; Liu, Z.; Hu, X. Magnetic Anomaly Detection Based on Energy-Concentrated Discrete Cosine Wavelet Transform. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 9700210. [\[CrossRef\]](#)
8. Cao, Y.; Ji, R.; Huang, X.; Lei, G.; Shao, X.; You, I. Empirical mode decomposition-empowered network traffic anomaly detection for secure multipath TCP communications. *Mob. Netw. Appl.* **2022**, *27*, 2254–2263. [\[CrossRef\]](#)
9. Keshun, Y.; Zengwei, L.; Yingkui, G. A performance-interpretable intelligent fusion of sound and vibration signals for bearing fault diagnosis via dynamic CAME. *Nonlinear Dyn.* **2024**, *112*, 20903–20940. [\[CrossRef\]](#)
10. Akcay, S.; Atapour-Abarghouei, A.; Breckon, T.P. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Proceedings of the Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Revised Selected Papers, Part III 14*; Springer: Cham, Switzerland, 2019; pp. 622–637. [\[CrossRef\]](#)
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30*.
12. Zhang, H.; Zhang, S.; He, Q.; Kong, F. The Doppler Effect based acoustic source separation for a wayside train bearing monitoring system. *J. Sound Vib.* **2016**, *361*, 307–329. [\[CrossRef\]](#)
13. Kumar, J.; Sharma, S.; Bharti, A.K. Fault detection and classification in automobile engine based on its audio signature using support vector machine. In *Proceedings of the ICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering, Hyderabad, India, 9–10 April 2021; pp. 103–114*. [\[CrossRef\]](#)

14. Huang, X.; Teng, Z.; Tang, Q.; Yu, Z.; Hua, J.; Wang, X. Fault diagnosis of automobile power seat with acoustic analysis and retrained SVM based on smartphone. *Measurement* **2022**, *202*, 111699. [\[CrossRef\]](#)
15. Akbal, A. A local knit pattern-based automated fault classification method for the cooling system of the data center. *Appl. Acoust.* **2021**, *176*, 107888. [\[CrossRef\]](#)
16. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [\[CrossRef\]](#)
17. Zhu, W.; Liu, H.; Zhou, Y.; Gan, L.; Ma, Y. Wind turbine blade fault detection by acoustic analysis: Preliminary results. In Proceedings of the 2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xi'an, China, 17–20 August 2021; pp. 1–5. [\[CrossRef\]](#)
18. Zhang, J.; Liu, W.; Lan, J.; Hu, Y.; Zhang, F. Audio Fault Analysis for Industrial Equipment Based on Feature Metric Engineering with CNNs. In Proceedings of the 2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE), Wuhan, China, 4–6 November 2021; pp. 409–416. [\[CrossRef\]](#)
19. Chen, S.H.; Wang, J.C.; Lin, H.J.; Lee, M.H.; Liu, A.C.; Wu, Y.L.; Hsu, P.S.; Yang, E.C.; Jiang, J.A. A machine learning-based multiclass classification model for bee colony anomaly identification using an IoT-based audio monitoring system with an edge computing framework. *Expert Syst. Appl.* **2024**, *255*, 124898. [\[CrossRef\]](#)
20. Peng, B.; Wang, K.; Abdulla, W. Robust Classification of Urban Sounds in Noisy Environments: A Novel Approach Using SPWVD-MFCC and Dual-Stream Classifier. *Acoust. Aust* **2025**. [\[CrossRef\]](#)
21. Peng, B.; Li, D.; Wang, K.I.K.; Abdulla, W.H. Acoustic-Based Industrial Diagnostics: A Scalable Noise-Robust Multiclass Framework for Anomaly Detection. *Processes* **2025**, *13*, 544. [\[CrossRef\]](#)
22. Yan, H.; Zhan, X.; Wu, Z.; Cheng, J.; Wen, L.; Jia, X. Unsupervised anomalous sound detection method based on Gammatone spectrogram and adversarial autoencoder with attention mechanism. *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.* **2024**, 09544089241258027. [\[CrossRef\]](#)
23. Neto, W.A.d.O.; Guedes, E.B.; Figueiredo, C.M.S. Anomaly Detection in Sound Activity with Generative Adversarial Network Models. *J. Internet Serv. Appl.* **2024**, *15*, 313–324. [\[CrossRef\]](#)
24. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
25. Hojati, H.; Armanfard, N. Self-supervised acoustic anomaly detection via contrastive learning. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 3253–3257.
26. Guan, J.; Xiao, F.; Liu, Y.; Zhu, Q.; Wang, W. Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
27. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
28. Bai, J.; Chen, J.; Wang, M.; Ayub, M.S.; Yan, Q. SSDPT: Self-supervised dual-path transformer for anomalous sound detection. *Digit. Signal Process.* **2023**, *135*, 103939. [\[CrossRef\]](#)
29. Shams, S.; Dindar, S.S.; Jiang, X.; Mesgarani, N. SSAMBA: Self-Supervised Audio Representation Learning With Mamba State Space Model. In Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT), Macau, China, 2–5 December 2024; pp. 1053–1059. [\[CrossRef\]](#)
30. Qu, X.; Liu, Z.; Wu, C.Q.; Hou, A.; Yin, X.; Chen, Z. MFGAN: Multimodal Fusion for Industrial Anomaly Detection Using Attention-Based Autoencoder and Generative Adversarial Network. *Sensors* **2024**, *24*, 637. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Bi, Z.; Li, H.; Zhang, W.; Dong, Z. Variational bayesian clustering algorithm for unsupervised anomalous sound detection incorporating VH-BCL+. *Multimed. Tools Appl.* **2024**, *83*, 43777–43800. [\[CrossRef\]](#)
32. Khan, M.; Negi, A.; Kulkarni, A.; Phutke, S.S.; Vipparthi, S.K.; Murala, S. Phaseformer: Phase-based Attention Mechanism for Underwater Image Restoration and Beyond. *arXiv* **2024**, arXiv:2412.01456.
33. Purohit, H.; Tanabe, R.; Ichige, T.; Endo, T.; Nikaido, Y.; Suefusa, K.; Kawaguchi, Y. MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York, NY, USA, 25–26 October 2019; pp. 209–213.
34. 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 the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 313–317. [\[CrossRef\]](#)
35. Ono, N.; Harada, N.; Kawaguchi, Y.; Mesaros, A.; Imoto, K.; Koizumi, Y.; Komatsu, T. (Eds.) *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*; Zenodo: Geneva, Switzerland, 2020. [\[CrossRef\]](#)
36. Liu, H.; Zhu, W.; Zhou, Y.; Shi, L.; Gan, L. Nonintrusive wind blade fault detection using a deep learning approach by exploring acoustic informationa). *J. Acoust. Soc. Am.* **2023**, *153*, 538–547. [\[CrossRef\]](#)
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

38. Zenati, H.; Foo, C.S.; Lecouat, B.; Manek, G.; Chandrasekhar, V.R. Efficient gan-based anomaly detection. *arXiv* **2018**, arXiv:1802.06222.
39. Suefusa, K.; Nishida, T.; Purohit, H.; Tanabe, R.; Endo, T.; Kawaguchi, Y. Anomalous Sound Detection Based on Interpolation Deep Neural Network. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 271–275. [[CrossRef](#)]
40. Wichern, G.; Chakrabarty, A.; Wang, Z.Q.; Roux, J.L. Anomalous Sound Detection Using Attentive Neural Processes. In Proceedings of the 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 17–20 October 2021; pp. 186–190. [[CrossRef](#)]
41. Zeng, X.M.; Song, Y.; Dai, L.R.; Liu, L. Predictive AutoEncoders Are Context-Aware Unsupervised Anomalous Sound Detectors. In Proceedings of the Man-Machine Speech Communication, Hefei, China, 15–18 December 2022; Zhenhua, L., Ling, Z., Gao, J., Yu, K., Jia, J., Eds.; Springer, Singapore, 2023; pp. 101–113. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.