OPEN ACCESS
IOP Publishing | ISECAEP
Journal of Reliability Science and Engineering
J. Reliab. Sci. Eng. **2** (2026) 025301 (14pp)
https://doi.org/10.1088/3050-2454/ae5498

Mel spectrogram–driven deep learning framework for acoustic emission-based manufacturing monitoring in wire arc additive manufacturing

Fei Gao[1], Yishu Chen[1], Jing Lin[1], Yinmin Zhu[1], Yonghao Miao[1] and Jinghui Tian[2,*]

[1] School of Reliability and Systems Engineering, Beihang University, Xueyuan Road No.37, Haidian District, Beijing, People's Republic of China
[2] Ningbo Institution of Technology (NIT), Beihang University, Ningbo 315832, People's Republic of China

E-mail: jinghuitian@buaa.edu.cn, youfeigao@buaa.edu.cn, 18375152@buaa.edu.cn, linjing@buaa.edu.cn, jimmyz@buaa.edu.cn and miaoyonghao@buaa.edu.cn

Received 26 December 2025, revised 14 March 2026
Accepted for publication 18 March 2026
Published 31 March 2026

Abstract

Acoustic emission shows great potential for *in-situ* monitoring and defect diagnosis in wire arc additive manufacturing (WAAM). However, the AE signal produced during WAAM is heavily contaminated by arc discharge noise, making it difficult to directly extract defect-related features from raw data. This study proposes a Mel spectrogram–based deep learning framework for efficient WAAM health monitoring. The method introduces an enhanced time–frequency representation strategy that captures defect-related features more effectively under the high-noise conditions inherent in WAAM processes. By leveraging Mel spectrogram representations, the framework emphasizes informative low-frequency components while suppressing high-frequency noise, thereby improving feature interpretability and robustness. Convolutional neural network and vision transformer models are employed for defect classification and performance benchmarking. Experimental results demonstrate that the proposed approach achieves high diagnostic accuracy with substantially reduced computational cost, outperforming conventional short-time Fourier transform-based methods. The findings confirm that Mel spectrogram representations offer a more efficient and generalizable solution for health monitoring in WAAM.

Keywords: WAAM, acoustic emission, Mel spectrogram, CNN, manufacturing quality monitoring

* Author to whom any correspondence should be addressed.

1. Introduction

Within recent years, advantages including high deposition rate, low cost, high forming efficiency, high material utilization rate, and good manufacturing flexibility have prompted increased implementation of wire arc additive manufacturing (WAAM) in advanced manufacturing [1, 2]. Yet the complex interactions of the process parameters and the nonlinear thermodynamics dominating the WAAM process make it difficult to ensure the consistency and reliability of the products [3]. Main reasons for fault initiation and propagation are inadequate planning of the printing process, the unstable dynamics of weld pools, the accumulation of excessive heat, and the lack of supply of shielding gases. These problems introduce flaws such as porosity, residual stress, oxidation, and cracking, which reduce the mechanical properties and structural reliability of the fabricated parts [4]. Therefore, real time-monitoring is crucial to ensure stable performance and structural integrity of WAAM-fabricated components.

Conventional approaches used to characterize defects in AM products include radiographic [5, 6], thermal [7, 8], optical [9], and ultrasonic inspection methods [10], which are commonly applied during or after fabrication. However, these methods are often limited by the trade-off between real-time monitoring capability and detection accuracy, thus limiting their overall effectiveness in WAAM applications. Compared to methods mentioned above, acoustic emission (AE) is a highly reliable nondestructive (NDT) technique for real-time monitoring in WAAM, as AE signals inherently carry physical information generated throughout the manufacturing process [11]. As a result, AE techniques have been extensively used in damage detection and remaining useful life prediction [12, 13]. At present, there are two principal categories of AE analysis: (i) conventional parameter-based methods based on physical modeling, and (ii) data-driven approaches utilizing machine learning algorithms [14].

Before analyzing the signals, the extraction of fault-related components from noisy measurements is performed. Zhang *et al* [15] proposed a sparse representation framework with a generalized logarithm nonconvex penalty to reconstruct repetitive transient impulses from heavily noise-contaminated bearing signals, achieving higher reconstruction accuracy than traditional penalties. Qiu *et al* [16] formulated a joint sparse and low-rank optimization model on spectrogram matrices and developed a Moreau-envelope ADMM algorithm to separate weak fault features from strong background components. Jiao *et al* [17] constructed a time–frequency–based dictionary learning framework to reconstruct leakage-related acoustic components, thereby reducing the background interference and increasing the accuracy of leak localization. Collectively, these studies indicate that successfully extracting defect-related response components from noisy measurements substantially reduces the difficulty of subsequent feature construction and pattern recognition.

The conventional parameter-based approach is extensively used due to the physical interpretability of its parameters, which are directly related to material behavior and defect evolution. Subaşı *et al* [18] proved that variations in the laser power and scanning speed are closely associated with AE metrics, including amplitude, frequency distribution, and sound pressure level, demonstrating that AE characteristics can effectively convey the level of process stability and defect development. Gao *et al* [19] proposed a parameter-based fault diagnosing method for rolling bearings by combining empirical wavelet decomposition with correlated kurtosis (CK). AE signals were decomposed using the empirical wavelet decomposition method, and CK was employed to identify the frequency band that is most sensitive to different failure modes, after which corresponding features were extracted using the envelope demodulation. Cui *et al* [20] investigated fatigue crack development in steel structures by examining various AE components and their correlations. The results indicate that changes in amplitude, energy, and duration of AE signals are closely related to different stages of crack propagation. Yet there are several limitations concerning the parameter-based method. First, features of AE signals are sensitive to material, environmental, and structural layout changes, making it difficult to develop a unified set of parameters for different scenarios. Second, the informative features on damages that are vital in fault diagnosis may not be retained in the process of manual feature extraction and dimensionality reduction [14]. Third, AE signals often suffer from strong noise interference and weak inter-feature correlation [21], which is often the case in the WAAM environment due to the high-energy arc-induced noise. These issues increase the probability of false alarms and considerably limit the performance of traditional threshold-based detection methods that are based on the conventional AE parameters. Subsequently, the traditional parameter-driven methods cannot achieve high-precision fault detection in the WAAM process.

To address the challenges mentioned above and to improve diagnosis accuracy for industrial data streams, data-driven approaches based on machine learning have become increasingly popular in recent years [22]. Ju *et al* [23] utilized *k*-means and Gaussian mixture models to detect crack modes in AE signals. Liu *et al* [24] studied the damage evolution behavior of BFRP-strengthened concrete beam through AE coupled with unsupervised machine learning. Several parameters of AE signals were evaluated, and *k*-means clustering was used to stratify AE events into four clusters corresponding to various damage modes. Ji *et al* [25] proposed a multimodal structural health monitoring technique that combines the AE with ultrasonic guided wave to measure the stiffness of open-hole laminated composites, establishing a relationship between cumulative counts of AE and the change in the structural stiffness. Machine learning methods exhibit several prominent features. To begin with, the performance of machine learning methods tends to improve with the amount of available data. Machine learning models can further improve their diagnostic ability through parameter optimization to extract latent information that may not be easily extracted manually. Second, machine learning models have a certain

degree of generalizability and transferability and can be adapted to related areas through transfer learning. Despite these advantages, some limitations still exist. Most machine learning models are low in interpretability since most of them are black-box systems that extract features without physical explanation. Moreover, model selection and hyperparameter tuning are complicated tasks. The capability of different models varies considerably under different working conditions, resulting in inconsistent performance across different scenarios. Although it has been observed that multi-model feature fusion is shown to be effective in enhancing the classification accuracy, it also comes with the increase of computational expenses, mostly due to comprehensive tuning of many hyperparameters- including learning rates, batch sizes, and network depth. Lastly, similar to the traditional parameter-based method, the AE features extracted by machine learning frameworks are often vulnerable to interference by both operational and environmental noise produced by the WAAM process [21].

Unlike the individual scalar features, time-frequency representations provide a more detailed description of the dynamics of the signal, retaining more information that is crucial for fault diagnosis. Most of the existing methods use short-time Fourier transform (STFT) to create time-frequency spectrograms. In comparison to STFT, the Mel spectrogram highlights the low-frequency components in AE signals and adopts a perceptual frequency scale designed to match the perceptual properties of the human ear [26]. Some recent works have discussed how Mel spectrograms and Mel-frequency cepstral coefficients (MFCCs) can be used for AE-based damage detection. Ren *et al* [27] proposed a Mel spectrogram-CNN model for detecting the damage in filament-wound CFRP composites. Mel spectrograms were obtained by the Mel filter bank, which effectively preserves the low-frequency content of the signal and eliminates high-frequency noise. This spectrogram is then fed into the ResNet-50 network for classification. The results show that Mel spectrogram-based deep learning has the potential to identify different failure modes, including matrix cracking, debonding, and fiber fracture, with high accuracy. Yang *et al* [28] made use of MFCCs to study the fracture evolution of sandstone under varying loading conditions. AE signals were processed using FFT and Mel filter bank, highlighting low-frequency components and suppressing high-frequency noise. The first twelve MFCCs were extracted to capture key time–frequency and energy distribution features of the AE signals.

In WAAM processes, process-induced defects are frequently accompanied by low-frequency AE components that are sometimes perceptible to the human ear. Additionally, due to the sparse distribution of Mel filters in the high-frequency range, the Mel spectrogram suppresses unwanted high-frequency disturbances while preserving informative low-frequency content. Moreover, the Mel spectrogram provides a less complex representation compared with the STFT-based time frequency spectrogram, requiring fewer parameters in machine learning models, thus reducing its computational cost, which simplifies the fault diagnosis and feature learning process.

In this paper, a Mel spectrogram-based fault diagnosis method for WAAM is proposed. The method is validated on a WAAM experimental dataset and its performance and efficiency are proven to be superior to the conventional machine learning methods that use STFT-based features.

The organization of the paper is as follows. Section 2 presents a description of machine learning models used in the study. Section 3 describes the experimental configuration in detail. Section 4 presents a Mel spectrogram-based fault diagnosis method and assesses its performance through experimental validation. Section 5 concludes the paper with a summary of key findings and outlines directions for future work.

2. Data preparation

2.1. Experimental setup

The experimental equipment used consists of a fully customized cold metal transfer welding station, AE sensors, and a signal acquisition system, as shown in figure 1. The welding torch is synchronized to a three-axis positioning system to fabricate samples with high accuracy. The feedstock used in the fabrication process is AZ31 magnesium alloy with the chemical composition Mg-2.8 Al-0.56Zn-0.37Mn(wt). During the fabrication process, the scanning rate is 0.02 m min^{-1}, the constant voltage is 12.8 V, and the welding current is 146 A. Previous experiments have confirmed that these process parameters yield deposited samples with excellent mechanical properties.

During the WAAM operation, real-time AE signals are recorded using the DSSS-8A high-speed acquisition module with a sampling rate of 3 MHz. A broadband AE sensor (WSa, Physical Acoustics Inc.) is securely attached to the top surface of the steel substrate. This allows real-time recording of broadband AE signals during the entire deposition process. To enhance signal quality, a preamplifier is employed for amplification and noise reduction of the raw AE signals.

To study the influence of printing quality on AE signals, four coupons made of AZ31 magnesium alloy are fabricated under different shielding gas flow conditions. The fabricated specimens were composed of five deposited layers, with an overall length of 10 cm. The layers are deposited layer by layer, and the AE signals are recorded throughout the process. Samples of four different quality conditions were produced, i.e., normal, slight, medium, and poor, by adjusting the flow rate of the shielding gas, as shown in figure 2. All other process parameters are held constant to ensure internal homogeneity within each coupon while gas flow settings shifted from 10 l min^{-1} (normal), 25% reduction, 75% reduction, and 100% shut-off [21].

AE signals were continuously recorded throughout the production process. As each coupon consisted of five layers, the data was divided into five corresponding signal groups for each condition, as shown in figure 3. The raw AE waveforms showed periodic pulse structures with substantial noise interference induced by arc discharges. Due to the complex

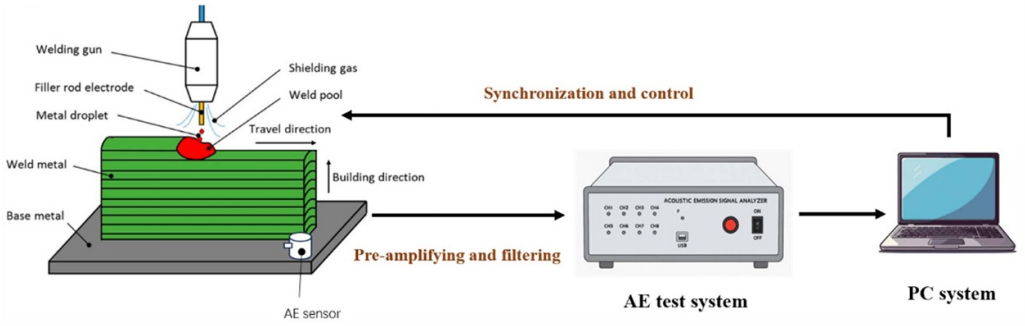

Figure 1. Experimental setup.

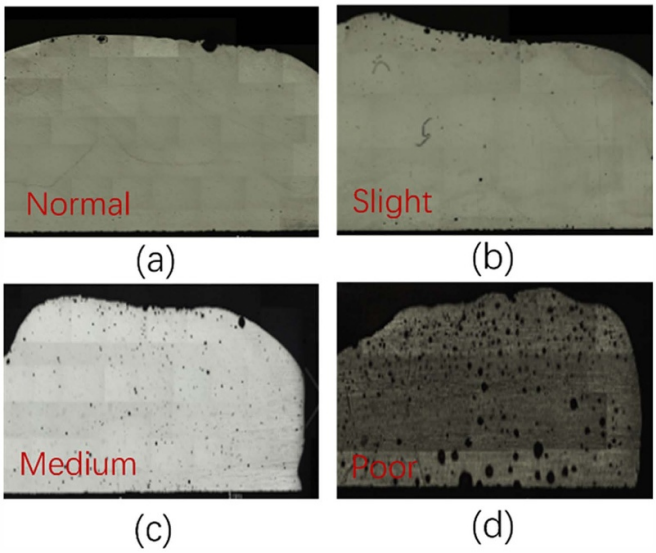

Figure 2. Coupons of different quality produced with gas flow settings of (a) 10 l min^{-1} (normal), (b) 25% reduction, (c) 75% reduction, and (d) 100% shut-off.

generation mechanism of AE signals and strong signal coupling, directly correlating coupon quality with time-domain waveforms or spectral features is challenging. Therefore, it is necessary to perform signal processing to extract defect-indicative information from raw AE signals.

2.2. Data preprocessing

Before feature extraction, preprocessing of the raw AE signals is necessary. First, a median filter with a window length of 500 is applied to suppress high-frequency noise. Since gaps exist between arc excitation events, many low-energy segments within the signal contain little or no useful information. Thus, a low-energy masking strategy is adopted to remove segments with persistently low energy over extended durations.

The signal envelope is then computed via the Hilbert transform, and a threshold of 0.008 is chosen based on manual inspection of the WAAM AE signals under normal conditions. This threshold is both higher than 75% the maximum value of the AE signal envelope and lower than most of the AE signal peaks under the normal condition, as shown in figure 4. The threshold serves as the minimum amplitude required for the detection of local peaks in the wave packet. Changes in gas flow would not affect the threshold, as lowering the gas flow would only increase the amplitude of AE event signals. Given that each excitation event lasts approximately 5–6 ms and the sampling frequency is 3 MHz, a local maxima detection strategy is used to locate effective peaks. During this process, a minimum peak height and a minimum distance between peaks are used to ensure the extracted peaks are both significant in amplitude and temporally independent, thus avoiding redundant detection due to noise or minor fluctuations.

To ensure each data sample contains exactly one excitation event in WAAM, a minimum inter-peak interval of 15 000 samples is defined. A signal segment of 15 000 points, centered at each detected peak and extracted using a rectangular window, is used to construct a single training sample. The overall process of the proposed method is shown in figure 5.

3. Mel spectrogram–driven deep learning framework with AE signal

For each data sample procured in section 2.2, time–frequency representations are computed: spectrograms are obtained via STFT, and Mel spectrograms are further derived by applying a Mel filter bank. A pretrained ResNet-50 model is employed to extract high-dimensional features and to compute inter-class distances between different gas flow settings. On this basis, convolutional neural network (CNN) and vision transformer (VIT) models are constructed to identify different defect states. Finally, the model performance is comprehensively evaluated in terms of accuracy, $F1$-score, confusion matrix, and runtime. The flowchart of the proposed method is shown as follows:

3.1. Mathematical framework of CNN

CNN is a widely used deep learning model that performs feature extraction by applying convolutional operations to input data, enabling effective feature representation [29]. By employing convolutional layers to extract features from input data, CNN is capable of learning signal patterns through optimization of convolutional kernel parameters. In this study, the CNN is mainly used to investigate the local feature extraction capability of Mel spectrogram representations.

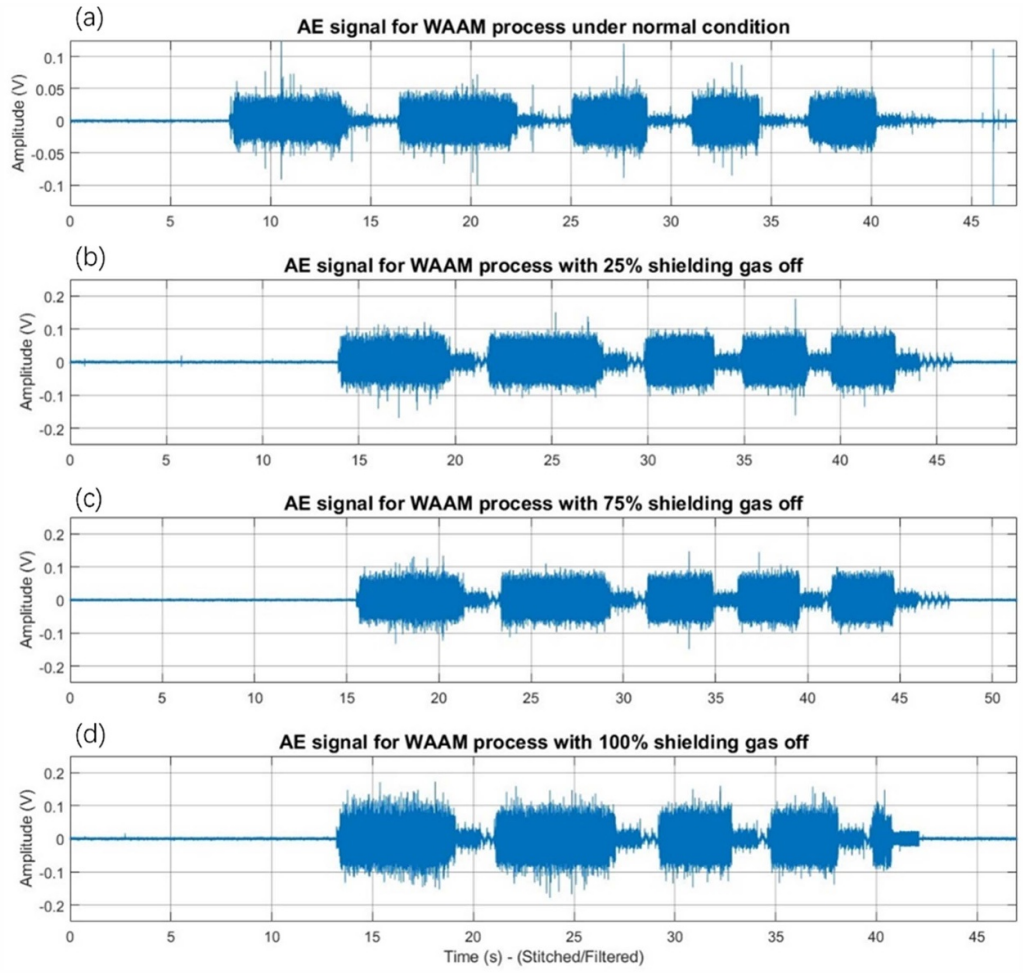

Figure 3. Full-length AE signal for WAAM process under shielding gas conditions of (a) 10 l min^{-1} (normal), (b) 25% reduction, (c) 75% reduction, and (d) 100% shut-off.

A typical CNN network consists of convolutional layers, pooling layers, and fully connected layers. The convolutional layers perform feature extraction by using kernels to collect information from local neighborhoods,

$$\mathbf{Z}^{(l)} = f\left(\mathbf{Z}^{(l-1)} * \mathbf{W}^{(l)} + \mathbf{b}^{(l)}\right) \tag{1}$$

where * denotes the convolution operation, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ represent the convolutional kernel weights and bias of the lth layer, and $f()$ is a nonlinear activation function (e.g., ReLU).

The output of the convolutional layers is then processed by pooling layers to reduce its size. The pooling layer performs down-sampling and feature selection, which reduces the computational complexity and enhances the translational invariance of the learned representations. Common down-sampling techniques include max pooling and average pooling, which retain either the strongest or the mean activation within a neighborhood [30],

$$\mathbf{Z}^{(l+1)} = \text{down}\left(\mathbf{Z}^{(l)}\right) \tag{2}$$

where $\text{down}(\cdot)$ denotes a downsampling function, commonly implemented using max pooling or average pooling.

Finally, the network uses fully connected layers to integrate features and perform classification, as shown in equation (3),

$$\mathbf{y} = f_{\text{fc}}\left(\mathbf{W}_{\text{fc}}\mathbf{x}_{\text{fc}} + \mathbf{b}_{\text{fc}}\right) \tag{3}$$

where $\mathbf{y}$ denotes the output vector, $\mathbf{W}_{\text{fc}}$ is the weight matrix of the fully connected layer, $\mathbf{b}_{\text{fc}}$ is the corresponding bias term, f_{fc} represents the activation function, and $\mathbf{x}_{\text{fc}}$ is the input vector.

Owing to its powerful hierarchical feature learning capability and computational efficiency, CNN has been extensively used in areas including image recognition, speech processing, and natural language analysis.

In this study, a lightweight CNN architecture is employed for signal feature extraction and classification. The model consists of three convolution–pooling modules followed by two fully connected layers, as shown in figure 6. Its compact structure and low computational complexity make it suitable for small- to medium-scale datasets.

The network is designed to process single-channel input signal maps of size $1 \times 28 \times 30$. The first convolutional layer uses 32 3×3 kernels to produce an output feature map with the shape of $[B, 32, 28, 30]$, where B denotes the batch size. A 2×2 max pooling operation is then applied, reducing the

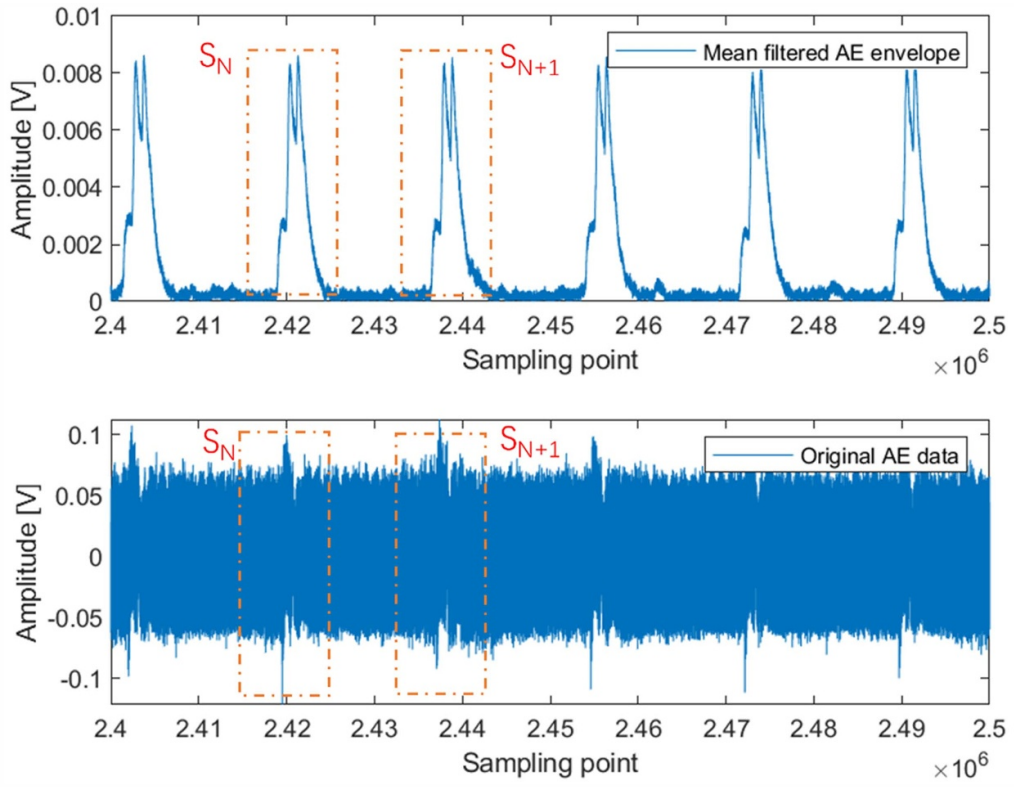

Figure 4. AE signal segmentation under the normal condition.

spatial resolution to $[B, 32, 14, 15]$. The second convolutional layer makes use of 64 kernels of the same size, followed by another 2×2 max pooling layer, yielding a feature map of $[B, 64, 7, 7]$. The third convolutional layer further increases the channel dimension to 128, and after the final pooling operation, the output has a shape of $[B, 128, 3, 3]$. This output is then flattened into a vector of size 1152 and passed through a fully connected layer with 256 units. The final result is presented by an output layer consisting of 4 neurons corresponding to the four-class classification task.

3.2. Mathematical framework of VIT

Compared with CNN, which focuses on local features, ViT emphasize global representations by dividing the input image into fixed-size patches [31]. In this study, it is used to evaluate the global feature modeling capability of the Mel spectrogram representation.

In the VIT model, each patch is reshaped into a one-dimensional vector and mapped into a D -dimensional embedding space via linear projection, where D denotes the embedding size. The patch embeddings are defined as:

$$\mathbf{z}_p = \mathbf{W}_E \text{Flatten}\,(\mathbf{x}_p) + \mathbf{b}_E, \quad p = 1, 2, ..., N, \tag{4}$$

where $\mathbf{x}_p$ is the pth image patch, $\mathbf{W}_E \in \mathbb{R}^{D \times (P^2 C)}$ is a learnable linear projection matrix, $\mathbf{b}_E$ is the bias term, and N is the total number of patches.

Two learnable components $\mathbf{z}_{\text{cls}}$ and $\mathbf{E}_{\text{pos}}$ are added to the patch embeddings. $\mathbf{z}_{\text{cls}}$ is a token for classification and $\mathbf{E}_{\text{pos}}$ is a positional encoding for spatial information. The input to the Transformer encoder is thus formulated as:

$$\mathbf{Z}_0 = [\mathbf{z}_{\text{cls}}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] + \mathbf{E}_{\text{pos}}. \tag{5}$$

The input sequence is then fed into a stack of transformer encoder layers, each consisting of multi-head self-attention (MSA), position-wise feedforward networks (FFN), residual connections, and layer normalization. The mechanism of the MS layer is shown in equation (6),

$$\text{Attention}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{6}$$

Attention weights are derived from the dot product the dot product $\mathbf{Q}\mathbf{K}^\top$, which signifies the similarity between the query $\mathbf{Q}$ and key $\mathbf{K}$. This result is normalized by $\sqrt{d_k}$, where d_k is the key dimension, and transformed into attention coefficients via the softmax function. $\mathbf{V}$ is the linear projection of the original input into the value space. The final attention output is the weighted sum of the value vector $\mathbf{V}$ [32].

Finally, classification is performed based on the encoded $\mathbf{z}_{\text{cls}}$ token, which is passed through a multilayer perceptron (MLP) head.

In this study, visual features are extracted using the ViT-B/16 model, which is based on the original framework proposed by Dosovitskiy *et al* [31] and implemented via the PyTorch torchvision library. The designation 'B/16' refers to the base-scale model with an input patch size of 16×16.

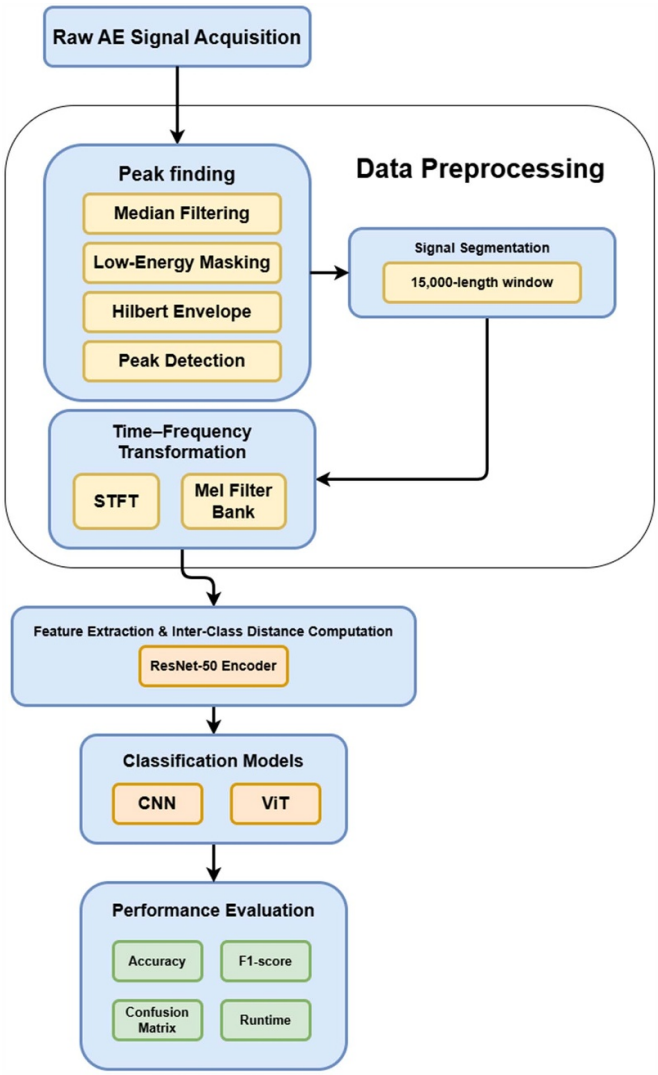

Figure 5. Flowchart of the proposed method.

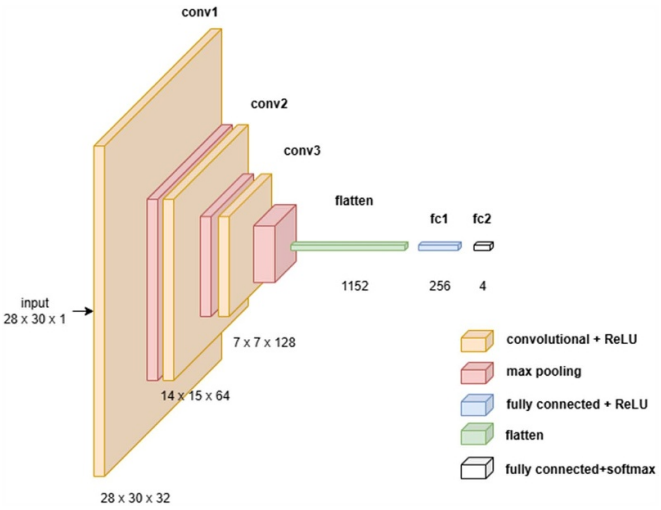

Figure 6. Architecture of the Mel CNN model.

The architecture is illustrated in figure 7. The input image of size $3 \times 224 \times 224$ is partitioned into 196 non-overlapping patches. Each patch is flattened and linearly projected into a 768-dimensional embedding space. A learnable CLS token and positional encoding are then added to the embedding. 12 transformer encoder layers, each consisting of MSA, a FFN with hidden size 3072, residual connections, and layer normalization are used to process the sequence. The final output is fed into a MLP head for classification. The model is fully attention-based, making it able to model global context and improve generalization performance.

3.3. STFT and Mel spectrogram

The Mel spectrogram is a time–frequency representation method based on the auditory perception properties of the human ear. The way humans perceive frequency is nonlinear, with lower frequency parts being more resolved than higher frequency parts. This is also the principle of the Mel spectrogram; it naturally has a higher resolution for low-frequency components [26].

To derive Mel spectrogram from AE signals, the raw time-domain signal is first preprocessed and divided into overlapping frames (framing). Each frame is windowed to reduce spectral leakage. The spectrogram is then obtained by undergoing STFT:

$$S(t,f) = \left| \int_{-\infty}^{\infty} x(\tau)\, w(\tau - t)\, \mathrm{e}^{-\mathrm{j}2\pi f\tau} \mathrm{d}\tau \right|^2 \tag{7}$$

where x is the time domain signal, w is the window function.

The STFT spectrogram is then processed by a Mel filter bank, which is a collection of band-pass triangular filters spaced on the Mel scale. Each filter emphasizes energy in a specific frequency band while attenuating others, and the filtered energy is aggregated to produce the Mel spectrogram:

$$M(t,m) = \int_{0}^{\infty} S(t,f)\, H_m(f)\, \mathrm{d}f \tag{8}$$

where $H_m(f)$ denotes the frequency response of the mth Mel filter.

Compared to the original STFT spectrogram, the Mel spectrogram compresses the frequency dimension by reducing the number of frequency bins, thereby lowering data dimensionality and enhancing computational efficiency. More importantly, it emphasizes frequency features in the low-frequency range. And suppresses high-frequency noise, which enhances its performance in signal classification and condition monitoring tasks. An example of the advantages of the Mel spectrogram is shown in figure 8. When the signal is limited to the low-frequency range, a large part of the STFT spectrogram consists of noise. Whereas in the Mel spectrogram, defect features are expanded, and the high-frequency part without useful information is compressed.

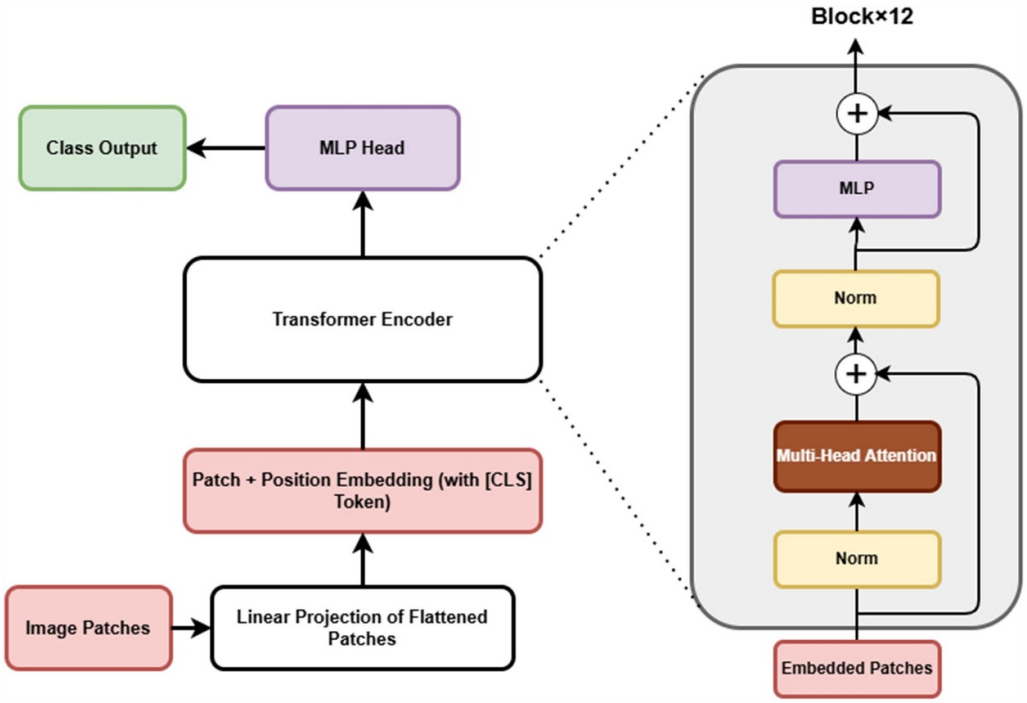

Figure 7. Architecture of ViT-B/16 model.

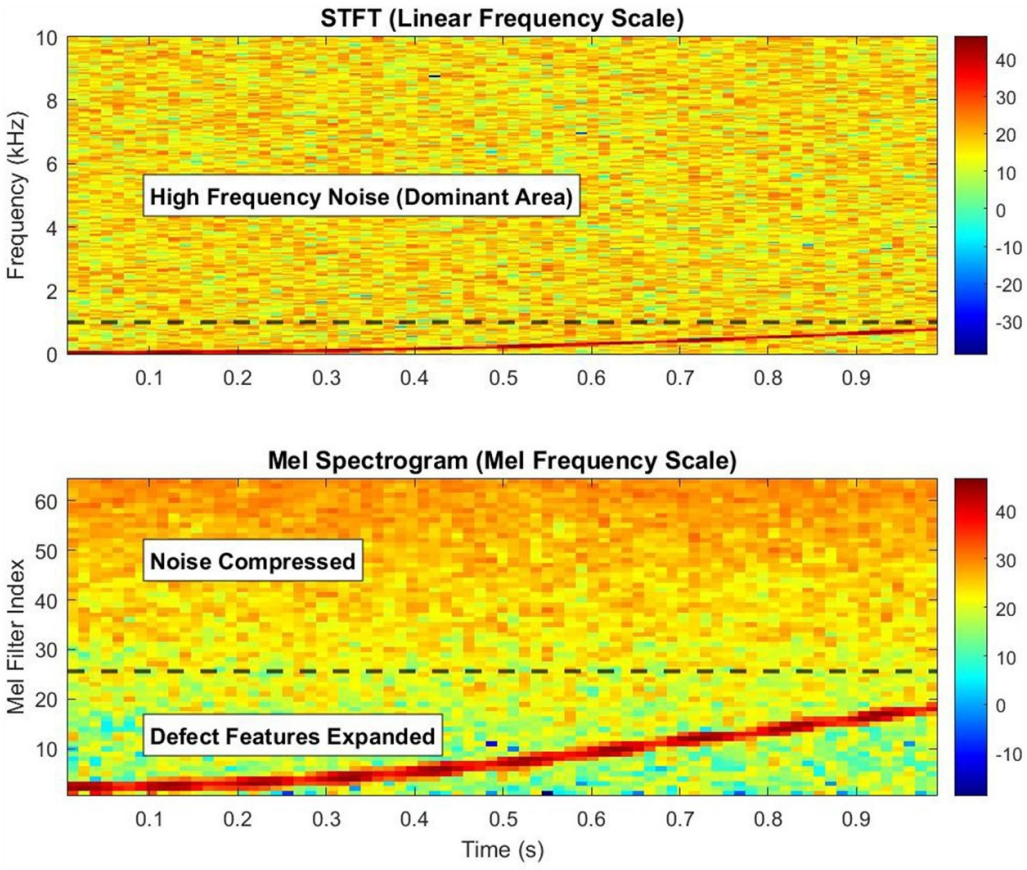

Figure 8. Comparison of STFT spectrogram and Mel spectrogram.

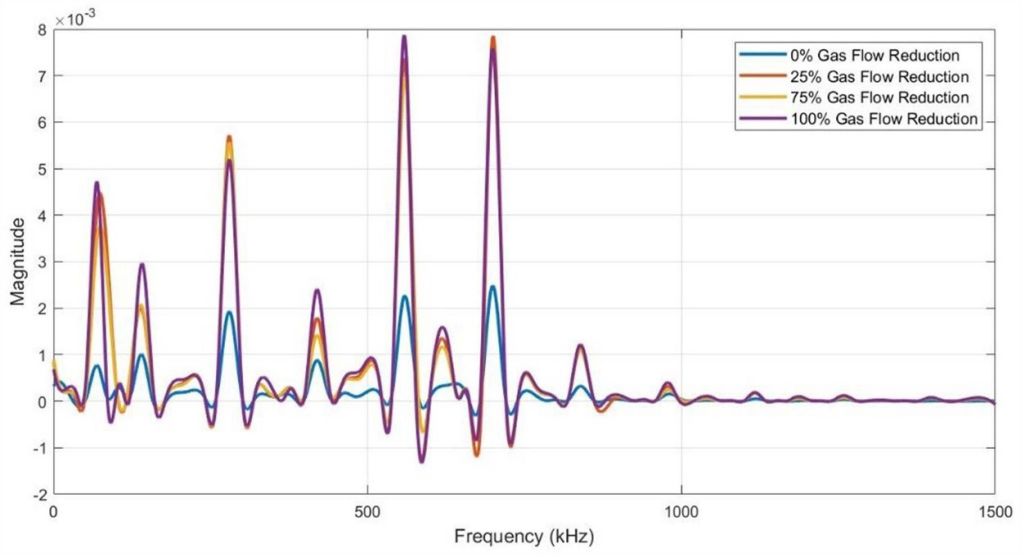

Figure 9. Spectral analysis of different conditions.

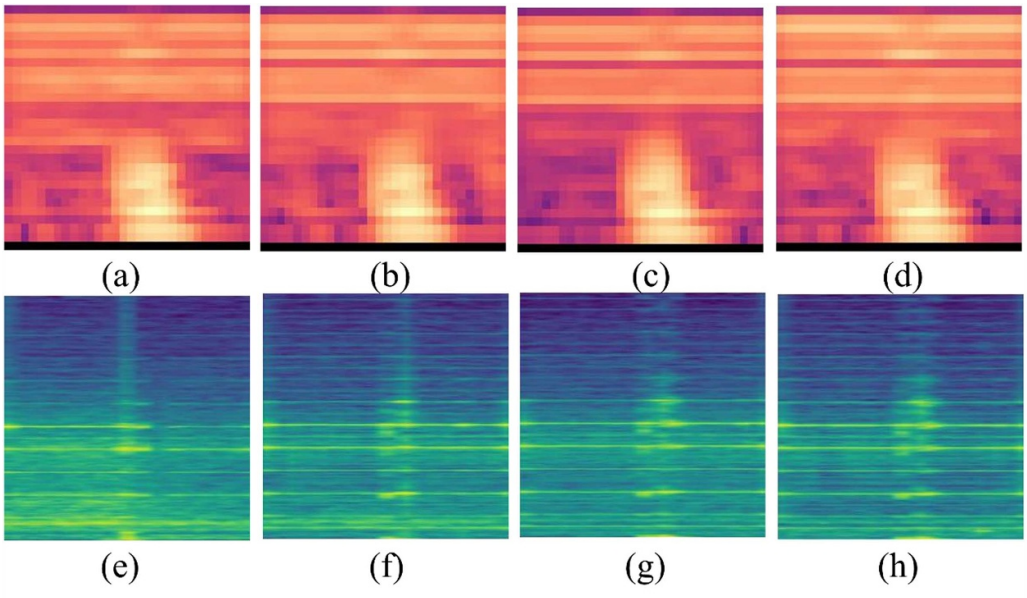

Figure 10. Mel spectrogram under different (a) normal, (b) 25% reduction, (c) 75% reduction, (d) 100% shut-off gas flow settings; STFT spectrogram under different (e) normal, (f) 25% reduction, (g) 75% reduction, (h) 100% shut-off gas flow settings.

4. Results and discussion

Spectral analysis is performed under each different condition, and the results are shown in figure 9. It can be seen that apart from the normal condition, differences between other conditions mainly lie in 0–500 kHz, which is the low-frequency range when considering a sampling rate of 3 MHz. Therefore, the Mel scale is suitable for the classification task.

The librosa library is used to perform STFT on each segmented waveform. A Hanning window is applied to reduce spectral leakage, and the time–frequency representation is obtained. Given the high sampling frequency, the number of FFT points is set to 2048, with a hop length of 512. Based on the STFT output, the Mel spectrogram is computed using 28 Mel filters. These filters perform a nonlinear mapping from the linear frequency axis to the Mel scale axis. This transformation increases frequency resolution in the low-frequency range and suppresses noise in the high-frequency range, leading to a more compact spectral representation and improved feature robustness. An example of the STFT spectrogram and Mel spectrogram is shown in figure 10.

To assess the sensitivity of different time–frequency representations to process variations, normal-condition samples (i.e., gas flow reduction = 0) of STFT and Mel spectrograms were respectively averaged to create a baseline. The KL divergence was calculated between each spectrogram and its class-mean

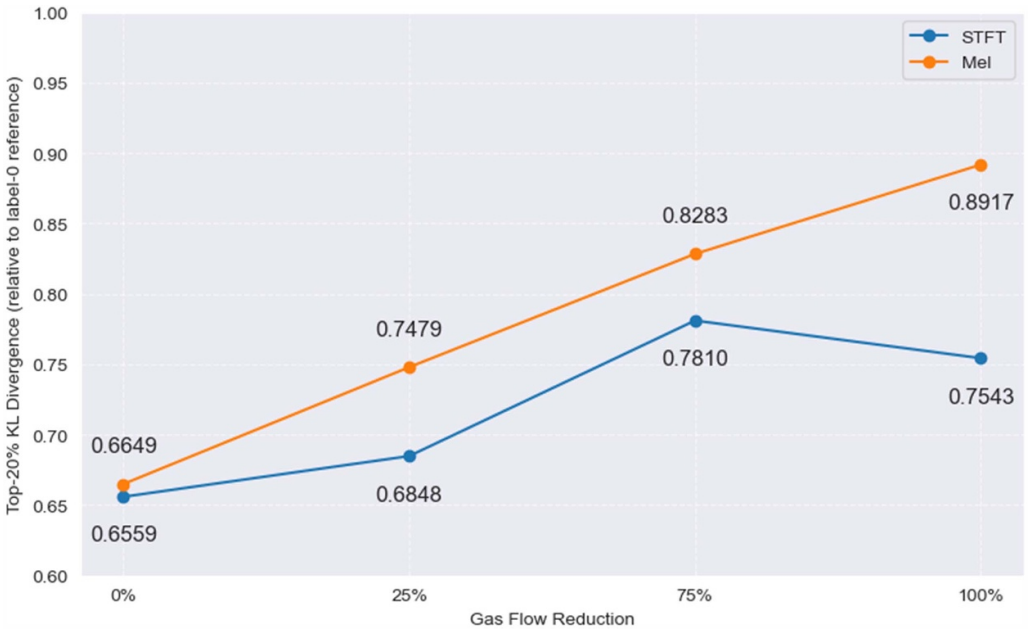

Figure 11. KL divergence of Mel and STFT spectrograms.

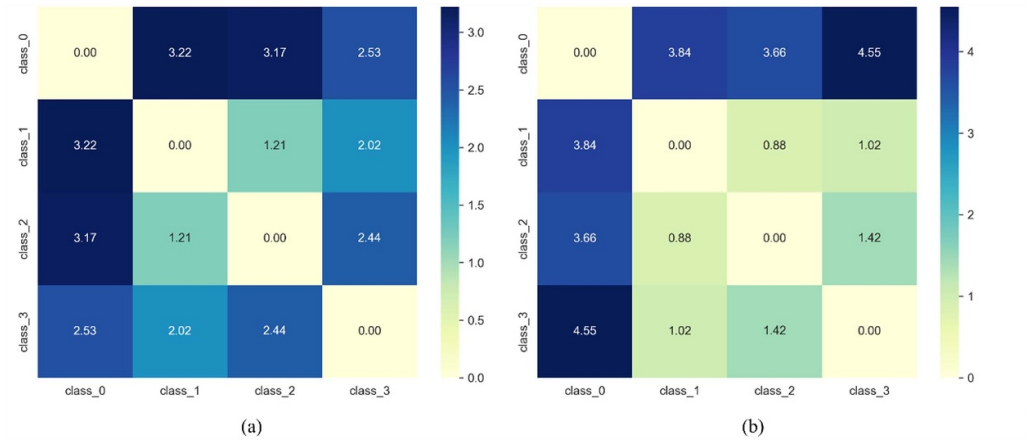

Figure 12. Inter-class distance matrix (a) Mel (b) STFT.

baseline by considering only the top 20% energy components, thereby excluding low-energy regions that are highly susceptible to WAAM noise. This operation suppresses WAAM-related noise and emphasizes systematic changes induced by process degradation. The results in figure 11 show that the KL divergence computed from the Mel spectrogram increases monotonically and with a relatively stable progression as the gas-flow reduces. In contrast, the STFT-based KL values exhibit a non-monotonic and less regular pattern. This behavior indicates that the Mel representation provides a more reliable characterization of the underlying energy distribution and offers enhanced sensitivity to process-related changes.

4.1. Inter-class distance computation

After preprocessing, inter-class distances were evaluated for Mel spectrograms and STFT spectrograms using a ResNet-50–based feature encoder. A pretrained ResNet-50 was loaded with its fully connected layer removed to retain only the feature extractor. All images were resized to 224 $\times$ 224, and a 512-dimensional feature vector was obtained for each sample. Class centroids were then computed, and their pairwise Euclidean distances formed the matrix shown in figure 12, where class 0 to class 3 represent shielding gas flow from normal to 100% reduction. The results show that Mel-based features yield larger inter-class separations—especially when the shielding gas flow is abnormal—whereas STFT-based features exhibit weaker separability.

4.2. CNN comparison

The CNN architectures for Mel spectrograms and STFT spectrograms share the same structure and functionality. For both models, the batch size is 32, and the number of epochs is 25; Adam is used as the optimizer with cross-entropy loss, and the final classification head is a four-dimensional MLP. The only

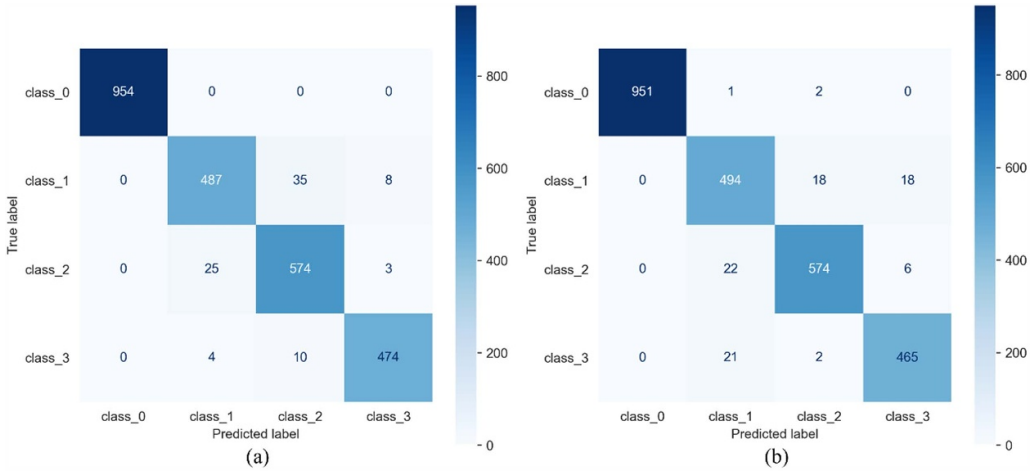

Figure 13. CNN confusion matrix (a) STFT (b) Mel.

Table 1. Classification results of CNN for STFT spectrogram.

Class	Precision	Recall	$F1$-score
Class_0	1	1	1
Class_1	0.9438	0.9189	0.9312
Class_2	0.9273	0.9535	0.9402
Class_3	0.9773	0.9713	0.9743
Macro avg	0.9621	0.9609	0.9614
Weighted avg	0.9671	0.967	0.967
Accuracy		**0.967**	
Run time (s)		**521**	

Table 2. Classification results of CNN for Mel spectrogram.

Class	Precision	Recall	$F1$-score
Class_0	1	0.9969	0.9984
Class_1	0.9182	0.9321	0.9251
Class_2	0.9631	0.9535	0.9583
Class_3	0.9509	0.9529	0.9519
Macro avg	0.9581	0.9588	0.9584
Weighted avg	0.9652	0.965	0.9651
Accuracy		**0.965**	
Run time (s)		**212**	

difference lies in the input tensor shape: the STFT network receives a tensor with the shape of [32, 1025, 30], whereas the Mel network receives a tensor with the shape of [28, 30, 32]. The classification results and confusion matrices are shown in figure 13, tables 1, and 2. Accuracy and run time are bolded in tables 1–4 to highlight the primary performance metrics.

Both models achieve high accuracy (up to 0.96). Of the four classes, the STFT network has a relatively poor classification ability for class 1, and the Mel network has a relatively poor classification ability for class 3. It must be noted that Mel network trains significantly faster due to its lower input dimensionality: the training time is 2 min 14 s for the Mel model versus 7 min 49 s for the STFT model. In the preprocessing phase, generating Mel spectrograms takes 1 min 18 s, whereas computing STFT takes 52 s. In terms of end-to-end runtime, which includes preprocessing and training, the Mel-spectrogram approach is markedly faster than the STFT approach, while still providing an effective local feature representation for WAAM fault diagnosis.

Table 3. Classification results of VIT for STFT spectrogram.

Class	Precision	Recall	$F1$-score
Class_0	0.97	0.94	0.95
Class_1	0.58	0.71	0.64
Class_2	0.78	0.46	0.58
Class_3	0.67	0.89	0.76
Macro avg	0.75	0.75	0.73
Weighted avg	0.79	0.77	0.76
Accuracy		**0.77**	

4.3. VIT comparison

This study employs the ViT-B/16 model from torchvision as the visual encoder. Because the available WAAM dataset is relatively small and insufficient for training a deep model from scratch, pretrained weights are adopted to provide a better initialization and improve model convergence. Mel and STFT spectrograms are resized to 224 × 224 and fed into the model. Training uses the Adam optimizer with cross-entropy loss, an initial learning rate of 1×10^{-3}, and head dropout of 0.2. The original classification head is replaced with a four-class head for the present task. During the first stage, only this head is trained while the remaining layers are frozen. The best checkpoint is retained for subsequent use.

In the second stage, all layers are unfrozen and fine-tuned with a learning rate of 2×10^{-5}, which is subjected to a scheduled decay. The two-stage procedure yields the classification results shown in tables 3 and 4. Classification accuracy with STFT inputs is markedly lower than with Mel inputs, which is consistent with section 4.1, where Mel representations exhibited greater inter-class separability. The results prove that the Mel spectrogram provides more informative and

Table 4. Classification results of VIT for Mel spectrogram.

Class	Precision	Recall	$F1$-score
Class_0	1	1	1
Class_1	0.81	0.83	0.82
Class_2	0.87	0.88	0.87
Class_3	0.92	0.9	0.91
Macro avg	0.9	0.9	0.9
Weighted avg	0.92	0.92	0.92
Accuracy		**0.92**	

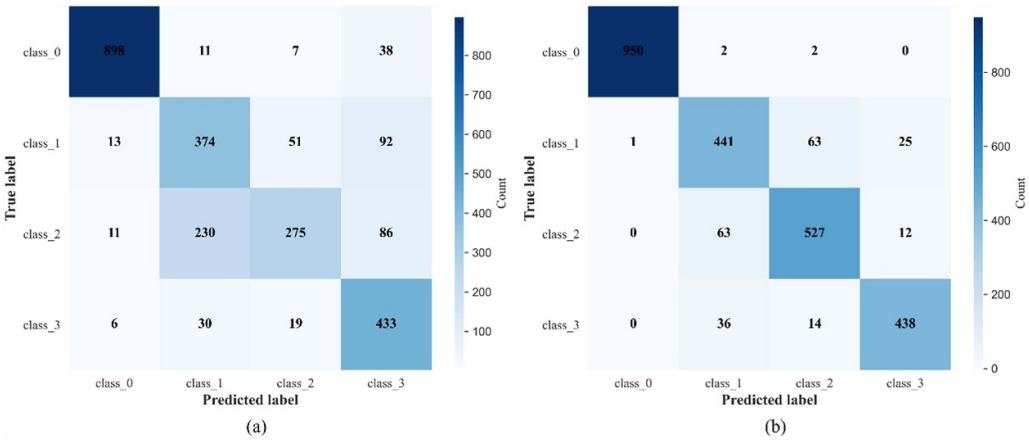

Figure 14. VIT confusion matrix (a) STFT (b) Mel.

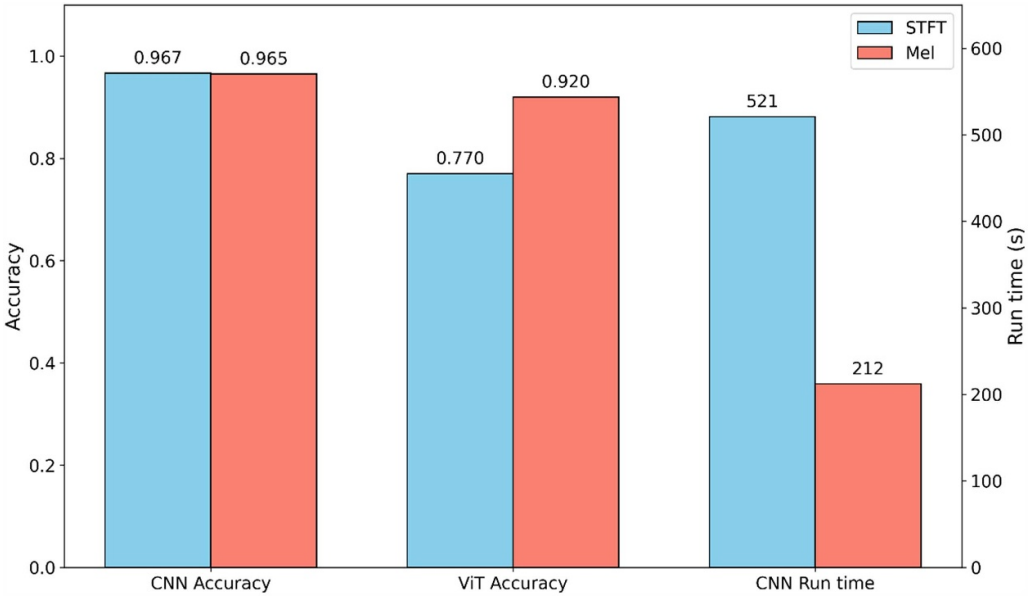

Figure 15. Overall comparison of Mel and STFT spectrogram.

discriminative global features for WAAM fault identification. The run time for Mel and STFT representations is similar as they are both resized to 224 × 224.

Overall, the ViT model exhibits degraded performance compared to the CNN in this task; this is mainly due to the limited availability of WAAM data. The low number of training samples restricts the model's ability to fine-tune deep attention layers effectively, causing most pretrained weights to remain close to their original values from the torchvision initialization. As a result, the transformer fails to achieve full task-specific adaptation, as shown in its final classification accuracy, which is only 0.92. The confusion matrix of the VIT model is shown in figure 14. Overall results of CNN and VIT for Mel and STFT spectrograms are shown in figure 15.

5. Conclusions

This work presents a health monitoring framework for WAAM based on Mel spectrogram representations. During data preprocessing, average filtering and Hilbert envelope peak searching are employed to enhance feature characterization and highlight AE events within the full-length recorded signals. By constructing time–frequency representations, most of the information generated during the WAAM process is captured, thereby minimizing information loss. Considering the high-energy, high-frequency noise and the low signal-to-noise ratio inherent in WAAM, the Mel spectrogram is adopted for feature extraction due to its ability to provide finer resolution and better discriminability in the low-frequency range. Experimental results indicate that the method greatly lowers the number of network parameters and computational complexity. It also increases the processing speed and separability between features of various defect modes, leading to better fault-classification accuracy and an increase in reliability for AE monitoring.

Adaptive Mel filter design and time-frequency feature fusion will be considered in future work to enhance performance. In addition, real-time implementation and transfer learning of various materials and working conditions in the WAAM process will also be studied to improve the generalization and industrial applicability of the proposed framework. Transfer learning can also be applied to the model for the framework to adapt to more challenging arc modes, such as spray transfer, in future work. Pretraining can be performed on the current dataset to learn general features of various defects, and fine-tuning can help the framework adapt to noises induced by different arc modes.

Acknowledgments

This work is supported by the Foundation of Science and Technology on Reliability and Environmental Engineering Laboratory (Grant No. WDZC6142004240401), which is highly appreciated by the authors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Gain S and Veeman D 2025 A review on advances and challenges in wire arc additive manufacturing: process parameters, microstructural evolution and material performance across alloys *J. Alloys Compd.* **1029** 180735

[2] Shah A, Aliyev R, Zeidler H and Krinke S 2023 A review of the recent developments and challenges in wire arc additive manufacturing (WAAM) process *J. Manuf. Mater. Process.* **7** 97

[3] Ding C, Wang J, Zhang S, Yang S and Ma Y 2025 A novel active learning stochastic Kriging metamodel for improving reliability and stability of additive manufacturing processes *Reliab. Eng. Syst. Saf.* **260** 111043

[4] Wu B, Pan Z, Ding D, Cuiuri D, Li H, Xu J and Norrish J 2018 A review of the wire arc additive manufacturing of metals: properties, defects and quality improvement *J. Manuf. Process.* **35** 127–39

[5] Thompson A, Maskery I and Leach R 2016 X-ray computed tomography for additive manufacturing: a review *Meas. Sci. Technol.* **27** 072001

[6] Bhat G A, Smagulova D and Jasiunienė E 2026 Optimization of nondestructive evaluation techniques for bonded components through model-assisted POD analysis *IEEE Trans. Reliab.* **75** 555–69

[7] Choudhary A, Goyal D and Letha S S 2021 Infrared thermography-based fault diagnosis of induction motor bearings using machine learning *IEEE Sens. J.* **21** 1727–34

[8] Mulaveesala R, Arora V and Dua G 2021 Pulse compression favorable thermal wave imaging techniques for non-destructive testing and evaluation of materials *IEEE Sens. J.* **21** 12789–97

[9] Vilar N, Artigas R, Bermudez C, Thompson A, Newton L, Leach R, Duocastella M and Carles G 2022 Optical system for the measurement of the surface topography of additively manufactured parts *Meas. Sci. Technol.* **33** 104001

[10] Li C, Guo H, Wei X, He W, Nie X, Zhang T and Fang Y 2025 SaimVAE: an unlabeled intelligent ultrasonic NDT method for composite materials *Measurement* **249** 117059

[11] Wei W, Liu Y, Wu J, Wei Z, Zhou Z and Long Y 2025 In-situ monitoring method of femtosecond laser welding between glass and copper with acoustic emission *Measurement* **240** 115568

[12] He J, Yang F, Wang H, Sun X, Zhu Y, Wang Y and Guan X 2025 A physics-based acoustic emission energy method for mixed-mode impact damage prediction of composite laminates *Ultrasonics* **145** 107490

[13] Wang X, Yue Q and Liu X 2024 Reliable arrival time picking of acoustic emission using ensemble machine learning models *Mech. Syst. Signal Process.* **215** 111442

[14] Zhao S, Li G and Wang C 2024 Bridge cable damage identification based on acoustic emission technology: a comprehensive review *Measurement* **237** 115195

[15] Zhang Z, Huang W, Liao Y, Song Z, Shi J, Jiang X, Shen C and Zhu Z 2022 Bearing fault diagnosis via generalized logarithm sparse regularization *Mech. Syst. Signal Process.* **167** 108576

[16] Qiu T, Huang W, Zhang Z, Wang J and Zhu Z 2024 A new approach for sparse optimization with Moreau envelope to extract bearing fault feature *Mech. Syst. Signal Process.* **216** 111493

[17] Jiao J, Zhang J, Ren Y, Li G, Wu B and He C 2023 Sparse representation of acoustic emission signals and its application in pipeline leak location *Measurement* **216** 112899

[18] Subasi L, Oren S, Dursun G, Sen C and Orhangul A 2024 In-situ acoustic signals correlation of process parameters in laser powder bed fusion *Proc. CIRP.* **126** 668–73

[19] Gao Z, Lin J, Wang X and Xu X 2017 Bearing fault detection based on empirical wavelet transform and correlated kurtosis by acoustic emission *Materials* **10** 571

[20] Cui J, Qu X, Lv C, Du J and Wang H 2025 Multi-parameter acoustic emission analysis for fatigue crack evaluation in structural health monitoring *Measurement* **256** 118529

[21] Gao F, Li C, Ji D, Hua J and Lin J 2025 Spectral characterization method for wire-arc additive manufacturing monitoring with broadband AE signals *Meas. Sci. Technol.* **36** 035002

[22] Zhuang J, Wu J, Pei G, Li W, Xin G, Ma C, Feng K, Zhang L and Yang C 2025 A neuro-fuzzy approach with hypergraph

convolution for fault diagnosis in industrial devices *J. Reliab. Sci. Eng.* **1** 035301
[23] Ju S, Li D and Jia J 2022 Machine-learning-based methods for crack classification using acoustic emission technique *Mech. Syst. Signal Process.* **178** 109253
[24] Liu K, Wulan T, Yao Y, Bian M and Bao Y 2024 Assessment of damage evolution of concrete beams strengthened with BFRP sheets with acoustic emission and unsupervised machine learning *Eng. Struct.* **300** 117228
[25] Ji D, Li W, Liu Z, Liu H, Gao F, Wang M, Chang B and Lin J 2025 Quantitative stiffness evaluation and damage mechanism investigation in open-hole composites via multi-modal SHM data fusion *Mech. Syst. Signal Process.* **237** 112998
[26] Abdul Z K and Al-Talabani A K 2022 Mel frequency cepstral coefficient and its applications: a review *IEEE Access* **10** 122136–58
[27] Ren X, Wang J, Liang Y, Ma L and Zhou W 2024 Acoustic emission detection of filament wound CFRP composite structure damage based on Mel spectrogram and deep learning *Thin-Walled Struct.* **198** 111683
[28] Yang F, Li Z, Liu S, He X, Song D, Li N, Wang H and Sobolev A 2025 Characterizing the fracture evolution of sandstone by using Mel-frequency cepstral coefficients of acoustic emission *Eng. Fract. Mech.* **324** 111254
[29] Alzubaidi L, Zhang J, Humaidi A J, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel M A, Al-Amidie M and Farhan L 2021 Review of deep learning: concepts, CNN architectures, challenges, applications, future directions *J. Big Data* **8** 53
[30] Hamed Alizadeh Moghaddam S, Gazor S, Homayouni S and Karami F 2025 Integrating local and global features in a convolutional vision transformer for wetland mapping *IEEE Sens. J.* **25** 8674–83
[31] Dosovitskiy A *et al* 2021 An image is worth 16x16 words: transformers for image recognition at scale (arXiv:2010.11929)
[32] Qu L, Li X, Yang T and Wang S 2025 Radar-based continuous human activity recognition using multidomain fusion vision transformer *IEEE Sens. J.* **25** 9946–56

Fei Gao received BS and PhD degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2013 and 2018, respectively. He is currently an Associate Professor with the School of Reliability and Systems Engineering, Beihang University, Beijing, China. His research interests include structural health monitoring (SHM), nondestructive testing (NDT), and guided wave propagation.

Yishu Chen received a BS degree in reliability and safety engineering from Beihang University, Beijing, China, in 2022. He is currently pursuing a PhD degree in control science and engineering at Beihang University, Beijing, China. His research interests include structural health monitoring(SHM), nondestructive testing (NDT), and digital twin for fault diagnosis.

Jing Lin (Senior Member, IEEE) received BS, MS and PhD degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 1993, 1996 and 1999, respectively. From 2009 to 2018, he was a Professor with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China. He is currently the Dean of the School of Reliability and Systems Engineering, Beihang University, Beijing, China.

Yinmin Zhu received a BS degree in safety engineering from Beihang University, Beijing, China, in 2024. He is currently pursuing a PhD degree in electronic information engineering with Beihang University, Beijing, China. His research interests include structural health monitoring(SHM), nondestructive testing (NDT), and intelligent signal processing technology.

Yonghao Miao received BS degree in mechanical engineering and automation from the Wuhan University of Technology, Wuhan, China, in 2013, and a PhD degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2018. He is currently an Associate Professor with the School of Reliability and Systems Engineering, Beihang University, Beijing, China.

Jinghui Tian received a BS degree in measurement and control technology and instrumentation from Henan Polytechnic University, Jiaozuo, China, in 2018, and a PhD degree in mechanical design and theory from Yanshan University, Qinhuangdao, China, in 2024. He was a Visiting PhD Student with the Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy, from 2022 to 2023. He is currently a Postdoctoral Research Fellow with the Ningbo Institute of Technology, Beihang University, Ningbo, China. His research interests include transfer learning, information fusion, machinery condition monitoring, and intelligent fault diagnosis.