2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

LCAA: Lightweight Convolutional Attention Autoencoder for Acoustic Anomaly Detection

Yuxue Wang[1] and Chenhao Ye[1]

[1] School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China
chaechaexue@gmail.com

Abstract. Industrial machinery monitoring is pivotal in modern manufacturing, where unexpected equipment failures could incur significant economic and operational costs. In this work, we introduce LCAA, a novel unsupervised framework tailored for acoustic anomaly detection in industrial environments. Our approach synergistically combines convolutional neural networks with multi-head attention mechanisms within a compact autoencoder architecture, enabling the effective capture of both temporal and frequency domain features inherent in acoustic signals. By selectively focusing on the most informative components of the input, the proposed model enhances feature extraction, leading to improved detection accuracy and faster convergence compared to traditional methods. Extensive experiments on multiple benchmark datasets demonstrate that LCAA not only outperforms state-of-the-art baselines in detecting subtle anomalies but also maintains a minimal parameter footprint, thereby facilitating real-time deployment on resource-constrained edge devices. This study contributes a robust and efficient solution for proactive maintenance strategies, promoting enhanced operational reliability and reduced downtime in industrial systems.

Keywords: Acoustic anomaly detection, Convolutional neural networks, Attention mechanisms, Autoencoder, Industrial monitoring..

1 Introduction

Industrial systems and machinery are the backbone of modern manufacturing processes, where unplanned downtime due to equipment failure could result in significant economic losses. Traditional maintenance approaches such as reactive maintenance (fix after failure) or schedule-based preventive maintenance are increasingly inadequate for today's complex industrial environments [1]. This has driven the emergence of condition-based monitoring (CBM) and predictive maintenance systems that could detect anomalous equipment behavior before catastrophic failures occur [2,3].

Acoustic-based condition monitoring has gained significant attention in recent years due to its non-intrusive nature and rich information content [4]. Sound signatures emitted by machinery contain valuable diagnostic information about their operational states,

as mechanical defects often manifest as distinctive acoustic patterns [5]. However, conventional approaches to acoustic-based anomaly detection typically rely on supervised learning techniques that require extensive labeled datasets for each failure mode [6,7]. This presents a substantial challenge in industrial environments where anomaly data is inherently scarce and expensive to obtain, as machinery must be deliberately damaged to collect such data [8].

To address these limitations, unsupervised learning approaches have emerged as promising alternatives that could identify anomalies without requiring labeled examples of fault conditions [9,10]. Recent studies have demonstrated the effectiveness of various unsupervised techniques for industrial anomaly detection, including autoencoders [11], generative adversarial networks [12], and one-class classification methods [13]. However, these approaches often face challenges related to model complexity, convergence speed, and computational efficiency—particularly critical factors for deployment on resource-constrained edge devices in industrial settings [14,15].

Deep autoencoders have shown particular promise for anomaly detection due to their ability to learn compact representations of normal data distributions [16]. By reconstructing the input data through a bottleneck layer, autoencoders could effectively capture the underlying data manifold of normal operational states. The reconstruction error then serves as an anomaly score, with higher errors indicating potential anomalies [17]. However, standard autoencoder architectures may struggle to capture temporal dependencies and complex patterns in acoustic data, limiting their effectiveness for machinery monitoring [18].

Attention mechanisms, first introduced in the context of neural machine translation [19], have revolutionized numerous domains by enabling models to focus selectively on the most relevant parts of the input. In the context of anomaly detection, attention mechanisms could help identify salient features in the acoustic signals that are most indicative of normal or abnormal operation [20]. The integration of multi-head attention with autoencoders presents an opportunity to enhance feature extraction while maintaining computational efficiency.

In this paper, we propose a novel unsupervised acoustic anomaly detection system for industrial equipment monitoring that addresses the aforementioned challenges. Our approach combines a lightweight autoencoder architecture with multi-head attention mechanisms and convolutional neural networks to effectively capture both temporal and frequency domain features in acoustic signals. The proposed model achieves superior performance in terms of both anomaly detection accuracy and convergence speed compared to state-of-the-art baselines, while maintaining a minimal parameter footprint suitable for edge deployment.

The key contributions of our work are as follows:

- We introduce a compact, attention-enhanced autoencoder architecture specifically designed for unsupervised acoustic anomaly detection in industrial settings.
- We demonstrate that our proposed approach outperforms existing baselines across multiple benchmark datasets in terms of initial performance and convergence rate.

- We validate the practical applicability of our model through successful deployment on resource-constrained edge hardware (Arduino Nano 33 BLE Sense), making real-time acoustic monitoring accessible for a wide range of industrial applications.

2 Methodology

2.1 Problem Formulation

Acoustic anomaly detection could be formulated as identifying deviations from normal acoustic patterns. Given a set of acoustic signals from normal operating equipment, we aim to learn a model that could distinguish between normal and anomalous signals without requiring examples of anomalies during training.

Let $\mathcal{X} = x_1, x_2, \dots, x_N$ represent a set of acoustic signals recorded from machinery operating under normal conditions. Each signal x_i is first transformed into a log-Mel spectrogram $S_i \in \mathbb{R}^{T \times F}$, where T is the number of time frames and F is the number of frequency bands. Traditional approaches typically process the entire spectrogram S_i as input to reconstruct the same, resulting in models with large parameter counts and computational requirements that are prohibitive for edge deployment.

To address this limitation, we propose a novel frame prediction approach. Instead of reconstructing the entire input spectrogram, we partition each spectrogram S_i into consecutive segments$s_i^1, s_i^2, \dots, s_i^K$, where each segment $s_i^j \in \mathbb{R}^{(2w+1) \times F}$ contains $2w + 1$ consecutive frames centered around a time step t. Our key insight is to use the surrounding $2w$ frames to predict the center frame, effectively transforming the problem from full reconstruction to center frame prediction:

$$\mathbf{x}_{\text{in}} = \left[s_{i,t-w}, \dots, s_{i,t-1}, s_{i,t+1}, \dots, s_{i,t+w}\right] \in \mathbb{R}^{2w \times F} \tag{1}$$

$$\mathbf{x}_{\text{target}} = s_{i,t} \in \mathbb{R}^{F} \tag{2}$$

where $s_{i,t}$ represents the frame at time step t in segment s_i^j. This formulation significantly reduces the model's parameter count while focusing the learning task on modeling the temporal dependencies between adjacent frames.

During inference, the anomaly score for a test signal x_{test} is computed as the reconstruction error between the predicted center frames and the actual center frames:

$$\text{score}(x_{\text{test}}) = \frac{1}{K} \sum_{j=1}^{K} \| f_\theta\left(\mathbf{x}_{\text{in}}^j\right) - \mathbf{x}_{\text{target}}^j \|_2^2 \tag{3}$$

where f_θ represents our proposed LCAA model with parameters θ, and K is the number of segments in the test signal.

2.2 Lightweight Convolutional Attention Autoencoder Architecture

The LCAA architecture combines the strengths of autoencoders, convolutional neural networks, and attention mechanisms to efficiently capture both local and global dependencies in acoustic signals. Fig.1 illustrates the overall structure of our proposed model.

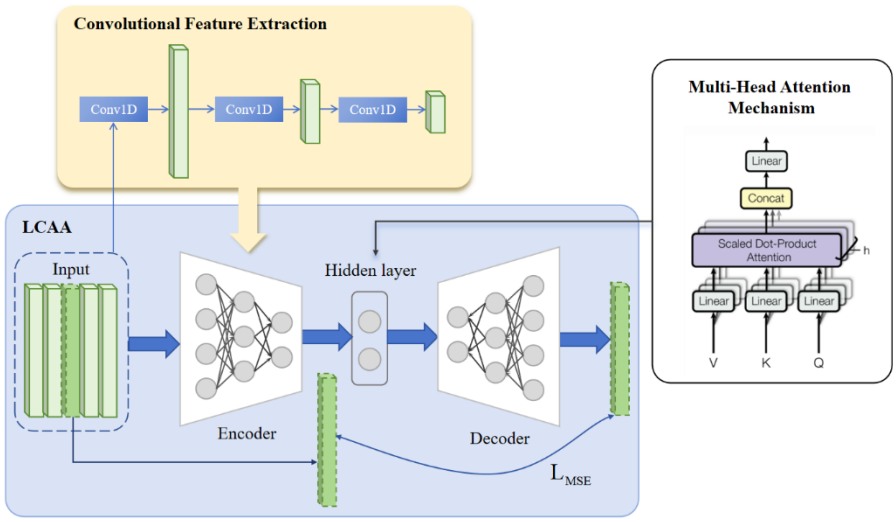

Fig. 1. The architecture of the proposed Lightweight Convolutional Attention Autoencoder (LCAA).

Encoder-Decoder Framework. The core of our architecture follows the autoencoder paradigm, consisting of an encoder that maps the input to a latent representation and a decoder that reconstructs the target from this representation. However, unlike traditional autoencoders that aim to reconstruct the input itself, our model is designed to predict the center frame from the surrounding frames.

The encoder E_{θ_e} maps the input frames $x_{\mathbf{in}} \in R^{2w \times F}$ to a latent representation $z \in R^d$:

$$z = E_{\theta_e}(x_{\mathbf{in}}) \tag{4}$$

The decoder D_{θ_d} then maps this latent representation to the predicted center frame $\hat{\mathbf{x}}_{\text{target}} \in \mathbb{R}^F$:

$$\hat{\mathbf{x}}_{\text{target}} = D_{\theta_d}(\mathbf{z}) \tag{5}$$

This center frame prediction approach significantly reduces the model's complexity compared to full reconstruction methods, making it more suitable for edge deployment.

Additionally, by focusing on predicting a single frame rather than reconstructing the entire input, the model could better capture the temporal dependencies between adjacent frames, leading to more precise predictions.

Convolutional Feature Extraction. While fully connected layers are commonly used in autoencoder architectures, they may fail to capture local patterns and temporal dependencies effectively in acoustic signals. To address this limitation, we incorporate convolutional layers in our encoder to extract hierarchical local features.

For each input $\mathbf{x}_{\text{in}} \in \mathbb{R}^{2w \times F}$, we apply a series of 1D convolutional operations along the time dimension:

$$h_1 = \sigma(W_1 * x_{\text{i}} + b_1) \tag{6}$$

$$h_l = \sigma(W_l * h_{l-1} + b_l), \quad \text{for } l = 2,3 \tag{7}$$

where $*$ denotes the convolution operation, W_l and b_l are the weights and biases of the $l-th$ convolutional layer, σ is the activation function (ReLU in our implementation), and h_l represents the feature maps at the $l-th$ layer.

We design a three-layer convolutional module with carefully chosen kernel sizes and strides to gradually extract multi-level features from the input data. The convolutional layers use kernel sizes of (3,1), (3,1), and (2,1) with strides of (1,1), (1,1), and (1,1) respectively. This configuration allows the model to capture temporal patterns at different scales.

To enhance the model's representational capacity, we combine the features extracted by the convolutional layers with the outputs of fully connected layers at the same scale, creating a multi-scale feature fusion mechanism:

$$f_l = h_l \oplus \text{FC}_l(x_{\text{i}}), \quad \text{for } l = 1,2,3 \tag{8}$$

where $\oplus$ denotes channel-wise concatenation, and FC_l represents the $l-th$ fully connected layer. This fusion mechanism allows the model to benefit from both the local feature extraction capabilities of convolutional neural networks(CNNs) and the global mapping abilities of fully connected layers.

Multi-Head Attention Mechanism. While convolutional layers excel at capturing local patterns, they may struggle to model long-range dependencies in the data. To overcome this limitation, we incorporate multi-head attention mechanisms in the bottleneck layer of our autoencoder.

Given the fused features f_3 from the final convolutional layer, we first apply a projection layer to obtain the query, key, and value matrices:

$$Q = f_3 W^Q, \quad K = f_3 W^K, \quad V = f_3 W^V \tag{9}$$

where $W^Q, W^K, W^V \in R^{d_{model} \times d_k}$ are learnable projection matrices.

The attention mechanism computes the weighted sum of values, where the weights are determined by the compatibility of the query with the corresponding keys:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

To capture different aspects of the input data, we employ multi-head attention with h parallel attention heads:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{11}$$

where

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{12}$$

and $W_i^Q, W_i^K, W_i^V \in R^{d_{model} \times d_k/h}$ and $W^O \in R^{d_{model} \times d_{model}}$ are learnable parameters.

The multi-head attention module enables the model to jointly attend to information from different representation subspaces, thereby capturing complex relationships between different parts of the input sequence. This is particularly beneficial for acoustic anomaly detection, where anomalies may manifest as subtle deviations in the temporal or frequency domains.

Training Objective. The training objective of our model is to minimize the reconstruction error between the predicted center frame and the actual center frame. We employ the mean squared error (MSE) loss for this purpose:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \| \hat{\mathbf{x}}_{\text{target}}^i - \mathbf{x}_{\text{target}}^i \|_2^2 \tag{13}$$

where $\hat{\mathbf{x}}_{\text{target}}^i$ is the predicted center frame for the i-th sample, and $\mathbf{x}_{\text{target}}^i$ is the corresponding ground truth.

During training, we only use normal acoustic signals to optimize the model parameters, following the common practice in unsupervised anomaly detection. The model learns to capture the underlying patterns of normal operation, and deviations from these patterns during inference are flagged as potential anomalies.

3 Experiments

3.1 Training Set

Dataset and Preprocessing. The proposed method is evaluated on the MIMII dataset [21], which contains sounds generated by four types of industrial machines, namely Valve, Pump, Fan and Slider. Each recording is a mono 10-second long audio file of

the target operating machinery sound mixed with ambient noise. Both the test and training samples have a sampling rate of 16 kHz, a bit rate of 256k, and the samples are encoded as 16bit signed integer PCM. Fig. 2 illustrates the log-Mel spectrograms of sample sounds from four machine types.

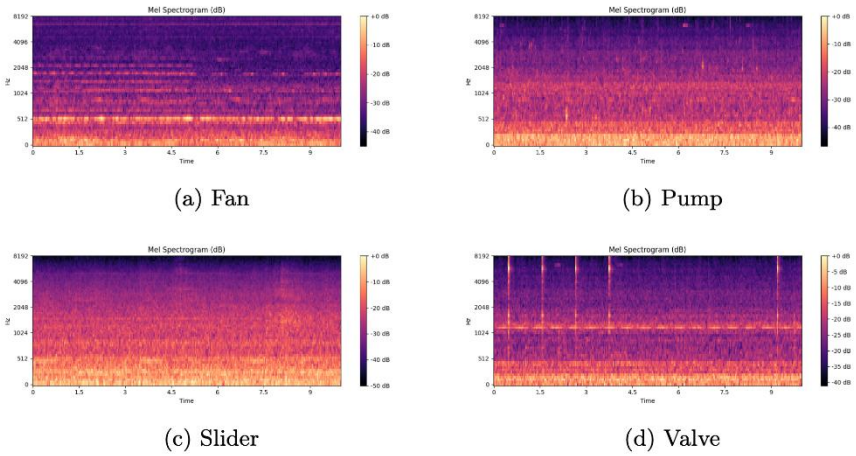

Fig. 2. Examples of log-Mel spectrograms of the original sound

For data preprocessing, the raw audio recordings are converted into a logarithmic Mel-band energy feature vector with 64 bands; using a 1024-bin Fast Fourier Transform with a jump size of 512. each audio recording is converted into a 64 × 313 frequency-time matrix. This matrix is then converted into an overlapping window consisting of five consecutive frames and after extracting the center frame, it is sent to the lightweight self-encoder architecture proposed in this paper to determine whether it is an anomalous sample or not based on the generated reconstruction loss.

As can be seen from the figure, within these four machine type sounds, the valve sound has a distinct non-smoothness.

Baseline System. We used the model provided by the DCASE2020 competition as a baseline system [22]. The baseline system is a simple autoencoder-based anomaly score calculator. The architecture consists of a fully connected neural network (FCN) layer, followed by three hidden FCN layers, and one output FCN layer. Each hidden layer contains 128 hidden units, and the encoder output dimension is 8. The rectified linear unit (ReLU) activation function is applied after each FCN layer, except for the output layer of the decoder. The training process is halted after 100 epochs, with a batch size of 512. The ADAM optimizer is utilized, and the learning rate is fixed at 0.001.

3.2 Results

In this paper, the AUC metric, the number of model parameters, and the number of convergence rounds are used to illustrate model performance, lightweighting, and training efficiency, respectively.

Model Performance. AUC is an important metric to measure the classification performance of a model, and the AUC for each specific machine and the average value for each machine type are given in Table 1. For comparison, the baseline system results are also provided in the table.

In terms of average AUC per machine type, LCAA is much better than the baseline system in most machine types.

Table 1. Dataset Description

Method	Fan	Pump	Slider	Valve
Baseline	0.6583	**0.7289**	0.8476	0.6628
LCAA	**0.8318**	0.6330	**0.9280**	**0.8719**
Isolation Forest	0.7993	0.5691	0.7355	0.5222

Lightweight and Training Efficiency. In the Industrial Internet of Things (IIoT), real-time anomaly detection applications are often deployed on resource-constrained embedded devices such as the Arduino Nano 33 BLE Sense, which has limited storage (e.g., 1MB Flash and 256KB SRAM). To meet these constraints, we designe a lightweight model by streamlining convolutional layers and optimizing the attention mechanism, reducing parameter count to just one-quarter of the baseline model. As shown in Fig. 3, this significantly lowers storage and computation demands while improving performance, making it ideal for IIoT deployment.

The efficiency gains extend to training. Fig. 3 highlights that our model converges in 30 epochs, compared to the baseline's 85 epochs, reducing training costs and time overhead. This accelerates convergence, combined with superior AUC metrics, demonstrates that our solution achieves higher accuracy with fewer resources, fulfilling real-time IIoT requirements.

3.3 Ablation Study

To evaluate the contribution of different components in our proposed LCAA model, we conducted a series of ablation experiments. We systematically removed key components of our architecture—specifically the CNN module and the multi-head attention mechanism—to assess their individual impact on performance. Additionally, we compared our approach against a baseline model that uses a traditional autoencoder structure without our center frame prediction technique. Table 2 presents the AUC-ROC

scores for each configuration across different machinery types and their corresponding fault conditions.

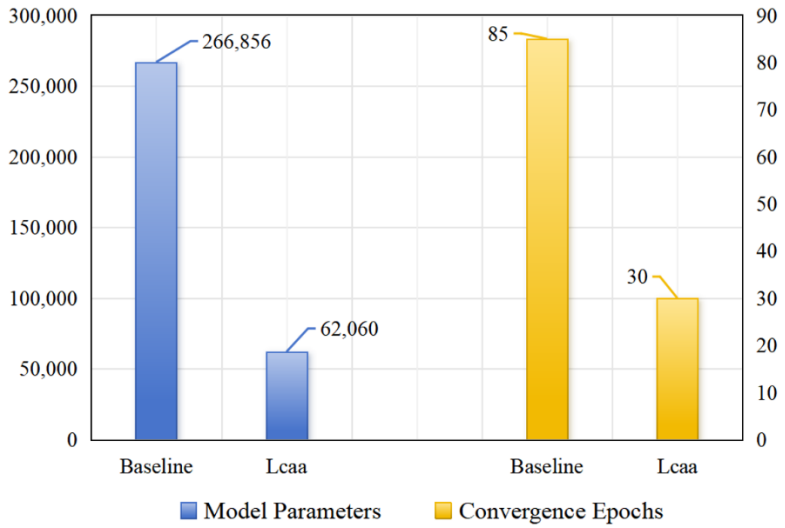

Fig. 3. Comparison in model parameters and convergence epochs

Table 2. Ablation study: AUC-ROC scores for different model configurations across various machinery types and fault conditions.

Machine Type	Full LCAA	w/o CNN	w/o Attention	Baseline
Slider(Avg.)	**0.9280**	0.9161	0.9231	0.8476
Fan (Avg.)	**0.8318**	0.8239	0.8283	0.6583
Pump (Avg.)	0.6330	0.6264	0.6244	**0.7289**
Valve (Avg.)	**0.8719**	0.8621	0.8699	0.6628

3.4 Parameter Sensitivity

We thoroughly investigate the effects of key hyperparameters on LCAA performance on the Slider. First, we systematically adjust the number of attention heads in the attention mechanism by gradually increasing it from 1 to 16 (as shown in Fig.4(a)). The experimental results show that with the increase in the number of attention heads, the model is able to capture more diverse feature relationships and the performance is gradually improved. When the number of attention heads reaches 8, the model performance reaches its peak. However, continuing to increase the number of attention heads does not lead to further performance improvement, but may instead lead to an increase in computational complexity and a potential risk of overfitting.

Next, we explore the effect of the embedding dimension in the attention mechanism on the performance of LCAA. By adjusting the size of the embedding dimension (as shown in Fig. 4(b)), we find that the model exhibited the best performance when the embedding dimension was set to 64. However, when the embedding dimension is further increased to 128, the performance do not improve significantly and even show a slight decrease in some cases. This phenomenon may be related to the limited generalization ability of the model at higher dimension.

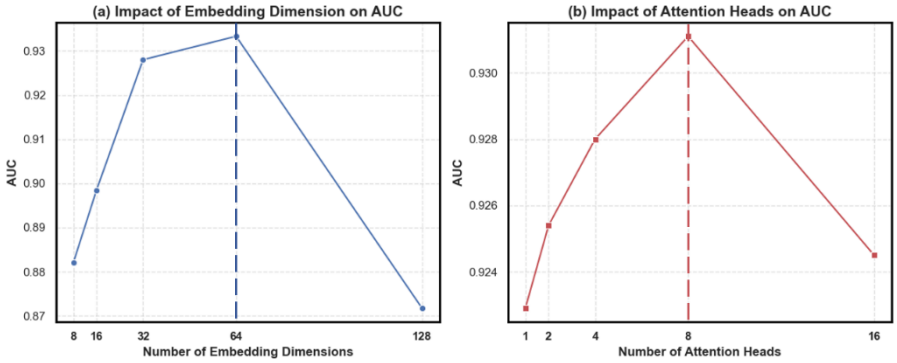

Fig. 4. Parameter sensitivity analysis on Slider: (a) Number of attention heads, and (b) Number of embedding dimension

The experimental results indicate that the number of attention heads and the embedding dimension are key hyperparameters affecting the performance of LCAA, and their proper configuration has a significant impact on the final performance of the model. Although the optimal configuration is 8 for the number of attention heads and 64 for the embedding dimension from the perspective of model performance, considering that the model should be deployed on resource-constrained embedded devices, we ultimately choose to set the number of attention heads to 4 and the embedding dimension to 32. This compromise guarantees the performance of the model while significantly lowering the computational resource requirements, making it more suitable for deployment in real-world applications.

4 Conclusion

In this paper, we presented LCAA, a Lightweight Convolutional Attention Autoencoder for acoustic anomaly detection in industrial equipment. By combining a center frame prediction approach with convolutional neural networks and multi-head attention mechanisms, our model achieves superior anomaly detection performance while maintaining minimal parameter counts suitable for edge deployment. Experimental results on the MIMII dataset demonstrate that LCAA outperforms baseline methods across most machine types, with significant improvements in AUC scores for fans, sliders,

and valves. The model not only exhibits faster convergence during training but also requires substantially fewer parameters, making it particularly suitable for resource-constrained Industrial IoT environments. The ablation studies confirm that both the convolutional feature extraction and attention mechanisms contribute meaningfully to performance gains. Overall, LCAA offers an efficient and effective solution for real-time acoustic anomaly detection in industrial settings where computational resources are limited but reliable monitoring is critical.

References

1. Carvalho, T.P., et al.: A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering 137, 106024 (2019)
2. Jardine, A.K.S., Lin, D., Banjevic, D.: A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing 20(7), 1483–1510 (2006)
3. Lei, Y., et al.: Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical Systems and Signal Processing 104, 799–834 (2018)
4. Randall, R.B.: Vibration-based condition monitoring: industrial, automotive and aerospace applications. Wiley (2021)
5. Tang, L., et al.: A survey of mechanical fault diagnosis based on audio signal analysis. Measurement 220, 113294 (2023)
6. Liu, X., et al.: Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering 35(1), 857–876 (2021)
7. Verma, R., Nagar, V., Mahapatra, S.: Introduction to supervised learning. Data Analytics in Bioinformatics: A Machine Learning Perspective, 1–34 (2021)
8. Michau, G., Fink, O.: Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. Knowledge-Based Systems 216, 106816 (2021)
9. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys 41(3), 1–58 (2009)
10. Pang, T., et al.: Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2043–2052 (2021)
11. Zhang, L., et al.: Self-supervised variational graph autoencoder for system-level anomaly detection. IEEE Transactions on Instrumentation and Measurement 72, 1–11 (2023)
12. Li, M., et al.: GAN compression: Efficient architectures for interactive conditional GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5284–5294 (2020)
13. Ruff, L., et al.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE 109(5), 756–795 (2021)
14. Kong, L., et al.: Edge-computing-driven internet of things: A survey. ACM Computing Surveys 55(8), 1–41 (2022)
15. Zeng, J., Liang, Z.: A deep Gaussian process approach for predictive maintenance. IEEE Transactions on Reliability 72(3), 916–933 (2022)
16. Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407 (2019)
17. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE 2(1), 1–18 (2015)

18. Deng, M., et al.: Intelligent fault diagnosis of rotating components in the absence of fault data: A transfer-based approach. Measurement 173, 108601 (2021)
19. Vaswani, A., et al.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017)
20. Xu, J., et al.: Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642 (2021)
21. Purohit, H., et al.: MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. arXiv preprint arXiv:1909.09347 (2019)
22. Koizumi, Y., et al.: Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), pp. 81–85 (2020)