Dual-Channel Acoustic Temporal–Spectral Representation and Noise-Perturbation Learning for Mining Conveyor Fault Diagnosis

Zhenyu Wang[1], Taixiang Li[2,*]

[1] *Zijin Zhixin Zhikong (Xiamen) Technology CO.,Ltd., Xiamen, China*
wang_zhenyu@zijinmining.com
[2] *Big Data Institute, Central South University, Changsha, China*
[*] *taixiangl@csu.edu.cn*

ABSTRACT

This study develops an acoustic fault diagnosis framework for mining conveyor idlers, addressing early-stage mechanical degradation under imbalanced field data and non-ideal industrial acoustic conditions. A dual-channel temporal–spectral representation is constructed by combining multi-scale log-Mel spectrograms and raw waveforms to capture complementary spectral patterns and fine-grained temporal dynamics. A Dual-Stream Cross-Attention Convolutional Recurrent Neural Network (DSCA-CRNN) is proposed to model cross-stream dependencies and enhance feature fusion. Class-weighted focal learning, training-time raw-waveform perturbation, and a triplet-based contrastive objective are employed to improve minority-class discrimination and embedding compactness. Experiments on a self-collected conveyor auscultation dataset with 2,495 segments across three health states show that DSCA-CRNN achieves an accuracy of 0.952 and a macro-F1 score of 0.900, outperforming representative machine learning and deep learning baselines. Severe-fault recognition reaches an F1-score of 0.803. Ablation studies and feature-space analysis confirm the effectiveness of recurrent temporal modeling, cross-stream attention fusion, and contrastive representation shaping. The proposed auscultation-based framework provides a practical window-based solution for offline or near-real-time conveyor condition monitoring.

Keywords: Acoustic signal processing; Conveyor fault diagnosis; Deep learning; Contrastive learning; Noise augmentation; Automation

1. INTRODUCTION

Conveyor belt systems are the backbone of bulk material transportation in modern mining and mineral processing industries, where they operate continuously under harsh environmental conditions such as heavy dust, vibration, and high mechanical load (Alharbi et al., 2023; Bzinkowski et al., 2024). The operational reliability of these systems directly affects the safety and efficiency of ore handling and production. Consequently, early fault detection and condition monitoring are essential to prevent unplanned downtime and ensure stable operation (Soares et al., 2024; Khalifa et al., 2025). Conventional monitoring approaches—such as vibration analysis, infrared thermography, or visual inspection—have been widely applied in mining operations (Farhat et al., 2023; X. Wu et al., 2023). However, these methods often require expensive sensors, complex installation, and clean optical access, which make them impractical for dusty and acoustically noisy mining environments.

In contrast, acoustic sensing offers a promising non-contact alternative for health monitoring of conveyor components, particularly idlers and rollers. By capturing and analyzing the subtle sound variations generated during belt movement, it becomes possible to identify abnormal mechanical behaviors before severe failures occur. This concept, often referred to as "stethoscope-style" auscultation monitoring, provides a cost-effective and easily deployable sensing route for mining sites (Chen et al., 2025; X. Li et al., 2024). The apparent tension is that mining sites are acoustically noisy, whereas the proposed method also relies on acoustic signals. In this work, this challenge is addressed at two levels: close-range auscultation is used to increase the signal-to-background ratio during data acquisition, and the model combines complementary time- and frequency-domain representations with class-aware representation learning to reduce sensitivity to local spectral fluctuations. Nevertheless, the present experiments should be interpreted as a close-range field validation rather than a complete validation under arbitrary microphone placements or controlled external noise levels.

Recent advances in machine learning and deep learning have significantly enhanced the ability to model such complex acoustic dynamics (Qiu et al., 2023; Zhao et al., 2019; Jiao et al., 2020). Traditional machine learning methods, such

https://doi.org/10.36001/IJPHM.2026.v17i2.4841

as Support Vector Machines (SVM), Random Forests (RF), and K-Nearest Neighbors (KNN), have been applied to extract handcrafted time- and frequency-domain features for fault classification (M. Wang et al., 2023; Soares et al., 2025). However, their performance heavily depends on feature design and lacks adaptability to unseen operating conditions. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated strong capabilities in end-to-end representation learning for acoustic and vibration-based fault diagnosis (Lei et al., 2016; J. Wang et al., 2019; Zhang et al., 2018; Y. Liu et al., 2023; Z. Li et al., 2024; J. Liu et al., 2025). CNNs effectively capture local spectral patterns, while RNNs model sequential dependencies within time-series data. Hybrid architectures such as CRNNs further combine these strengths, enabling simultaneous learning of spatial–temporal dependencies (Y. Liu et al., 2023; Z. Li et al., 2024). However, real-world industrial data, particularly from mining conveyors, typically suffer from severe class imbalance, where fault samples are much rarer than normal conditions, leading to biased training and poor generalization on minority classes (Soares et al., 2024; Sung et al., 2024; Kim et al., 2025).

To address class imbalance, various imbalance-learning strategies have been proposed in recent studies (Qiu et al., 2023; Zhang et al., 2018). Data-level approaches, such as oversampling and undersampling, attempt to balance class distributions by adjusting sample ratios. Algorithm-level methods, including cost-sensitive learning, focal loss, and class-balanced reweighting, modify the training objective to emphasize minority classes (Sung et al., 2024). In the context of deep learning, data augmentation and transfer learning have also shown promise in alleviating imbalance by increasing the diversity and representativeness of scarce fault samples (Z. Li et al., 2024; J. Liu et al., 2025). However, relying only on resampling or static class weighting may still be insufficient when scarce fault samples exhibit complex temporal–spectral variability (Chen et al., 2025). The present work therefore does not reject class weighting; rather, it combines class-weighted focal learning with perturbation-based regularization and contrastive embedding constraints so that minority classes are emphasized both in the decision loss and in the learned representation space.

To overcome these limitations, this study proposes an intelligent acoustic diagnosis framework tailored for conveyor belt systems in mining environments. The framework leverages deep learning–based feature extraction for end-to-end fault recognition while addressing data imbalance and temporal–spectral complexity in industrial sound signals. Specifically, a Dual-Stream Cross-Attention Convolutional Recurrent Neural Network (DSCA-CRNN) is developed to jointly learn high-level temporal–spectral representations by combining log-Mel spectrograms with raw waveform features. The Mel branch captures frequency–energy distributions, whereas the raw branch preserves fine-grained temporal structures, enabling complementary feature learning across heterogeneous acoustic patterns. Furthermore, class-weighted focal learning, training-time raw-waveform perturbation, and a triplet contrastive objective are combined to mitigate minority-class under-representation without relying solely on synthetic duplication. The proposed framework is evaluated through five-fold cross-validation on a self-collected close-range conveyor auscultation dataset from an operating mining site.

The main innovations of this work are summarized as follows:

- A practical auscultation-based acoustic monitoring framework was established for mining conveyor systems, enabling non-contact and cost-effective fault diagnosis through sound analysis.
- A class-imbalance-aware training strategy was studied by combining class-weighted focal learning, raw-waveform noise perturbation, and triplet contrastive regularization to support minority-class representation learning.
- A dual-branch temporal–spectral representation was proposed by integrating raw waveform features with log-Mel spectrograms, providing complementary temporal and spectral cues and significantly improving fault discriminability.

2. MATERIALS AND METHODS

2.1. Problem Definition

In this study, the objective is to develop an intelligent diagnostic framework capable of identifying the operational condition of mining conveyor belt idlers through acoustic signals. Mechanical defects in idlers—such as bearing wear, misalignment, or shell cracking—induce subtle yet distinguishable variations in the emitted sound. The task can therefore be formulated as a supervised multi-class classification problem. Let $\mathbf{x}_i \in \mathbb{R}^T$ denote a raw acoustic waveform of length T, and let $y_i \in \{0, 1, 2\}$ represent its corresponding condition label, where the three classes correspond to *normal*, *warning*, and *severe fault*. The goal is to learn a discriminative mapping function

$$f_\theta : \mathbf{x}_i \rightarrow y_i, \tag{1}$$

parameterized by θ, that can accurately infer the operational state of previously unseen audio samples.

A major challenge in this problem lies in the characteristics of real mining-site auscultation data. The recorded signals are highly non-stationary and affected by environmental acoustic variability, which can obscure fault-related acoustic patterns. Moreover, the dataset exhibits severe class imbalance: normal samples are overwhelmingly abundant, whereas warning and severe-fault cases are scarce and often display substantial intra-class variability. Such imbalance can cause conventional learning models to overfit majority patterns and show reduced sensitivity to early-stage or subtle fault conditions.

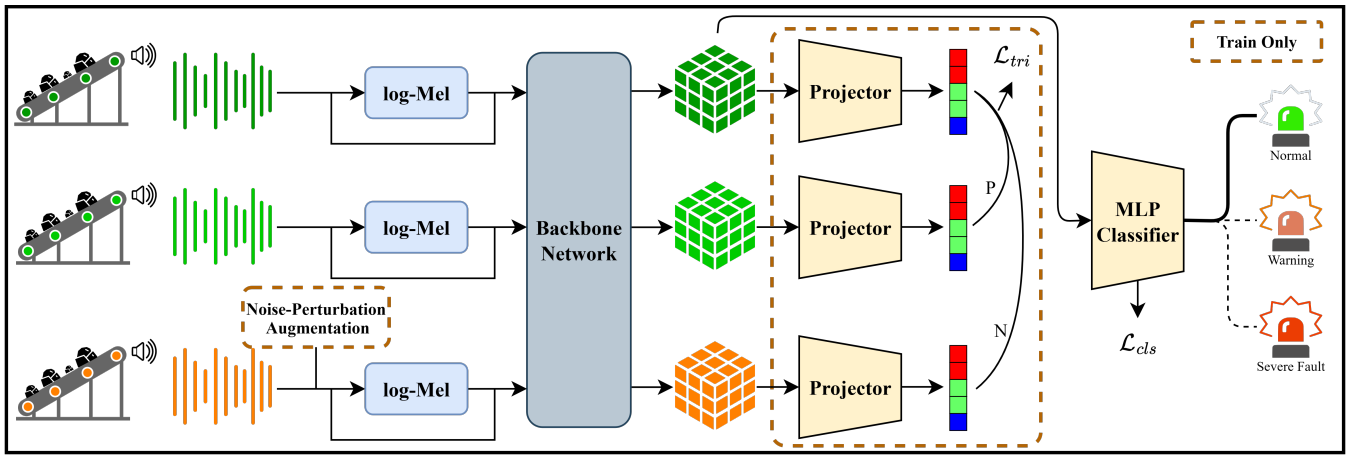


Figure 1. Overall framework of the proposed DSCA-CRNN-based acoustic diagnosis system. Multi-scale log-Mel spectrograms and raw waveform signals are used as dual-channel inputs to capture complementary spectral and temporal characteristics. The DSCA-CRNN backbone extracts hierarchical temporal-spectral features through dual-stream convolutional encoding, cross-stream attention fusion, and bidirectional recurrent modeling. During training, the three illustrated paths correspond to anchor, positive, and negative samples for triplet contrastive learning; they share the same DSCA-CRNN parameters and are shown separately only to illustrate the training procedure. The classification head and projection head are jointly optimized using classification loss ($\mathcal{L}_{cls}$) and triplet contrastive loss ($\mathcal{L}_{tri}$). During inference, a single input window passes through one shared backbone and the classification head only.

Consequently, the core difficulty is to extract robust temporal–spectral representations from industrial audio while preserving discriminability across highly imbalanced fault categories.

During data preprocessing, each acoustic waveform is segmented into fixed-length windows, from which two complementary representations are derived: the log-Mel spectrogram that captures the signal's frequency–energy distribution, and the raw waveform that retains the fine-grained temporal structure.

The Mel branch employs a 2D convolutional encoder to extract hierarchical spectral features, whereas the raw branch uses a 1D convolutional encoder to model time-domain structures. Both feature sequences are downsampled to a uniform temporal resolution and concatenated before being fed into a bidirectional GRU for high-level temporal modeling. The network outputs one of three condition categories: *normal*, *warning*, and *severe fault*.

To address limited and imbalanced data, class-aware training regularization and five-fold cross-validation are incorporated, improving model generalization and ensuring reliable statistical performance.

2.2. Algorithm Framework

The overall workflow of the proposed DSCA-CRNN-based acoustic diagnosis system is illustrated in Fig. 1. The framework takes dual-channel inputs consisting of multi-scale log-Mel spectrograms and raw waveform signals to capture complementary spectral and temporal characteristics of conveyor idler acoustics. The DSCA-CRNN backbone extracts hierarchical temporal-spectral features through dual-stream convolutional encoding, cross-stream attention fusion, and bidirectional recurrent modeling. During training, the extracted features are simultaneously fed into a classification head for condition recognition and a projection head for triplet contrastive feature learning. The three paths shown in Fig. 1 represent anchor, positive, and negative samples processed by parameter-sharing instances of the same DSCA-CRNN; they are separated only to visualize triplet construction during training. During inference, only a single input window and the classification head are used; the projection head and triplet sampling procedure are disabled. This distinction clarifies that the triplet objective is a training-time representation constraint rather than three independent inference networks.

2.2.1. Dual-channel feature representation

Acoustic signals emitted by conveyor idlers contain rich temporal and spectral information influenced by rolling friction, mechanical impacts, and ambient noise. To comprehensively characterize these complementary aspects, a dual-channel feature representation is constructed based on multi-scale log-Mel spectrograms and raw waveform features.

The multi-scale Mel channel is obtained by computing three Mel spectrograms under different analysis resolutions, using distinct combinations of $(n_{\text{mels}}, n_{\text{fft}}, \text{hop})$. This captures narrow-band harmonic structures (e.g., 32-Mel), mid-resolution spectral envelopes (64-Mel), and broad-band excitation components (128-Mel). Each Mel map is normalized and resized to a fixed spatial dimension, after which the three maps are stacked to form a unified three-channel

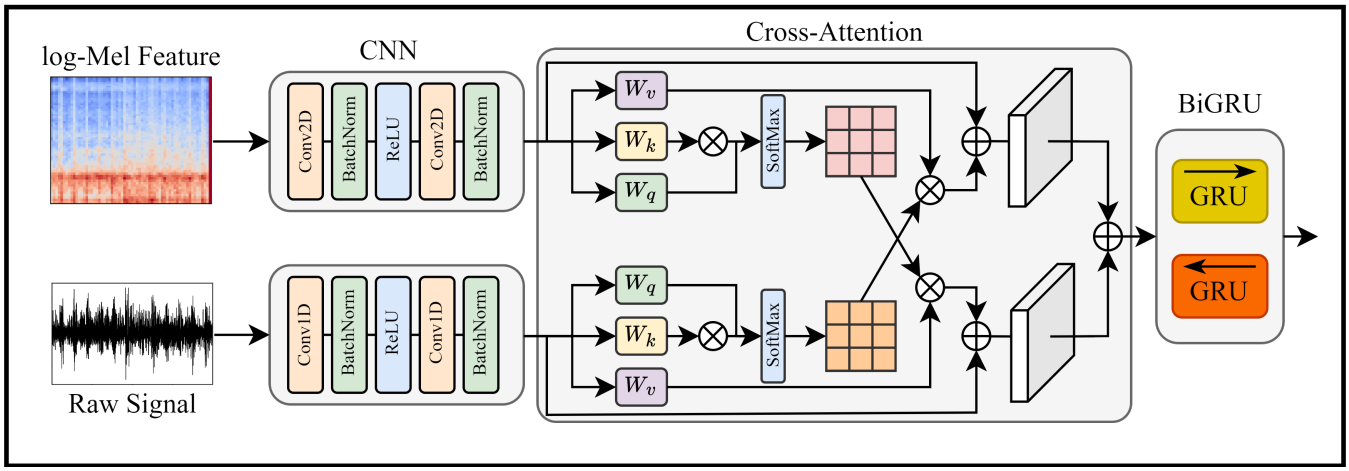


Figure 2. Architecture of the proposed DSCA-CRNN backbone. The network integrates dual-stream convolutional feature extractors, a cross-stream attention fusion module, and a bidirectional GRU-based temporal encoder with multi-level dropout regularization. The projection head is used only during training to compute the triplet objective; inference uses a single shared backbone and the classification head.

time-frequency representation:

$$\mathbf{X}_{mel} = \text{Concat}[\text{Mel}_{32},\ \text{Mel}_{64},\ \text{Mel}_{128}] \in \mathbb{R}^{3\times H\times W}.$$

In implementation, each 2-s audio window is sampled at 16 kHz and therefore contains 32,000 waveform samples before feature extraction. The three Mel maps are computed using $(32, 512, 256)$, $(64, 1024, 512)$, and $(128, 2048, 1024)$ for $(n_{\text{mels}}, n_{\text{fft}}, \text{hop})$, respectively, and each map is resized to 64×64. Thus, the Mel branch input is $\mathbf{X}_{mel} \in \mathbb{R}^{3\times 64\times 64}$. A standard STFT representation could also be used for mechanical signals; however, the log-Mel representation provides a compact, smoothed frequency-energy description that is widely used in acoustic modeling and reduces the dimensionality of high-frequency bins while preserving multi-scale spectral patterns (Gong et al., 2021; Z. Li et al., 2024). To avoid relying solely on this frequency-warped representation, the raw waveform branch is retained as a complementary time-domain path.

While Mel spectrograms encode the long-term spectral distribution of acoustic energy, they inevitably lose fine-grained temporal cues due to windowing and time-frequency averaging. To preserve these transient dynamics, the raw waveform $\mathbf{x}(t)$ is employed as a second feature channel. The raw signal retains short-duration impulses, impact signatures, and micro-vibration patterns that often reflect early-stage mechanical degradation but are difficult to capture from Mel features alone. For the raw branch, the same 2-s window is resampled to 2048 Hz, producing a fixed-length input $\mathbf{X}_{raw} \in \mathbb{R}^{1\times 4096}$. This lower-rate raw representation keeps the temporal waveform structure while limiting computational cost.

By jointly leveraging multi-scale Mel spectrograms and raw waveform inputs, the proposed dual-channel representation provides a richer and more discriminative description of conveyor idler acoustics. The Mel channel captures global frequency structures across multiple resolutions, while the raw channel preserves high-resolution temporal information. This complementary design enables the subsequent DSCA-CRNN backbone to simultaneously exploit stable spectral patterns and transient time-domain signatures, thereby improving discriminability under complex field conditions.

2.2.2. Backbone

To effectively learn discriminative acoustic representations from dual-channel temporal-spectral inputs, a Dual-Stream Cross-Attention Convolutional Recurrent Neural Network (DSCA-CRNN) is designed as the backbone of the proposed framework. The architecture integrates convolutional, attention, and recurrent components to jointly capture local spectral textures, high-resolution temporal structures, and long-range acoustic dependencies. Fig. 2 illustrates the conceptual architecture.

The DSCA-CRNN backbone consists of three functional stages: (1) dual-stream convolutional feature extraction, (2) cross-stream attention fusion, and (3) bidirectional recurrent temporal modeling.

(1) Dual-stream convolutional feature extraction Two independent feature extraction branches are adopted to process the multi-scale log-Mel spectrogram and the raw waveform, respectively. The *Mel stream* receives a three-channel Mel representation, where each channel corresponds to a Mel spectrogram computed under a different analysis resolution. A 2D CNN is used to extract spectral-temporal textures from this stacked Mel input.

The *raw stream* processes the waveform directly using a 1D CNN, preserving high-resolution temporal signatures such as transient impacts and micro-vibration patterns that may be smoothed out by time-frequency transformations.

Each stream applies two convolutional blocks (2D for Mel, 1D for raw), each consisting of convolution, batch normalization, ReLU, max pooling, and dropout ($p = 0.2$). Let

$$\mathbf{X}_{mel} \in \mathbb{R}^{3\times H\times W}, \qquad \mathbf{X}_{raw} \in \mathbb{R}^{1\times T},$$

denote the input Mel tensor and raw waveform. The extracted feature maps are given by

$$\mathbf{F}_{mel} = f_{cnn}^{2D}(\mathbf{X}_{mel}), \qquad \mathbf{F}_{raw} = f_{cnn}^{1D}(\mathbf{X}_{raw}), \quad (2)$$

where f_{cnn}^{2D} and f_{cnn}^{1D} denote the Mel-stream and raw-stream convolutional encoders. In single-stream configurations, only one branch is used; in fusion mode, both branches operate concurrently.

(2) Cross-attention fusion module To integrate complementary information from the Mel and raw streams, a dual-path Cross-Attention Fusion (CAF) module is introduced. The convolutional feature maps $\mathbf{F}_{mel}$ and $\mathbf{F}_{raw}$ are first flattened along the spatial (Mel) or temporal (raw) dimension to form sequence embeddings:

$$\mathbf{S}_{mel} = \mathrm{Flatten}(\mathbf{F}_{mel}), \qquad \mathbf{S}_{raw} = \mathrm{Flatten}(\mathbf{F}_{raw}),$$

where $\mathbf{S}_{mel}, \mathbf{S}_{raw} \in \mathbb{R}^{B\times L\times D}$ denote the Mel and raw feature sequences with unified length L and embedding dimension D.

Cross-attention is applied bidirectionally to model inter-stream dependencies:

$$\mathrm{Attn}_{mel\to raw} = \mathrm{Softmax}\left(\frac{Q_{mel}K_{raw}^T}{\sqrt{d}}\right), \quad (3)$$

$$\mathrm{Attn}_{raw\to mel} = \mathrm{Softmax}\left(\frac{Q_{raw}K_{mel}^T}{\sqrt{d}}\right), \quad (4)$$

followed by attended feature aggregation:

$$\mathbf{A}_{mel} = \mathrm{Attn}_{raw\to mel}\, V_{raw}, \quad (5)$$

$$\mathbf{A}_{raw} = \mathrm{Attn}_{mel\to raw}\, V_{mel}. \quad (6)$$

The fused representation is generated via concatenation or a learnable gating projection:

$$\mathbf{S}_{fused} = g([\mathbf{A}_{mel},\, \mathbf{A}_{raw}]), \quad (7)$$

where $g(\cdot)$ denotes a 1×1 convolution or gating unit. This module enables the network to dynamically emphasize the most informative cross-stream dependencies.

(3) Temporal modeling and classification The fused sequence is fed into a bidirectional gated recurrent unit (Bi-GRU) to capture long-range temporal dependencies:

$$\mathbf{H} = \mathrm{BiGRU}(\mathbf{S}_{fused}), \quad (8)$$

where $\mathbf{H}$ denotes the encoded temporal sequence. The final hidden state is regularized by dropout ($p = 0.3$) and passed into two parallel heads: a classification head for idler condition recognition, and a projection head optimized with the triplet loss to enhance inter-class separability.

Overall, the DSCA-CRNN backbone offers a unified temporal-spectral modeling framework that integrates convolutional feature extraction, cross-stream attention fusion, and recurrent temporal encoding.

2.2.3. Loss Function

The model is jointly optimized using a hybrid objective that combines a class-weighted focal loss for classification and a triplet-based contrastive loss for feature discrimination. Focal loss reduces the dominance of easy majority-class samples, while class weights compensate for the highly uneven label distribution (Lin et al., 2017). The triplet objective regularizes the projection head by encouraging within-class compactness and between-class separation (Schroff et al., 2015). The overall goal is to improve both decision accuracy and embedding separability under imbalanced conditions.

For a given training sample $\mathbf{x}_i$ with ground-truth label $y_i \in \{1, 2, \ldots, C\}$, the network produces class logits $\mathbf{p}_i = f_\theta(\mathbf{x}_i)$ and a latent embedding vector $\mathbf{z}_i$ from the penultimate layer. The class probabilities are obtained via a softmax operation:

$$\hat{y}_{i,c} = \frac{\exp(p_{i,c})}{\sum_{j=1}^{C}\exp(p_{i,j})}. \quad (9)$$

The classification loss is defined as a class-weighted focal loss:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c(1-\hat{y}_{i,c})^\gamma \tilde{y}_{i,c}\log(\hat{y}_{i,c}), \quad (10)$$

where w_c is the class weight inversely proportional to the frequency of class c, γ is the focal factor, and $\tilde{y}_{i,c}$ denotes the label-smoothed target. In this study, $\gamma = 2.0$ and the label-smoothing factor is 0.05.

To further enhance inter-class separability in the latent space, a triplet contrastive loss $\mathcal{L}_{tri}$ is incorporated. Each training batch forms anchor–positive–negative triplets $(\mathbf{z}_a, \mathbf{z}_p, \mathbf{z}_n)$, where the anchor and positive belong to the same class, and the negative belongs to a different class. The triplet loss enforces

a margin m between intra-class and inter-class distances:

$$\mathcal{L}_{tri} = \frac{1}{K}\sum_{k=1}^{K} \max\left(0, \|\mathbf{z}_a^k - \mathbf{z}_p^k\|_2^2 - \|\mathbf{z}_a^k - \mathbf{z}_n^k\|_2^2 + m\right). \quad (11)$$

This term encourages embeddings from the same class to cluster closely, while pushing apart samples from different classes, resulting in a more discriminative feature manifold.

The total loss function combines these two objectives as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha\,\mathcal{L}_{tri}, \quad (12)$$

where α is a balance coefficient controlling the relative contribution of the contrastive component. In this study, α is empirically set to 0.25 based on validation performance, and the triplet margin is set to $m = 1.0$.

2.2.4. Training-Time Targeted Gaussian Noise Perturbation

To further enhance the diversity of scarce fault samples, a targeted Gaussian noise-perturbation strategy is applied to the raw waveform branch during training. The present implementation injects zero-mean Gaussian white noise directly into the raw waveform as a noise-based regularizer (Bishop, 1995), while leaving the Mel branch unchanged. This differs from SpecAugment-style masking (Park et al., 2019), which masks the Mel representation and is included only as an ablation comparator rather than in the final proposed configuration. For a selected minority-class raw waveform tensor $\mathbf{X}_{raw}$, the perturbed representation is written as

$$\tilde{\mathbf{X}}_{raw} = \mathbf{X}_{raw} + \beta \cdot \mathcal{N}(0, \sigma^2), \quad (13)$$

where β controls the perturbation strength and σ is the noise standard deviation. In the implementation, the effective Gaussian noise scale is set to 0.01.

The perturbation is applied only during training to minority-class samples, namely warning and severe-fault conditions, to reduce overfitting caused by limited fault data. This targeted augmentation expands their intra-class variability while preserving the original class labels. It should be interpreted as a class-imbalance-oriented regularization mechanism rather than as a complete denoising solution or controlled noise-robustness test. In particular, it does not replace controlled SNR stress testing with recorded industrial noise, which is identified as an important direction for future work.

3. RESULTS AND DISCUSSION

3.1. Dataset

The dataset used in this study was collected from an operational industrial conveyor-belt system under normal production conditions. The audio recordings were acquired near the idler rollers during normal production operation, allowing close-range acquisition of the emitted sound while reducing interference from surrounding equipment (see Fig. 3). This acquisition protocol reflects a practical auscultation-style inspection setting, but it also means that the present dataset does not cover all possible microphone placements or intentionally degraded low-SNR conditions.

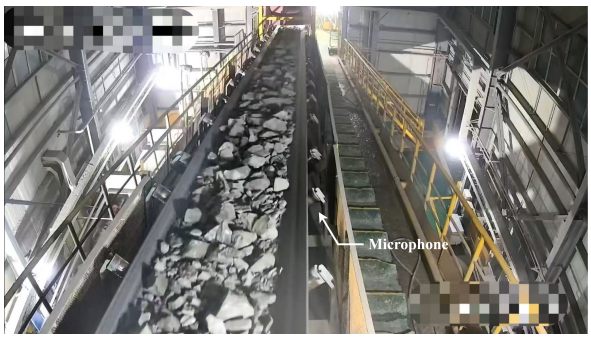


Figure 3. Industrial conveyor-belt environment in which acoustic signals were collected.

Each raw audio recording lasted 10 seconds and was sampled at 16 kHz. To increase the number of training samples and to capture the temporal characteristics more effectively, all recordings were segmented into non-overlapping 2-second clips. Each clip therefore contains 32,000 samples before feature extraction; after preprocessing, the model receives a $3 \times 64 \times 64$ multi-scale Mel tensor and a 1×4096 resampled raw waveform tensor. After preprocessing, the final dataset contained 2,495 samples, including 1,615 normal, 640 warning, and 240 severe-fault segments.

To ensure a fair evaluation, we first selected 20% of the samples as an independent test set. The selection was performed at the recording level to prevent clips originating from the same 10-second audio from appearing in different subsets. The remaining 80% of the data was then used to conduct five-fold cross-validation, where in each fold, one subset was used for validation and the remaining four subsets were used for training. This evaluation protocol enables a robust and unbiased assessment of the model's generalization performance.

3.2. Performance Over Related Works

To comprehensively evaluate the effectiveness of the proposed DSCA-CRNN framework, its performance was compared with a set of representative baseline and state-of-the-art methods for acoustic-based fault diagnosis, as summarized in Table 1. These methods include traditional machine learning classifiers (SVM and RF), convolutional neural networks (ResNet, CRNN, AudioMamba and IMPACT), attention-based transformer models (AST, FedFormer, PatchTST, TimesNet), and single-stream variants of the proposed architecture.

Accuracy measures the proportion of correctly classified samples over the entire test set. Because the dataset is imbalanced,

Table 1. Performance comparison between the proposed DSCA-CRNN and related methods on the conveyor acoustic dataset.

Method	Input Representation	Accuracy	Macro-F1	F1 (Normal)	F1 (Warning)	F1 (Severe Fault)
SVM	Log-Mel spectrogram	0.862±0.008	0.739±0.012	0.962±0.003	0.719±0.019	0.536±0.021
RF	Log-Mel spectrogram	0.849±0.007	0.689±0.021	0.940±0.004	0.721±0.013	0.406±0.056
ResNet (He et al., 2016)	Log-Mel spectrogram	0.928±0.009	0.843±0.019	0.991±0.003	0.863±0.017	0.675±0.044
CRNN (Shi et al., 2016)	Log-Mel spectrogram	0.931±0.009	0.850±0.017	0.992±0.004	0.867±0.017	0.691±0.037
AST (Gong et al., 2021)	Log-Mel spectrogram	0.902±0.015	0.817±0.025	0.973±0.008	0.811±0.030	0.665±0.046
AudioMamba (Erol et al., 2024)	Log-Mel spectrogram	0.896±0.007	0.796±0.018	0.976±0.006	0.804±0.007	0.609±0.049
IMPACT (Han et al., 2025)	Log-Mel spectrogram	0.925±0.013	0.860±0.028	0.980±0.003	0.858±0.025	0.743±0.058
FedFormer (Zhou et al., 2022)	Raw	0.785±0.017	0.643±0.035	0.902±0.008	0.590±0.034	0.436±0.116
PatchTST (Nie et al., 2022)	Raw	0.820±0.016	0.724±0.032	0.916±0.009	0.650±0.054	0.604±0.069
TimesNet (H. Wu et al., 2022)	Raw	0.702±0.026	0.532±0.027	0.846±0.020	0.441±0.066	0.309±0.058
DSCA-CRNN	Log-Mel spectrogram	0.946±0.016	0.873±0.043	0.986±0.004	0.896±0.030	0.748±0.111
DSCA-CRNN	Raw	0.852±0.042	0.740±0.067	0.948±0.021	0.712±0.069	0.561±0.123
DSCA-CRNN	**Log-Mel + Raw waveform**	**0.952±0.011**	**0.900±0.026**	**0.992±0.001**	**0.905±0.022**	**0.803±0.058**

macro-F1 is also reported as the unweighted average of the class-specific F1-scores, giving equal importance to normal, warning, and severe-fault classes. Class-specific F1-scores are further included to show whether minority fault categories are recognized reliably. Unless otherwise stated, the tabulated results are reported from five-fold cross-validation.

All comparison models were trained and evaluated using the same data partitioning protocol and consistent evaluation metrics. Fold-level summaries are reported when complete, and hyperparameters were selected based on validation performance to ensure a fair comparison.

Traditional machine learning models achieve moderate performance, with SVM and RF obtaining macro-F1 scores of 0.739 and 0.689, respectively. Although they perform well on the majority normal class, their performance on severe faults remains limited, indicating insufficient sensitivity to minority fault patterns.

Deep learning models based on log-Mel spectrograms show clear performance improvements. ResNet and CRNN achieve macro-F1 scores of 0.843 and 0.850, respectively, while AST reaches 0.817. These models effectively capture spectral characteristics, leading to improved recognition of warning and severe-fault conditions. Among the more recent baselines, IMPACT obtains a macro-F1 of 0.860 and a severe-fault F1-score of 0.743, showing the benefit of pretrained audio representations. AudioMamba reaches a macro-F1 of 0.796. In contrast, the raw-waveform baselines FedFormer, PatchTST, and TimesNet obtain lower macro-F1 scores than the log-Mel models, suggesting that direct raw-sequence modeling without explicit acoustic spectral inductive bias is insufficient for this conveyor auscultation task.

The proposed DSCA-CRNN further improves performance across all metrics. When using log-Mel spectrograms alone, DSCA-CRNN achieves a macro-F1 of 0.873, outperforming all baseline models. The raw-only variant achieves a macro-F1 of 0.740, confirming that raw temporal features contain useful fault information but are less discriminative when used independently.

By integrating both log-Mel spectrograms and raw waveform representations, the full dual-channel DSCA-CRNN achieves the best overall performance, with an accuracy of 0.952 and a macro-F1 score of 0.900. Notably, the F1-score for severe faults reaches 0.803, which is substantially higher than all comparison methods. The warning-class F1-score of 0.905 further demonstrates enhanced sensitivity to early-stage faults.

3.3. Ablation Study

To investigate the contribution of each key component in the proposed DSCA-CRNN framework, a comprehensive ablation study was conducted. Specifically, we evaluated the effects of recurrent temporal modeling (Bi-GRU), different dual-branch fusion strategies, the contrastive loss term, and training-time raw-waveform perturbation. All ablation models were trained and evaluated under the same experimental protocol, including identical data splits, five-fold cross-validation, and evaluation metrics, ensuring a fair comparison.

3.3.1. Effect of recurrent temporal modeling

To examine the role of recurrent temporal modeling, we compared the full model (with Bi-GRU) against a no-RNN variant that removes the Bi-GRU module and replaces sequence modeling with a simple temporal pooling strategy. Specifically, the no-RNN variant aggregates the feature sequence using the average of mean and max pooling over the temporal dimension, which provides a lightweight alternative but does not explicitly model long-range temporal dependencies.

Table 2. Effect of removing Bi-GRU and using temporal pooling.

Temporal Module	Accuracy	Macro-F1
Temporal pooling	0.903±0.015	0.803±0.029
Bi-GRU	**0.952±0.011**	**0.900±0.026**

As shown in Table 2, incorporating Bi-GRU leads to a noticeable performance improvement over simple temporal pooling, with accuracy increasing from 0.903 to 0.952 and macro-F1 improving substantially from 0.803 to 0.900.

While pooling-based aggregation summarizes global statistics of the feature sequence, it discards temporal ordering information and cannot model the dynamic evolution of acoustic patterns. In contrast, Bi-GRU explicitly captures bidirectional temporal dependencies across frames, enabling the model to better exploit gradual changes and transient fault-related cues embedded in the signal.

The larger improvement in macro-F1 indicates that recurrent temporal modeling is particularly beneficial for minority fault categories, where discriminative patterns often unfold progressively rather than appearing as isolated spectral events. This result confirms that sequential dependency modeling plays a crucial role in enhancing class-level discrimination under imbalanced industrial conditions. A limitation is that the Bi-GRU is bidirectional and therefore requires the complete 2-s input window before producing a prediction; the model is suitable for window-based offline or near-real-time diagnosis, but not for causal sample-by-sample streaming.

3.3.2. Effect of fusion method

To assess the effectiveness of the cross-stream attention fusion module, we compared the proposed cross-attention fusion with three alternative strategies: simple feature concatenation, gating-based fusion, and a joint self-attention baseline implemented by applying a Transformer encoder to concatenated Mel and raw tokens. In this baseline, every concatenated token attends to all tokens through standard self-attention; unlike cross-attention, it does not impose directed query–key/value roles between the two streams.

Table 3. Effect of different feature fusion strategies.

Fusion Strategy	Accuracy	Macro-F1
Concatenation	0.927±0.010	0.841±0.020
Gating	0.916±0.006	0.817±0.015
Joint Self-Attention	0.934±0.013	0.856±0.025
Cross-attention	**0.952±0.011**	**0.900±0.026**

As shown in Table 3, different fusion strategies lead to noticeable performance variations. The simple concatenation strategy achieves an accuracy of 0.927 and a macro-F1 score of 0.841, indicating that combining dual-channel features without explicit interaction modeling already provides benefits over single-stream configurations. However, this approach treats the two feature streams independently and relies on subsequent layers to implicitly learn cross-channel relationships, which may limit effective information exchange.

The gating-based fusion slightly underperforms concatenation, with a macro-F1 of 0.817. This suggests that a simple learnable weighting mechanism may not be sufficient to capture complex temporal–spectral dependencies in industrial acoustic signals, especially under severe class imbalance.

The joint self-attention variant improves cross-stream interaction compared with gating, but remains below the proposed cross-attention design. In contrast, the proposed cross-attention fusion significantly improves performance, achieving the highest accuracy of 0.952 and a macro-F1 of 0.900. The absolute improvement of 0.059 in macro-F1 compared with concatenation demonstrates that modeling inter-stream dependencies before recurrent temporal encoding enables the network to dynamically align and emphasize complementary temporal and spectral cues. This is particularly beneficial for minority fault categories, where subtle acoustic variations require fine-grained cross-modal interaction.

These results confirm that effective feature fusion is not merely a matter of combining representations, but requires structured interaction modeling. The proposed cross-attention mechanism provides a more expressive and adaptive fusion strategy, leading to superior discriminability under imbalanced industrial conditions.

3.3.3. Effect of noise-perturbation augmentation

The next set of experiments compares no augmentation, targeted Gaussian noise perturbation, and SpecAugment masking on the original close-range test set. All configurations use the same Cross-Attention DSCA-CRNN backbone. During training, Gaussian perturbation is applied to minority-class raw waveform inputs, whereas SpecAugment masks minority-class Mel representations. They are evaluated as imbalance-oriented regularizers and should not be interpreted as controlled noise-robustness tests.

Table 4. Effect of training-time augmentation strategies

Augmentation	F1 (Normal)	F1 (Warning)	F1 (Severe Fault)
w/o	0.992±0.006	0.879±0.034	0.744±0.047
SpecAugment masking	0.992±0.004	0.885±0.032	0.770±0.055
Gaussian noise	**0.992±0.001**	**0.905±0.022**	**0.803±0.058**

As shown in Table 4, both augmentation methods preserve normal-class performance at 0.992, while targeted Gaussian noise perturbation gives the highest F1-scores for the imbalanced fault categories. Relative to no augmentation, Gaussian perturbation improves the warning-class F1 from 0.879 to 0.905 and the severe-fault F1 from 0.744 to 0.803. It also outperforms SpecAugment masking, which obtains F1-scores of 0.885 and 0.770 for the warning and severe-fault classes, respectively. These results suggest that additive perturbations of minority-class raw waveforms better regularize scarce fault samples in this setting, while retaining the discriminative structure needed for fault-level separation. Therefore, the final configuration uses targeted Gaussian perturbation as a class-imbalance-oriented regularizer rather than as evidence of general noise robustness; robustness to stronger background interference should still be validated in future work using test sets corrupted at controlled SNR levels with recorded industrial noise.

Table 5. Effect of the triplet contrastive loss.

Triplet Weight	Accuracy	Macro-F1	F1 (Normal)	F1 (Warning)	F1 (Severe Fault)
$\alpha = 0$	0.949±0.007	0.890±0.013	**0.993±0.002**	0.901±0.014	0.777±0.029
$\alpha = 0.05$	0.950±0.004	0.883±0.008	0.992±0.004	0.898±0.008	0.774±0.022
$\alpha = 0.10$	0.946±0.019	0.895±0.037	0.992±0.004	0.894±0.038	0.787±0.074
$\alpha = 0.15$	0.940±0.015	0.872±0.027	0.993±0.004	0.883±0.033	0.738±0.046
$\alpha = 0.20$	0.939±0.016	0.867±0.033	0.992±0.004	0.883±0.032	0.726±0.067
$\boldsymbol{\alpha = 0.25}$	**0.952±0.011**	**0.900±0.026**	0.992±0.001	**0.905±0.022**	**0.803±0.058**
$\alpha = 0.30$	0.937±0.014	0.866±0.025	0.993±0.005	0.876±0.030	0.731±0.048

3.3.4. Effect of triplet contrastive loss

To quantify the contribution of the triplet-based contrastive objective, we evaluated different values of the triplet loss weight α while keeping the same dual-channel inputs, cross-attention fusion, and Bi-GRU temporal encoder.

As shown in Table 5, the best overall performance is obtained at $\alpha = 0.25$, where macro-F1 increases from 0.890 to 0.900 and severe-fault F1 increases from 0.777 to 0.803 compared with $\alpha = 0$. The gains are modest in overall accuracy because normal samples are already recognized reliably, but the improvement on severe faults indicates that the contrastive projection head helps shape a more discriminative minority-class embedding space. Larger or smaller weights do not provide further improvement, suggesting that a moderate contrastive regularization strength is preferable for this dataset.

3.4. Computational Cost and Deployment Scope

Because practical condition monitoring requires not only high accuracy but also manageable computational cost, the computational profile of the proposed model is summarized in Table 6. The inference cost was measured using batch size 1, and the reported time corresponds to the neural-network forward pass after feature tensors have been prepared.

Table 6. Computational profile of the proposed DSCA-CRNN.

Item	Value
Trainable parameters	1.39M
Checkpoint size	5.33 MB
Input window	2.0 s
Mel input size	$3 \times 64 \times 64$
Raw input size	1×4096
Forward time, RTX 4090 GPU	1.32 ms/sample
Algorithmic latency	one complete 2.0-s window

The parameter count and inference time indicate that the network is lightweight enough for window-based diagnostic use. However, the bidirectional GRU uses both past and future frames inside the 2-s window, so the method should be regarded as an offline or near-real-time window-level diagnostic model rather than a causal streaming detector. Future work will investigate causal recurrent or temporal-convolutional variants for lower-latency online monitoring.

3.5. Confusion Matrix Analysis

To further analyze the classification behavior of the proposed DSCA-CRNN, the confusion matrix on the held-out test partition is presented in Fig. 4.

As shown in Fig. 4, the confusion matrix exhibits strong diagonal dominance, indicating that most samples are correctly classified across all operational states.

For the normal condition, 321 out of 323 samples are correctly identified, with only two samples misclassified as warning and none misclassified as severe. This demonstrates high reliability in recognizing healthy operation while avoiding false severe-fault warnings.

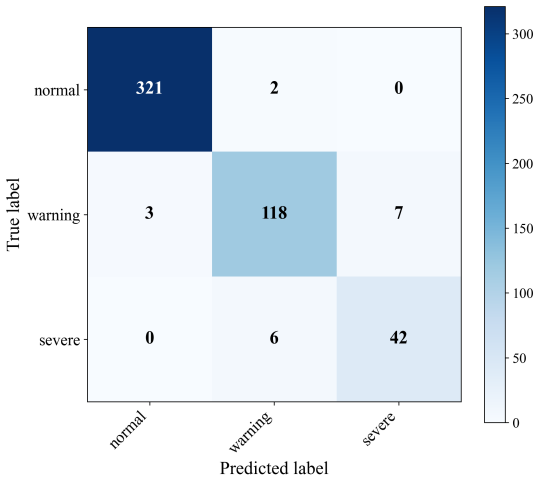


Figure 4. Confusion matrix of the proposed DSCA-CRNN on the test set.

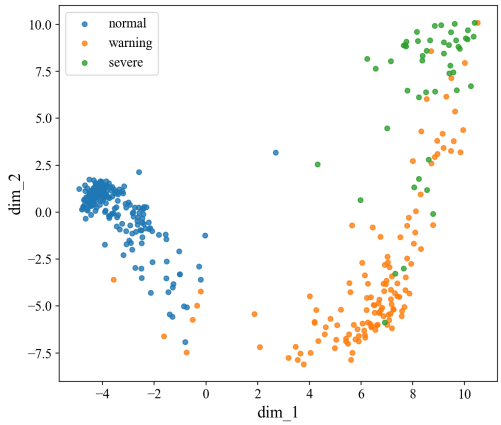
For the warning category, 118 out of 128 samples are correctly classified. A small number of warning samples are misclassified as normal (3 cases) or severe (7 cases), which can be attributed to the transitional and ambiguous acoustic characteristics of early-stage faults.

Importantly, for severe faults, 42 out of 48 samples are correctly identified, yielding a high recall rate. Notably, no severe samples are misclassified as normal, and no normal samples are predicted as severe. This absence of critical cross-level confusion significantly reduces the risk of missed severe faults

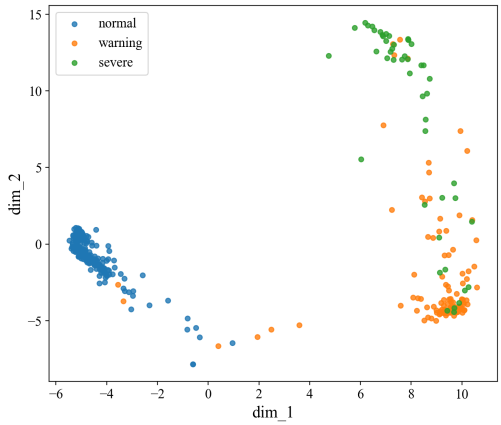
or false critical warnings, which is essential for safety-critical industrial applications.

3.6. Feature Space Visualization

To further analyze the influence of contrastive learning on representation structure, the feature embeddings extracted from the penultimate layer were projected into a two-dimensional space using Principal Component Analysis (PCA), as shown in Fig. 5. For comparison, Fig. 5(a) illustrates the embedding distribution without contrastive loss ($\alpha = 0$), while Fig. 5(b) shows the result with contrastive loss weight $\alpha = 0.25$. Silhouette score (Rousseeuw, 1987) is further used as an objective cluster-separation metric.



(a) Without contrastive loss ($\alpha = 0$)



(b) With contrastive loss ($\alpha = 0.25$)

Figure 5. PCA projection of learned feature embeddings under different contrastive loss weights.

As observed in Fig. 5(a), when the contrastive component is disabled, the feature clusters corresponding to different operational conditions exhibit partial overlap, particularly between warning and severe states. The intra-class dispersion of warning samples is relatively large, and the decision boundaries between adjacent fault levels appear less distinct.

Table 7. Silhouette scores of learned embeddings under different contrastive loss weights on the fold-0 test embeddings. Higher values indicate more compact and better separated clusters.

Feature Space	$\alpha = 0$	$\alpha = 0.25$
64-D projection embedding	0.568	**0.742**
2-D PCA projection	0.703	**0.759**

After introducing contrastive learning (Fig. 5(b)), the feature space becomes more structured. Intra-class clusters are noticeably more compact, while inter-class margins are significantly enlarged. The severe class forms a well-separated cluster with minimal overlap with normal samples, and the warning class exhibits clearer separation from both normal and severe conditions. This indicates that the contrastive objective effectively reduces intra-class variance while increasing inter-class discrimination.

These visualization results provide intuitive evidence that contrastive learning reshapes the embedding manifold into a more separable structure. The silhouette scores in Table 7 provide a quantitative complement: in the original 64-D projection space, the score increases from 0.568 to 0.742 after introducing the triplet objective. By explicitly enforcing relative distance constraints in the latent space, the model achieves improved feature compactness and class-level discrimination, which is consistent with the quantitative performance gains observed in previous sections.

4. CONCLUSION

This study presented a practical auscultation-based acoustic diagnosis framework for mining conveyor idler condition monitoring under imbalanced field-data conditions. A dual-channel temporal–spectral representation was constructed by jointly leveraging multi-scale log-Mel spectrograms and raw waveforms. On top of this representation, a Dual-Stream Cross-Attention CRNN (DSCA-CRNN) was developed to capture complementary spectral patterns and fine-grained temporal dynamics, while explicitly modeling cross-stream dependencies through a cross-attention fusion module. In addition, class-weighted focal learning and a triplet-based contrastive objective were integrated to improve feature discriminability, especially for minority fault categories.

Experiments on a self-collected close-range conveyor auscultation dataset demonstrate the effectiveness of the proposed approach. Compared with traditional machine learning and deep learning baselines, DSCA-CRNN achieves the best overall performance, with an accuracy of 0.952 and a macro-F1 score of 0.900. Notably, the proposed method substantially improves severe-fault recognition (F1 = 0.803), which is critical for safety-related industrial applications. Ablation studies

further verify that recurrent temporal modeling with Bi-GRU enhances sequential dependency learning compared with simple temporal pooling, that cross-attention provides a clear advantage over concatenation, gating, and joint self-attention variants in modeling inter-stream interactions, and that the triplet objective improves severe-fault F1. The confusion matrix analysis confirms that misclassifications mainly occur between adjacent fault levels and that no severe samples are predicted as normal. Moreover, PCA-based feature visualization and silhouette scores show that introducing the contrastive objective yields more compact intra-class clusters and larger inter-class margins.

Despite these promising results, several limitations remain. First, the dataset was collected from a specific conveyor system using close-range auscultation, and broader generalization to different sites, microphone placements, conveyor configurations, and deliberately lower-SNR conditions requires further investigation. Second, the present targeted Gaussian perturbation experiment evaluates regularization on the original test distribution and does not constitute a controlled noise-robustness validation. Future work should include stress tests using recorded industrial noise mixed at controlled SNR levels and comparisons with signal-level denoising or realistic noise-mixing baselines. Third, the Bi-GRU requires a complete 2-s window and therefore supports window-level offline or near-real-time diagnosis rather than causal streaming. Incorporating finer-grained fault taxonomy, multi-site validation, domain adaptation, and causal lightweight architectures will be important directions for future research.

DECLARATION OF CONFLICTING INTEREST

The author declares that there is no conflict of interest regarding the publication of this paper.

FUNDING

This research did not receive any specific grant from funding agencies.

DATA AVAILABILITY

The data and trial information supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Alharbi, F., Luo, S., Zhang, H., Shaukat, K., Yang, G., Wheeler, C. A., & Chen, Z. (2023). A brief review of acoustic and vibration signal-based fault detection for belt conveyor idlers using machine learning models. *Sensors*, *23*(4), 1902.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural computation*, *7*(1), 108–116.

Bzinkowski, D., Rucki, M., Chalko, L., Kilikevicius, A., Matijosius, J., Cepova, L., & Ryba, T. (2024). Application of machine learning algorithms in real-time monitoring of conveyor belt damage. *Applied Sciences*, *14*(22), 10464.

Chen, W., Shen, J., Zhu, H., & Xu, P. (2025). Interpretable fault diagnosis of coal-mine conveyor-belt idler bearings using multi-source fusion diagrams and mobile residual soft threshold–mobilevit. *Ain Shams Engineering Journal*, *16*(12), 103777.

Erol, M. H., Senocak, A., Feng, J., & Chung, J. S. (2024). Audio mamba: Bidirectional state space model for audio representation learning. *IEEE Signal Processing Letters*, *31*, 2975–2979.

Farhat, M. H., Gelman, L., Abdullahi, A. O., Ball, A., Conaghan, G., & Kluis, W. (2023). Novel fault diagnosis of a conveyor belt mis-tracking via motor current signature analysis. *Sensors*, *23*(7), 3652.

Gong, Y., Chung, Y.-A., & Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Han, C., Sim, Y., Jung, H., Lee, J., Lee, H., Kang, Y. S., … Jun, M. B.-G. (2025). Impact: Industrial machine perception via acoustic cognitive transformer. *arXiv preprint arXiv:2507.06481*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Jiao, J., Zhao, M., Lin, J., & Liang, K. (2020). A comprehensive review on convolutional neural network in machine fault diagnosis. *Neurocomputing*, *417*, 36–63.

Khalifa, R. M., Yacout, S., Bassetto, S., & Shaban, Y. (2025). Condition monitoring and warning of a belt drive system based on a logical analysis of data regression-based residual control chart. *Structural Health Monitoring*, *24*(3), 1657–1673.

Kim, J., Yi, I., & Suh, Y.-J. (2025). Domain generalized open-set fault detection and diagnosis for belt conveyor systems with prototype learning. *IEEE Access*.

Lei, Y., Jia, F., Lin, J., Xing, S., & Ding, S. X. (2016). An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Transactions on Industrial Electronics*, *63*(5), 3137–3147.

Li, X., Wu, D., Liu, Y., & Chen, Y. (2024). Belt conveyor idler fault diagnosis method based on multi-scale feature fusion and residual mask convolution attention. *Insight-Non-Destructive Testing and Condition Monitoring*, *66*(2), 82–93.

Li, Z., Wang, H., Liang, W., & Yao, L. (2024). Audio fault diagnosis of belt conveyors based on improved variational modal decomposition and improved adaptive noise reduction convolutional network in strong noise environments. *Measurement Science and Technology*,

35(10), 106126.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Liu, J., Fu, S., Liu, F., & Cheng, X. (2025). Intelligent fault diagnosis of belt conveyor rollers using a polar knn algorithm with audio features. *Engineering Failure Analysis*, *168*, 109101.

Liu, Y., Miao, C., Li, X., Ji, J., Meng, D., & Wang, Y. (2023). A dynamic self-attention-based fault diagnosis method for belt conveyor idlers. *Machines*, *11*(2), 216.

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Qiu, S., Cui, X., Ping, Z., Shan, N., Li, Z., Bao, X., & Xu, X. (2023). Deep learning techniques in intelligent fault diagnosis and prognosis for industrial systems: A review. *Sensors*, *23*(3), 1305.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 815–823).

Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, *39*(11), 2298–2304.

Soares, J. L., Costa, T. B., do Nascimento, G. S., Sousa, W. S., de Figueiredo, J. M., Braga, D. S., … Mesquita, A. L. (2025). Decision tree and anova as feature selection from vibration signals to improve the diagnosis of belt conveyor idlers. *Signals*, *6*(3), 42.

Soares, J. L., Costa, T. B., Moura, L. S., Sousa, W. S., Mesquita, A. L., Mesquita, A. L., … Braga, D. S. (2024). Fault diagnosis of belt conveyor idlers based on gradient boosting decision tree. *The International Journal of Advanced Manufacturing Technology*, *132*(7), 3479–3488.

Sung, S.-H., Hong, S., Choi, H.-R., Park, D.-M., & Kim, S. (2024). Enhancing fault diagnosis in iot sensor data through advanced preprocessing techniques. *Electronics*, *13*(16), 3289.

Wang, J., Mo, Z., Zhang, H., & Miao, Q. (2019). A deep learning method for bearing fault diagnosis based on time-frequency image. *IEEE Access*, *7*, 42373–42383.

Wang, M., Shen, K., Tai, C., Zhang, Q., Yang, Z., & Guo, C. (2023). Research on fault diagnosis system for belt conveyor based on internet of things and the lightgbm model. *Plos one*, *18*(3), e0277352.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2022). Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*.

Wu, X., Wang, C., Tian, Z., Huang, X., & Wang, Q. (2023). Research on belt deviation fault detection technology of belt conveyors based on machine vision. *Machines*, *11*(12), 1039.

Zhang, W., Li, C., Peng, G., Chen, Y., & Zhang, Z. (2018). A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical systems and signal processing*, *100*, 439–453.

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning* (pp. 27268–27286).