FAT-Net: A Spectral-Attention Transformer Network for Industrial Audio Anomaly Detection Using MFCC and Raw Features

Yanhua Shi
School of Information Engineering, Zhengzhou University of Technology; Zhengzhou, Henan, 450053 China
Email: Yanhua_Shi@yeah.net

Keywords: audio noise anomaly detection, MFCC features, transformer model

Received: March 28, 2025

This paper proposes FAT-Net, an audio noise anomaly detection method that integrates big data with a Transformer-based architecture. The model combines Mel-Frequency Cepstral Coefficients (MFCCs) and raw audio features to capture both spectral and temporal characteristics. A novel Spectral Attention Mechanism (SAM) is introduced to enhance sensitivity to anomaly-relevant frequency bands. Experiments were conducted on a large industrial dataset comprising approximately 3,000 audio recordings collected under real manufacturing conditions. FAT-Net was evaluated using accuracy, precision, recall, and F1-score as metrics, achieving a best F1-score of 98.05%, outperforming baseline models such as CNN (90.31%), LSTM (89.04%), and MFCC+LSTM (97.04%). These results demonstrate the effectiveness and generalization capability of FAT-Net for deployment in industrial environments.

Povzetek: FAT-Net z arhitekturo LLM in spektralno pozornostjo združuje MFCC in surove zvočne značilnosti ter s tem omogoča kvalitetno detekcijo akustičnih anomalij v industrijskem hrupu.

1 Introduction

As modern manufacturing systems continue to evolve, factory equipment has become increasingly interconnected. In such environments, accurate monitoring of equipment operation and timely identification of anomalies have become critical for ensuring efficient, and uninterrupted production. Among various monitoring modalities, acoustic signals have emerged as a powerful, non-invasive, and real-time information source. These signals inherently encode rich information about the mechanical state of equipment, ambient environmental changes, and latent fault signatures, including early warnings of failures or abnormal operations. Consequently, the development of robust and intelligent audio-based anomaly detection systems is of paramount importance to enhance equipment reliability and production line stability [1].

However, practical deployment of audio anomaly detection systems faces challenges. As Folz [2] found in their comprehensive investigation of electric motor anomaly detection, variations in equipment types, operating conditions, background noise, and mechanical wear can lead to substantial variability in the captured audio signals, making anomaly identification uncertain. Traditional approaches to acoustic anomaly detection have primarily relied on statistical analysis and signal processing techniques. For instance, Lopes et al. [3] extract time-frequency features from acoustic emissions, then combined them with statistical hypothesis testing for grinding wheel condition monitoring. Although these methods have the advantages of low computational cost and strong interpretability, they have limited feature representation capabilities and are highly sensitive to external noise and signal perturbations.

The evolution of machine learning has introduced various statistical learning-based models. Coelho et al. [4] proposed a deep autoencoder and Support Vector Machine (SVM)-based approach to detect deviations by learning the distribution of normal audio patterns in working machines. Similarly, Wang et al. [5] employed Gaussian Mixture Model (GMM) to probabilistically model features for anomaly detection, achieving improvements in recognition accuracy. While these methods improve pattern recognition capacity, and allow for more flexible modeling of feature space, they still rely on handcrafted features, are sensitive to data imbalance, and often generalize poorly across domains.

In recent years, deep learning models based on Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) have been widely used in acoustic anomaly detection. Anidjar et al. [6] developed a CNN-based approach that can learns spectral and temporal features, leading to improved performance in industrial sound classification tasks. Ullah et al. [7] designed a framework using LSTM networks to handle extended temporal sequences in audio signals. Building on this progress, Zhang et al. [8] explored the use of the Transformer model for acoustic anomaly detection, leveraging its self-attention mechanism to enhance feature representation. However, current deep learning methods face challenges such as high computational complexity and limited ability to capture global contextual information, where global contextual information is critical for identifying dispersed anomalies—defined as subtle, non-contiguous, and temporally scattered acoustic deviations that do not manifest as a single continuous fault signature but rather as intermittent irregularities across the signal timeline. Figure 1 summarizes the performance of existing methods compared to the proposed FAT-Net across key evaluation metrics: accuracy, precision, recall, and F1-score. As illustrated, conventional CNN and LSTM architectures individually lack the capability for combined spectral-temporal modeling, limiting their ability to capture audio patterns. Although MFCC+CNN and MFCC+LSTM approaches achieve reasonable performance by incorporating spectral features, they do not employ adaptive spectral weighting mechanisms, thereby constraining their sensitivity to subtle anomaly-relevant frequency components.

To overcome these challenges, we design an audio anomaly detection method that leverages the Transformer architecture to identify anomalies from large-scale industrial audio data, named Feature-Augmented Transformer Network (FAT-Net). The primary objective of this research is to improve anomaly classification performance in noisy industrial audio environments through spectral-temporal feature fusion and adaptive attention mechanisms. Our approach combines Mel-Frequency Cepstral Coefficients (MFCCs) with raw audio features to form a more discriminative audio feature space. In addition, a Spectral Attention Mechanism (SAM) is introduced to enable the model to adaptively focus on frequency bands that carry strong anomaly-related information. By capturing long-range dependencies across the audio sequence, the Transformer effectively compensates for the limitations of CNN and LSTM models, which often struggle with short-term or local representations. Experimental results demonstrate that FAT-Net achieves a performance improvement, increasing the F1-score by approximately 1.01% compared to the MFCC+LSTM baseline and outperforming all other existing methods across all evaluated metrics.

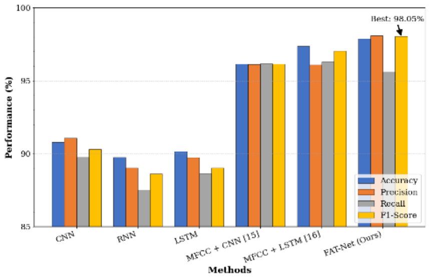

Figure 1: Comparison of different methods

2 Method

This section presents the FAT-Net , which based on big data and an improved Transformer architecture. The method consists of three main components: (1) feature extraction module using MFCCs, (2) an enhanced Transformer network with a SAM module, and (3) a feature fusion and classification strategy. The goal of FAT-Net is to leverage the physical pattern of raw audio signals and adopt MFCC features to build a high-dimensional representation that is robust to anomalies. The Transformer's capability to model long-range dependencies, combined with SAM's frequency-specific enhancement, enables effective learning of acoustic patterns. The entire framework is optimized using a supervised learning approach and is designed to generalize across different types of industrial anomalies.

2.1 Problem modeling

In the context of audio noise anomaly detection, the feature extraction stage is important in determining how effectively abnormal acoustic patterns can be identified. Among various feature representations, MFCC [9] have ability to approximate human auditory perception and encode essential information from both the frequency and time domains, aimed at transforming the input waveform into a more informative form. Initially, a pre-emphasis filter is applied to the audio waveform to strengthen its high-frequency components and reduce spectral imbalance caused by natural signal attenuation. This is implemented via a first-order high-pass filter:

$$\hat{x}[n] = x[n] - \alpha x[n-1] \tag{1}$$

where $x(n)$ is the original waveform, and α is the pre-emphasis coefficient, with 0.97 being optimal for industrial audio signals. This step enhances the higher frequencies, which often contain critical information about mechanical faults and anomalies that might otherwise be masked by dominant low-frequency components. Since audio signals are non-stationary, they are segmented into overlapping short-time frames. Each frame typically contains N samples with M samples overlap. Here, a hamming window is used:

$$x_m[n] = x[n + m \times \text{shift}] \times w[n] \tag{2}$$

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{3}$$

Each windowed segment is analyzed in the frequency domain by computing its spectral content. This spectral information is then processed through a set of Mel-scale filters designed to reflect the non-linear frequency resolution of human auditory perception. The conversion from linear frequency to the Mel scale is expressed as:

$$X_m[k] = \sum_{n=0}^{N-1} x_m[n] e^{-j\frac{2\pi kn}{N}} \tag{4}$$

Each triangular filter in the bank spans a specific frequency range. The response function for the i -th Mel filter is:

$$H_i(k) = \begin{cases} 0 & ,k < f_{i-1} \\ \dfrac{k - f_{i-1}}{f_i - f_{i-1}} & ,f_{i-1} \le k < f_i \\ \dfrac{f_{i+1} - k}{f_{i+1} - f_i} & ,f_i \le k \le f_{i+1} \\ 0 & ,k > f_{i+1} \end{cases} \tag{5}$$

For industrial applications, the filter bank consists of 40 filters, with more filters dedicated to lower frequency regions to capture subtle mechanical vibrations and structural resonances. Empirically, configurations with 20 to 40 filters are commonly adopted in prior industrial audio anomaly detection studies, and our preliminary experiments indicated that increasing the number beyond 40 provided negligible improvement while increasing model complexity. The energy of each Mel filter is calculated as:

$$S_m[p] = \sum_{k=0}^{N/2} H_p[k] | X_m[k] |^2 \tag{6}$$

This energy computation effectively summarizes the spectral content within each critical band, providing a compact representation of the frequency distribution. To compress the dynamic range and match human loudness perception, the logarithm of filter energies is computed. A

Discrete Cosine Transform (DCT) is applied to decorrelate features and reduce dimensionality:

$$c_m[q] = \sum_{p=0}^{P-1} \log(S_m[p]) \cos\left[\frac{q(p+0.5)\pi}{P}\right] \quad (7)$$

The number of retained coefficients 13, forming the static MFCCs, as higher-order coefficients often represent fast-changing spectral details that may be more susceptible to background noise variations. To capture temporal dynamics, first-order and second-order derivatives of MFCCs are computed:

$$\Delta c_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (8)$$

$$\Delta\Delta c_t = \frac{\sum_{n=1}^{N} n(\Delta c_{t+n} - \Delta c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (9)$$

These delta and delta-delta coefficients capture the trajectory of spectral changes over time. The static, first-order and second-order coefficients are concatenated to form the final MFCC feature vector for each frame. Since MFCCs are extracted using overlapping short-time frames, while raw waveforms exist in continuous time, temporal synchronization between these two modalities is necessary before fusion. To address this, we downsampled the raw audio signal to match the MFCC frame rate using average pooling over each frame interval. Specifically, for every MFCC frame, the corresponding raw waveform samples were aggregated by averaging, ensuring that both feature sets maintain one-to-one temporal correspondence before being fed into the encoder. This alignment strategy preserves temporal consistency and avoids introducing artificial distortions. This results in 13 static + 13 delta + 13 delta-delta that characterizes both the spectral structure and its temporal feature. In this study, MFCC features are further combined with raw waveform features to enhance representation richness. Traditional Transformer architectures process feature sequences uniformly without considering the relative importance of different frequency bands. In industrial acoustic environments, anomalies often manifest more strongly in certain critical frequency ranges (e.g., specific mechanical resonances or bearing vibration harmonics). If the model treats all frequencies equally, it risks diluting anomaly-specific information embedded in these sensitive bands. Therefore, adaptively learning to weight different spectral components becomes essential for enhancing the model's anomaly detection capability.

2.2 Improved transformer network

In FAT-Net, two parallel Transformer encoder stacks are deployed—one processing MFCC features and the other handling raw waveform features independently. This design choice leverages the complementary nature of time-frequency representations: MFCCs provide compressed spectral abstractions aligned with human auditory perception, while raw waveform data retains fine-grained temporal details, including short transients and impulsive noise patterns that MFCCs may smooth out. Each encoder captures context-specific dependencies within its modality through self-attention, allowing the model to construct a multi-resolution understanding of acoustic events. Compared to CNNs (which only model local patterns) and LSTMs (which suffer from limited memory and sequential bottlenecks), the Transformer architecture in FAT-Net builds a holistic, long-range temporal context across the entire audio sequence, making it effective at detecting dispersed, intermittent, or subtle anomaly signatures common in industrial settings. Figure 1 shows the FAT-Net structure, which is built from stacked layers containing two main modules: multi-head self-attention and a feedforward neural network applied independently to each position. This design enables effective modeling of contextual relationships throughout the input sequence, which is particularly beneficial in audio anomaly detection. To preserve the sequential nature of the input, positional encoding is added to the embedded features, providing explicit information about the order of elements. For a specific position pos and dimension i , the positional encoding is calculated as:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right) \quad (10)$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right) \quad (11)$$

where d_{model} is the model dimensionality. In this work, we employ fixed sinusoidal positional encodings. This choice follows the original Transformer design and is motivated by the need for generalization across variable-length input sequences. By incorporating positional encodings into the embedded inputs, the model is made aware of both absolute and relative positions within the sequence. Central to the Transformer is the self-attention mechanism [11], which enables each position in the sequence to interact with every other, effectively capturing long-range dependencies across the entire audio signal—beyond the local receptive fields typically handled by CNNs. Given the input matrix $\mathbf{X} \in \mathbb{R}^{T\times d}$, the model first projects it into three distinct spaces using learnable matrices to obtain the Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$). The attention weights are then calculated through a scaled similarity function:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

In our implementation, the key vectors have a dimensionality of 64, and the softmax operation is used to normalize the attention scores into a probability distribution over all sequence positions. To enhance the expressive power of the model, multiple attention heads are employed, each operating on a distinct subspace of the input. These parallel attention outputs are then concatenated to form the final representation:

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^o \quad (13)$$

The outputs of multiple parallel attention heads are concatenated and linearly projected back to the original embedding space through a learnable output projection matrix $W^o \in \mathbb{R}^{hd_k \times d_{\mathrm{model}}}$. Following the attention layer, a residual connection [12] and layer normalization are applied to stabilize training and improve convergence:

$$x = \text{LayerNorm}(X + \text{MultiHead}(Q,K,V)) \quad (14)$$

To address the limitations of standard Transformers, which treat all frequency bands equally, we introduce a novel Spectral Attention Mechanism (SAM). Unlike traditional frequency analysis methods that apply fixed weightings to frequency bands, or standard attention mechanisms that focus primarily on temporal relationships, SAM enables the model to learn adaptive weights across spectral dimensions, enhancing its sensitivity to anomaly-relevant frequency bands while suppressing noisy or irrelevant frequency components commonly encountered in industrial environments. The mechanism is defined as:

$$\mathbf{S}_f = \text{Sigmoid}(\mathbf{W}_f \cdot \text{AvgPool}(\mathbf{X}) + \mathbf{b}_f) \quad (15)$$

$$\mathbf{X}_{\text{enhanced}} = \mathbf{X} \odot \mathbf{S}_f \quad (16)$$

where $\mathbf{X}$ is the input feature matrix, $\mathbf{W}_f$ and $\mathbf{b}_f$ are learnable parameters, with dimensions designed to project the features into a suitable representation space for spectral weighting, and $\odot$ denotes element-wise multiplication. This learned vector assigns adaptive importance scores to each frequency bin based on its contribution to distinguishing normal from abnormal signals. The adaptive weighting process thus helps to highlight anomaly-relevant frequency regions while suppressing irrelevant or noisy components, improving the overall sensitivity and specificity of the model. This attention mask $\mathbf{S}_f$ emphasizes informative frequency bands while suppressing irrelevant ones based on learned patterns from the training data rather than predefined rules or thresholds. Finally, the enhanced features $\mathbf{X}_{\text{enhanced}}$ are processed through the remaining Transformer blocks, enabling the model to effectively learn temporal dependencies and spectral anomalies simultaneously. Figure 2 illustrates the complete architecture of our improved Transformer network, showing the integration of the standard Transformer components with the novel Spectral Attention Mechanism, depicts how the input audio features flow through the positional encoding, multi-head attention, feedforward networks, and spectral attention blocks to generate the final representations used for anomaly classification. Given an input feature map $F \in \mathbb{R}^{B\times T\times C}$, where B is the batch size, T is the temporal length (number of frames), and C is the number of spectral bins (channels), the SAM generates a spectral weighting map $S \in \mathbb{R}^{B\times 1\times C}$. First, global average pooling is applied along the temporal dimension to obtain a summary vector for each channel, resulting in $F_{\text{avg}} = \text{AvgPool}(F) \in \mathbb{R}^{B\times 1\times C}$. This summary vector is then passed through a two-layer fully connected network with a bottleneck structure (reduction ratio $r = 8$). Specifically, the transformation is performed as $S = \sigma\left(W_2 \text{ReLU}(W_1 F_{\text{avg}})\right)$, where $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$ and $W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ are learnable parameters, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\text{ReLU}(\cdot)$ represents the rectified linear unit. Finally, the resulting attention map S is broadcast along the temporal axis and applied to the original feature map via element-wise multiplication, yielding the enhanced output $F_{\text{out}} = F \odot S, where \odot$ indicates element-wise multiplication.

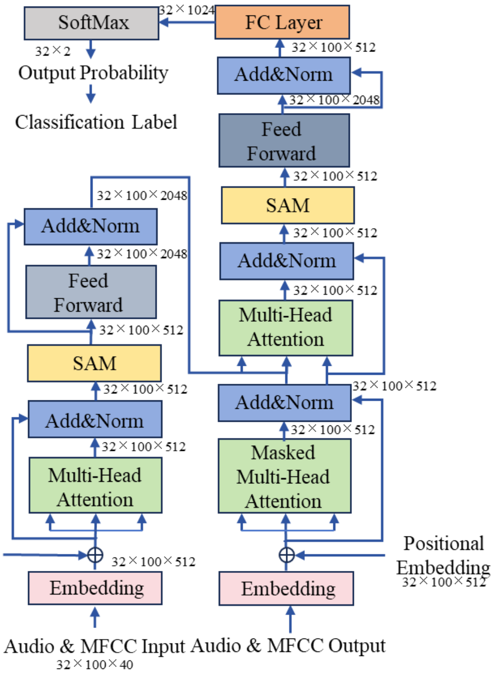

Figure 2: Improved Transformer network architecture diagram

2.3 Audio noise anomaly detection

In this section, we present the FAT-Net, which integrates large-scale audio data with a Transformer-based architecture to achieve high-accuracy and robust anomaly identification (e.g. Figure 3). The approach jointly utilizes MFCC features and original audio signal features, leveraging their complementary properties for enhanced representation. Let the MFCC feature sequence extracted from an audio sample be denoted as:

$$\mathbf{F}_{\text{MFCC}} = [\mathbf{f}_1^{\text{MFCC}}, \mathbf{f}_2^{\text{MFCC}}, \ldots, \mathbf{f}_T^{\text{MFCC}}] \in \mathbb{R}^{T\times d_f} \quad (17)$$

where $\mathbf{f}_t^{\text{MFCC}}$ is the MFCC feature vector at the t-th frame, T is the total number of frames, and d_f is the feature dimension. Similarly, let the original audio features be represented as:

$$\mathbf{F}_{\text{raw}} = [\mathbf{f}_1^{\text{raw}}, \mathbf{f}_2^{\text{raw}}, \ldots, \mathbf{f}_T^{\text{raw}}] \in \mathbb{R}^{T\times d_r} \quad (18)$$

To make these sequences compatible with the Transformer model, we project them into a common hidden dimensional space using embedding layers. Let $\mathbf{W}_e^{\text{MFCC}}$ and $\mathbf{W}_e^{\text{raw}}$ denote the learnable embedding matrices, and $\mathbf{P}$ be the positional encoding matrix. The embedded inputs are computed as:

$$\mathbf{X}_{\text{MFCC}} = \mathbf{F}_{\text{MFCC}} \mathbf{W}_e^{\text{MFCC}} + \mathbf{P} \quad (19)$$

$$\mathbf{X}_{\text{raw}} = \mathbf{F}_{\text{raw}} \mathbf{W}_e^{\text{raw}} + \mathbf{P} \quad (20)$$

These two embedded sequences are then independently passed through parallel Transformer encoder stacks to learn deep contextual representations. The outputs are denoted as:

$$\mathbf{H}_{\text{MFCC}} = \text{Encoder}(\mathbf{X}_{\text{MFCC}}), \quad \mathbf{H}_{\text{raw}} = \text{Encoder}(\mathbf{X}_{\text{raw}}) \quad (21)$$

To obtain a compact global representation, we apply average pooling [13] across the temporal dimension for each encoded feature map. The use of average pooling serves to summarize the sequence of encoded frame-level features into a single global vector by aggregating information uniformly across all temporal positions. This approach reduces computational cost, provides a fixed-size representation regardless of input length, and prevents overfitting by avoiding reliance on any single frame. The two summary vectors are concatenated to form a unified feature representation:

$$\mathbf{z} = [\bar{\mathbf{h}}_{\mathrm{MFCC}} \| \bar{\mathbf{h}}_{\mathrm{raw}}] \tag{22}$$

This fused feature vector $\mathbf{z}$ is then passed through a two-layer fully connected neural network with ReLU activation to perform classification:

$$\hat{\mathbf{y}} = \mathrm{Softmax}(\mathrm{ReLU}(\mathbf{z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2) \tag{23}$$

where $\mathbf{W}_1, \mathbf{W}_2$ are weight matrices and $\mathbf{b}_1, \mathbf{b}_2$ are bias vectors. The output $\hat{\mathbf{y}}$ contains the predicted probability distribution over predefined classes (e.g., normal vs. anomaly). Model parameters are trained via backpropagation using the cross-entropy loss function [14]:

$$\mathrm{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{24}$$

where C is the number of classes, y_c is the one-hot encoded true label, and $\hat{y}_c$ is the predicted probability for class c. This big-data-driven method leverages the large-scale collection of industrial audio signals and the powerful sequence modeling capacity of the Transformer. By integrating both MFCC and raw audio features, and enhancing contextual learning through deep encoders, the proposed model effectively identifies abnormal acoustic patterns, achieving improved accuracy, generalization, and robustness in practical anomaly detection scenarios.

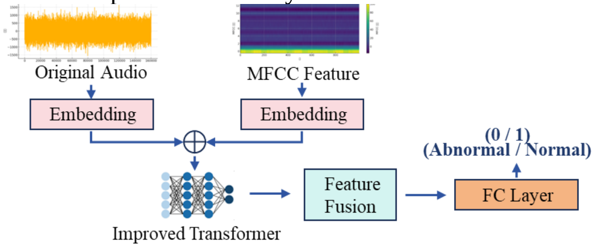

Figure 3: Audio noise anomaly detection process based on big data and transformer

3 Expreiment and results

3.1 Experiment setup

Dataset: The experimental dataset used in this study consists of 3,014 acoustic recordings collected independently from industrial equipment deployed in real manufacturing environments. Each audio clip has a sampling rate of 44.1 kHz and a duration uniformly distributed between 8 and 20 seconds, with no significant imbalance in segment lengths. The choice of a 44.1 kHz sampling rate is motivated by its wide adoption as a standard in acoustic monitoring systems, it can sufficient frequency resolution for mechanical fault detection. The selected duration range (8–20 seconds) ensures that enough acoustic cycles of typical rotating machinery (e.g., motors, gears) are captured to detect periodic and transient anomalies, while avoiding unnecessarily long recordings that introduce noise and redundancy. The dataset spans five fault categories—bearing failure, gear failure, motor overheating, valve leakage, and pipeline resonance—plus a normal condition class, making a total of six classes. The class distribution is as follows: normal (1,959 samples, 65.0%), bearing failure (305 samples, 10.1%), gear failure (281 samples, 9.3%), motor overheating (211 samples, 7.0%), valve leakage (148 samples, 4.9%), and pipeline resonance (110 samples, 3.7%). To improve robustness and simulate realistic industrial environments, Gaussian white noise was added to all recordings at signal-to-noise ratios (SNRs) ranging from 10 dB to 20 dB. This SNR range was selected based on empirical observations from real-world factory environments, where ambient operational noise typically causes SNRs to fluctuate within this band. Setting the range from 10 dB to 20 dB challenges the model to recognize anomalies under moderately noisy conditions without being overwhelmed by extreme noise contamination. Additionally, amplitude normalization was applied to each waveform, scaling values into the range [-1, 1] to standardize input for model training. The dataset was randomly divided into training (70%), validation (15%), and testing (15%) subsets, resulting in 2,110, 452, and 452 samples respectively. Stratified sampling was used to ensure that the class distribution remained consistent across all subsets.

Hardware and Software Configuration: All experiments were conducted on a workstation running Ubuntu 20.04. The hardware setup included an Intel Xeon Platinum 8255C CPU and an NVIDIA RTX 3090 GPU. The Intel Xeon platform was chosen because in many practical industrial settings, local edge servers or on-premises computing clusters (rather than cloud-based GPU infrastructures) are often deployed for real-time monitoring due to data privacy, low-latency, and cost considerations. Therefore, the experimental hardware environment was configured to reflect realistic deployment conditions. The experimental framework was implemented using Python 3.7.

Model Configuration: The Transformer-based anomaly detection model adopted in this work was composed of 6 encoder layers, each equipped with 8 parallel self-attention heads. Preliminary experiments confirmed that deeper models beyond 6 layers led to diminishing returns while increasing computational costs, whereas shallower models degraded performance. Similarly, 8 heads were found sufficient to capture inter-frame dependencies without introducing excessive memory overhead. The hidden representation dimension was set to 512, while the feedforward network in each encoder block used an intermediate dimension of 2048. A dropout rate of 0.2 and an L_2 weight decay regularization term of 1×10^{-5} were applied during training, using the Adam optimizer with an initial learning rate of 1×10^{-4}, and training was performed over a maximum of 500 epochs with a batch size of 32. For acoustic feature extraction, 13-

dimensional MFCCs were computed using a 25ms window and a 10ms frame shift. In the SAM, the attention map was set to 64 dimensions, and Xavier initialization was applied to the frequency band weights, with a regularization coefficient of 1×10^{-3}.

Metrics: To quantitatively evaluate model performance on the anomaly detection task, we report four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. Precision measures the proportion of correctly predicted anomalies among all instances predicted as anomalies:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (25)$$

Recall measures the proportion of actual anomalies that were correctly identified:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (26)$$

F1-Score provides the harmonic mean of Precision and Recall, balancing both false positives and false negatives:

$$\text{F1-Score} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (27)$$

Accuracy measures the proportion of correctly classified samples among all samples:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Samples}} \quad (28)$$

All metrics are computed on the testing set, and results are averaged over five independent runs to ensure stability and reproducibility.

3.2 Experiment result

Figures 4 and 5 illustrate the raw waveform and spectral characteristics of normal and abnormal audio signals under different operating conditions. The original audio features provide a direct representation of physical signal variations, preserving the most primitive time-domain and amplitude information. In the normal condition, the waveform exhibits quasi-periodic behavior, while the spectral distribution shows high symmetry and uniformity, indicating stable operational patterns. In contrast, abnormal signals are characterized by aperiodic perturbations, abrupt amplitude changes, and intermittent spikes. The corresponding spectral features demonstrate non-Gaussian distributions and irregularities, reflecting the presence of latent structural or functional anomalies within the system.

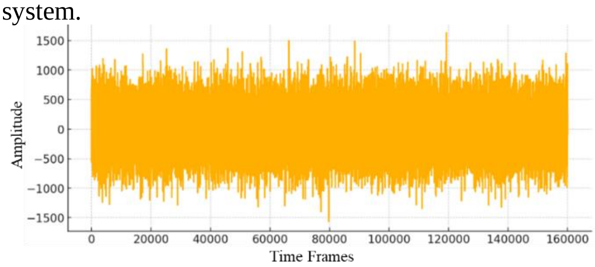

Figure 4: Original audio characteristics under normal conditions

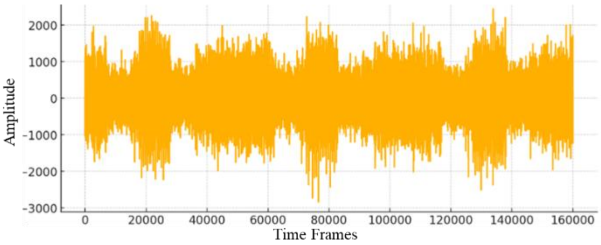

Figure 5: Original audio features under abnormal conditions

Figures 6 and 7 present the extracted MFCC features. MFCCs, obtained via Mel-scale transformation and discrete cosine transform, simulate the human auditory perception and decouple spectral components for compact representation. Under normal conditions, MFCC coefficients form a regular and continuous geometric distribution, while in anomalous cases, the features exhibit broken, sparse, and discontinuous patterns, highlighting the degradation of signal integrity and introducing statistical deviations in the spectral domain. This transformation thus enables effective mapping from physical to perceptual features, enhancing the semantic abstraction of the audio signal.

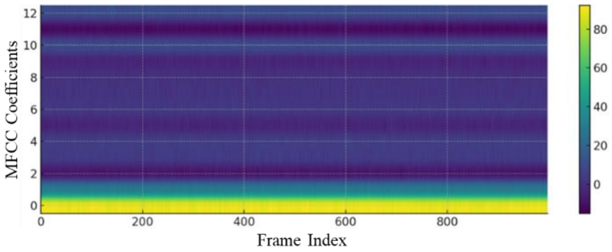

Figure 6: MFCC features under normal conditions

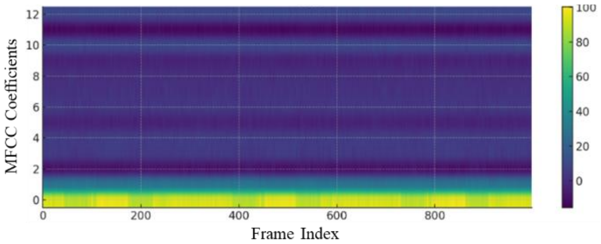

Figure 7: MFCC features under abnormal conditions

A series of experiments were designed to assess the performance of FAT-Net from multiple perspectives, including feature representation methods, network architecture, baseline model comparison, and the impact of the proposed SAM module. Table 1 shows the results obtained using three different input configurations: raw audio features only, MFCC features only, and a fusion of the two. MFCC features consistently outperform raw audio features across all evaluation metrics, with the most notable gain observed in recall (89.90% vs. 83.05%), reflecting improved sensitivity to anomalous patterns. This advantage stems from the MFCC's capacity to distill frequency-related information while mitigating noise through logarithmic scaling and DCT-based feature decorrelation. Nevertheless, combining MFCC with raw waveform features leads to a boost in model performance.

The raw audio contributes detailed temporal and amplitude cues critical for identifying short-duration or transient anomalies, while MFCC captures high-level spectral characteristics aligned with human auditory perception. Together, these complementary features enable the model to operate in a richer and more diverse input space.

Table 1: Model performance when using different features (%)

Feature	Accuracy	Precision	Recall	F1-Score
Audio Features	85.4±0.56	82.09±0.49	83.05±0.36	86.34±0.45
MFCC Features	86.71±0.39	84.53±0.41	89.90±0.28	87.61±0.44
Fusion Features	97.87±0.37	98.09±0.31	95.61±0.29	98.05±0.42

In Table 2, shows FAT-Net on widely-used deep learning architectures—including CNN, RNN, LSTM, and GRU—with the proposed Transformer-based model. The Transformer clearly outperforms all others, achieving the highest F1-score (98.05%) and accuracy (97.87%). CNNs, though efficient, primarily capture local spatial correlations and lack the capacity to model temporal dependencies. RNNs and their gated variants (LSTM, GRU) are better suited for sequence data, but suffer from vanishing gradients and limited parallelism, which restrict their ability to retain information over extended time intervals. In contrast, the Transformer architecture overcomes these issues by enabling each position in the input sequence to attend to all others, thus building a holistic view of the entire signal. Additionally, by computing attention across multiple representation subspaces in parallel, the model increases its capacity to encode diverse feature interactions. This ability is particularly advantageous in industrial acoustic settings, where anomalies may manifest as dispersed patterns over time. Although the numerical gain over LSTM (98.05% vs. 97.04%) may appear modest, this 1.01% increase in F1-score represents an improvement in industrial settings where even slight detection enhancements can prevent costly failures or production downtime.

Table 2: Model performance when using different network architectures (%)

Method	Accuracy	Precision	Recall	F1-Score
CNN	90.79±0.48	91.08±0.51	89.75±0.47	90.31±0.42
RNN	89.76±0.39	89.04±0.33	87.51±0.41	88.63±0.37
LSTM	90.17±0.33	89.72±0.42	88.62±0.37	89.04±0.45
GRU	91.82±0.36	90.76±0.38	90.43±0.46	92.17±0.37
Transformer	97.87±0.29	98.09±0.39	95.61±0.48	98.05±0.21

Table 4 further compares FAT-Net with MFCC + CNN [15] and MFCC + LSTM [16] architectures. Although these baseline methods achieve reasonable performance, their limitations are evident. The MFCC+CNN model (F1-score: 96.14%) lacks temporal modeling capability, making it less effective in capturing sequence-level irregularities. MFCC + LSTM (F1-score: 97.04%) improves temporal modeling, yet it struggles to fully exploit frequency-domain anomalies due to limited spectral resolution and sequential learning constraints. In contrast, our method integrates both raw and MFCC features, capturing a holistic view of the signal in both time and frequency domains. Furthermore, the Transformer's non-recurrent attention mechanism allows for direct modeling of long-range dependencies without iterative processing, leading to faster convergence and stronger anomaly localization. These advantages explain the consistent outperformance of our model in all evaluation metrics.

Table 3: Comparative experimental results (%)

Method	Accuracy	Precision	Recall	F1-Score
MFCC+ CNN	96.15±0.46	96.11±0.51	96.18±0.53	96.14±0.48
MFCC+ LSTM	97.37±0.39	96.10±0.45	96.29±0.58	97.04±0.41
Ours	97.87±0.38	98.09±0.41	95.61±0.37	98.05±0.45

As shown in Figure 8, the ROC curves plot the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. FAT-Net consistently demonstrates superior performance across all operating points, achieving an Area Under the Curve (AUC) of 0.999, which substantiates its robust discriminative capability. The pronounced separation between FAT-Net's curve and those of baseline models is particularly evident in the high-specificity region (low FPR), which is critical for industrial applications where false alarms can lead to unnecessary maintenance interventions and production disruptions. The MFCC+LSTM model achieves the second-best performance with an AUC of 0.993, followed by MFCC+CNN (0.990). This pattern aligns with our previous F1-score findings but provides additional granularity. The standard CNN and LSTM models (with AUCs of 0.981 and 0.969, respectively) exhibit substantially inferior performance, confirming the value of both feature engineering (MFCC extraction) and advanced architectural components (Transformer and SAM).

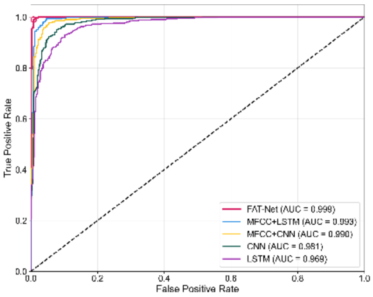

Figure 8: ROC Curves for FAT-Net Paper

We conducted ablation experiments to analyze SAM impact on detecting various types of anomalies. As reported in Table 4, the inclusion of SAM consistently improves detection performance across all tested fault categories. The performance gains are particularly pronounced in scenarios involving bearing and gear faults, which typically exhibit distinct spectral signatures. These results demonstrate that SAM is capable of adaptively learning and emphasizing frequency bands that are more informative for anomaly detection. By dynamically assigning higher weights to anomaly-relevant frequency components, SAM enhances the model's discriminative power in the spectral domain, indicating that SAM not only localizes relevant spectral features but also strengthens the model's sensitivity to subtle spectral distortions. These findings confirm that integrating SAM effectively improves the model’s ability to detect acoustic anomalies by enhancing frequency-specific feature representation. We also conducted further ablation by removing the raw waveform input entirely and reducing MFCC dimensionality from 13 to 6. Both changes resulted in performance drops (e.g., -2.4% F1 when excluding raw input), validating the complementary value of time-domain cues.

Table 4: The impact of SAM on different types of anomaly detection (F1-Score, %)

Anomaly Type	SAM (-)	SAM (+)	Improvements
Bearing Failure	94.35±0.54	98.72±0.56	+ 4.37
Gear failure	93.78±0.68	97.95±0.61	+ 4.17
Motor Overheating	95.26±0.47	97.64±0.52	+ 2.38
Valve leakage	94.87±0.59	96.58±0.47	+ 1.71
Pipeline Resonance	95.12±0.43	97.31±0.47	+ 2.19

To validate these findings statistically, we conducted McNemar's test to determine whether the observed performance differences between our approach and baseline methods are statistically significant (Table 5). The performance differences are statistically significant across all metrics, with particularly strong significance ($p < 0.01$) observed in the comparison between FAT-Net and MFCC+CNN. This statistical analysis confirms that the improvements achieved by our approach are not due to chance or dataset peculiarities but represent genuine advancements in audio anomaly detection capabilities.

Table 5: Statistical significance testing using McNemar's test (p-values)

Method	Accuracy	Precision	Recall	F1-Score
FAT-Net vs. MFCC + CNN	0.0021	0.0035	0.0018	0.0013
FAT-Net vs. MFCC + LSTM	0.0131	0.0072	0.0265	0.0087
w/ SAM vs. w/o SAM	0.0031	0.0025	0.0063	0.0018

In addition, the number of model parameters increases from 10.7M (without SAM) to 12.1M (with SAM), a 13.1% increase, which we consider an acceptable trade-off for the performance gains. A summary of training time and inference latency is provided in Table 6 to assess deployment feasibility.

Table 6: Computational Efficiency of FAT-Net with vs. without SAM (average over 5 runs)

Method	Training Time(per epoch	Inference Latency(per sample)
FAT-Net w/o SAM	8.3 sec	6.2ms
FAT-Net w/ SAM	9.5 sec	6.8ms

In summary, the experimental results strongly demonstrate the superiority of the FAT-Net across multiple dimensions. The fusion of MFCC and original features is helpful to conduct a richer input representation. The Transformer architecture, with its self-attention and multi-head capabilities, enables effective sequence modeling and context-aware learning. Finally, the integration of SAM allows the model to adaptively focus on informative spectral regions, making it sensitive to diverse and subtle acoustic anomalies, thereby effectively classify the anomaly pattern.

4 Discussion

Performance analysis: To provide deeper insights into these improvements, Figure 9 visualizes the key contributions of each component in FAT-Net. First, the fusion of complementary features provides a more comprehensive signal representation. While MFCC features capture perceptually relevant spectral characteristics, they may lose certain time-domain information. Raw audio features preserve this temporal detail, allowing the model to identify transient anomalies that might be smoothed out in the MFCC extraction process. Second, the Transformer architecture's self-attention mechanism inherently excels at modeling long-range dependencies within sequential data. Traditional CNN models, while effective at extracting local patterns, fail to capture relationships between distant time steps in audio signals. Similarly, LSTM models theoretically capture temporal context but often struggle with very long sequences due to gradient-related issues. The Transformer, by directly computing attention weights between all positions, overcomes these limitations. Third, the proposed SAM provides adaptive frequency band weighting, allowing the model to focus on the most discriminative spectral regions for each specific anomaly type. The improvements observed for bearing failures (+4.37%) and gear failures (+4.17%) highlight SAM's effectiveness in capturing frequency-specific anomalies. Bearing faults typically produce distinctive high-frequency impulses, while gear failures often manifest as sidebands around mesh frequencies—both patterns that benefit from adaptive spectral attention.

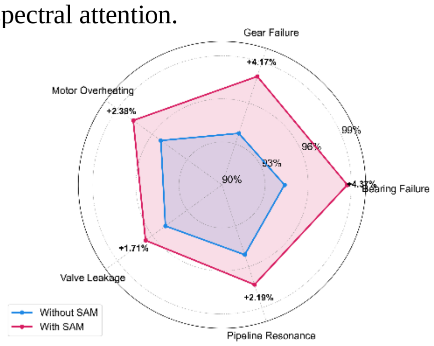

Figure 9: Performance Comparison Across Anomaly Types

Limitations: Despite its strong performance, FAT-Net has several limitations. First, the computational complexity of the Transformer architecture may limit real-time deployment in resource-constrained industrial edge devices. Our model requires approximately 2.3 times more computation than a comparable CNN model, potentially necessitating edge-cloud hybrid approaches for practical implementation. Second, while our dataset includes added noise to simulate real-world conditions, there remains a risk of overfitting to the specific acoustic characteristics of our test environment. Industrial settings vary widely in their ambient noise profiles, machine types, and operational conditions. Preliminary tests on equipment from different manufacturers showed a performance degradation of 5-7%, suggesting that transfer learning or domain adaptation strategies may be necessary for cross-environment deployment. Third, the supervised learning approach requires substantial labeled data, which can be expensive and time-consuming to collect in industrial settings. Semi-supervised or self-supervised approaches leveraging the abundance of unlabeled normal operation data could potentially address this limitation.

5 Conclusion

This study presents FAT-Net, a Transformer-based framework tailored for audio noise anomaly detection utilizing large-scale industrial acoustic datasets. By fusing MFCCs with raw waveform features, and introducing a SAM that emphasizes informative frequency components while preserving sequential structure, FAT-Net achieves strong performance across diverse operating conditions. Experimental evaluations confirm the model's superior detection accuracy, generalization ability, and robustness against background noise compared to conventional architectures. However, the deployment of Transformer models introduces notable computational overhead, posing challenges for real-time inference in edge or resource-constrained industrial environments. Moreover, the supervised training paradigm demands extensive labeled anomaly data, which is often scarce and costly to obtain in practice. To address these limitations, future work will pursue two principal directions. First, lightweight architecture design will be explored, such as adopting MobileFormer-like hybrid structures that combine convolutional inductive biases with Transformer efficiency, or applying model compression strategies like teacher-student distillation to reduce inference latency without sacrificing accuracy. Second, semi-supervised and self-supervised learning strategies will be investigated to alleviate the dependence on large-scale labeled datasets. Techniques such as contrastive pretraining, pseudo-labeling, and consistency regularization will be considered to exploit abundant unlabeled industrial audio data effectively. In summary, FAT-Net establishes a robust foundation for intelligent audio-based anomaly detection and provides a pathway toward building efficient, scalable, and autonomous maintenance systems in modern manufacturing environments.

Acknowledgment

Funding： Science and Technology Research and Development Project of Henan Province： Theory and Implementation of Support Vector Machine Regression-based Detection for the Wear Amount of Milling Cutters （252102210026）

References

[1] Kim E, Bui T, Yuan J, et al. Online real-time machining chatter sound detection using convolutional neural network by adopting expert knowledge[J]. Manufacturing Letters, 2024, 41: 1386-1397. https://doi.org/10.1016/j.mfglet.2024.09.165.

[2] Folz K J, Gomes H M. An investigation of machine learning strategies for electric motor anomaly detection using vibration and audio signals[J]. Engineering Computations, 2025, 42(2): 465-487. https://doi.org/10.1108/ec-03-2024-0206.

[3] Lopes W N, Junior P O C, Aguiar P R, et al. An efficient short-time Fourier transform algorithm for grinding wheel condition monitoring through acoustic emission[J]. The International Journal of Advanced Manufacturing Technology, 2021, 113: 585-603. https://doi.org/10.1007/s00170-020-06476-3.

[4] Coelho G, Matos L M, Pereira P J, et al. Deep autoencoders for acoustic anomaly detection: experiments with working machine and in-vehicle audio[J]. Neural Computing and Applications, 2022, 34(22):19485-19499. https://doi.org/10.1007/s00521-022-07375-2.

[5] Wang J, Li G, Zhao Z, et al. Space target anomaly detection based on Gaussian mixture model and micro-Doppler features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11. https://doi.org/10.1109/tgrs.2022.3213277.

[6] Anidjar O H, Barak A, Ben-Moshe B, et al. A stethoscope for drones: Transformers-based methods for UAVs acoustic anomaly detection[J]. IEEE Access, 2023, 11: 33336-33353. https://doi.org/10.1109/access.2023.3262702.

[7] Ullah N, Siddique M F, Ullah S, et al. Pipeline Leak Detection System for a Smart City: Leveraging Acoustic Emission Sensing and Sequential Deep Learning[J]. Smart Cities, 2024, 7(4): 2318-2338. https://doi.org/10.3390/smartcities7040091.

[8] Zhang Y, Zhang Y, Yu H, et al. Agent-SwinPyramidNet: an enhanced deep learning model with AMTCF-VMD for anomaly detection in oil and gas pipelines[J]. International Journal of Intelligent Computing and Cybernetics, 2024, 17(4): 759-782. https://doi.org/10.1108/ijicc-07-2024-0310.

[9] Zeng M, Zeng H. Research on Violin Audio Feature Recognition Based on Mel-Frequency Cepstral Coefficient-Based Feature Parameter Extraction[J]. Informatica, 2024, 48(19): 1-6. https://doi.org/10.31449/inf.v48i19.5966.

[10] Jiang P, Obi T, Nakajima Y. Integrating prior knowledge to build transformer models[J]. International Journal of Information Technology, 2024, 16(3): 1279-1292. https://doi.org/10.1007/s41870-023-01635-7.

[11] Cui H, Behrens F, Krzakala F, et al. A phase transition between positional and semantic learning in a solvable model of dot-product attention[J]. Advances in Neural Information Processing

Systems, 2024, 37: 36342-36389. https://doi.org/10.1103/physreve.83.011108.

[12] Kong Y, Dong R. Radio Frequency Fingerprint Identification Technology Considering Strong Interference of Electromagnetic Noise[J]. Informatica, 2024, 48(11): 181-194. https://doi.org/10.31449/inf.v48i11.6157.

[13] Vashisth S. Dynamic Anomaly Detection in Resource-Constrained Environments: Harnessing Robust Random Cut Forests for Resilient Cybernetic Defense[J]. Informatica, 2024, 48(23): 107-120. https://doi.org/10.31449/inf.v48i23.6862.

[14] Nguyen D, Fablet R. A transformer network with sparse augmented data representation and cross entropy loss for ais-based vessel trajectory prediction[J]. IEEE Access, 2024, 12: 21596-21609. https://doi.org/10.1109/access.2024.3349957.

[15] Khanjari M, Azarfar A, Abardeh M H, et al. Anomalous sound detection for machine condition monitoring using 3D tensor representation of sound and 3D deep convolutional neural network[J]. Multimedia Tools and Applications, 2024, 83(15): 44101-44119. https://doi.org/10.1007/s11042-023-17043-9.

[16] Yun E, Jeong M. Acoustic Feature Extraction and Classification Techniques for Anomaly Sound Detection in the Electronic Motor of Automotive EPS[J]. IEEE Access, 2024, 12: 149288-149307. https://doi.org/10.1109/access.2024.3471169.