Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085769

ARTICLE

Lightweight Attention CNN–BiLSTM Model with Contrastive Regularization and Explainable Time–Frequency Analysis for Bearing Fault Diagnosis

Sherif Said[1,2,*], **Emrah Aslan**[3,4] **and Yıldırım Özüpak**[5]

[1]College of Engineering, University of Sharjah, Sharjah, United Arab Emirates
[2]College of Engineering and Technology, American University of the Middle East, Al Ahmadi, Kuwait
[3]Department of Computer Engineering, Faculty of Engineering, Dicle University, Diyarbakır, Turkey
[4]Department of Computer Engineering, Faculty of Engineering and Architecture, Mardin Artuklu University, Mardin, Turkey
[5]Silvan Vocational School, Dicle University, Diyarbakır, Turkey

*Corresponding Author: Sherif Said. Email: ssaid@sharjah.ac.ae

Received: 20 May 2026; Accepted: 03 August 2026; Published: 22 October 2026

ABSTRACT: Reliable bearing fault diagnosis is essential for predictive maintenance, operational safety, and the reduction of unexpected downtime in rotating machinery. However, developing robust data-driven fault diagnosis models remains challenging when only small-scale laboratory datasets are available, because high classification accuracy may be affected by overfitting, segment-level data leakage, and insufficient interpretability. To address these issues, this study proposes a leakage-aware lightweight attention Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN–BiLSTM) framework using Log-Mel spectrogram representations of rotating-machinery sensor signals. The proposed architecture combines lightweight convolutional blocks for local time–frequency feature extraction, bidirectional long short-term memory layers for sequential dependency modeling, an attention mechanism for emphasizing discriminative temporal–spectral regions, and supervised contrastive regularization to improve class separability under limited-data conditions. To reduce the risk of overly optimistic evaluation, the experiments were conducted using repeated group-aware validation, where segments derived from the same original recording or adjacent temporal regions were prevented from being simultaneously used in training and testing whenever permitted by the dataset structure. The proposed model achieved an average classification accuracy of 98.1%, with high precision, recall, and F1-score across four bearing conditions. In addition, Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients, and SHapley Additive exPlanations (SHAP) analyses were used to interpret the learned decision regions. The attribution maps showed that the model mainly focused on physically meaningful time–frequency zones associated with fault-induced spectral energy variations rather than irrelevant background components. The results demonstrate that the proposed framework provides an accurate, compact, and interpretable solution for bearing fault diagnosis under controlled laboratory conditions, while further validation on larger and noisier industrial datasets remains necessary.

KEYWORDS: Bearing fault diagnosis; predictive maintenance; Log-Mel spectrogram; lightweight CNN–BiLSTM; attention mechanism

1 Introduction

Rotating machinery plays a critical role in industrial production, transportation, energy systems, and automated manufacturing [1]. The reliable operation of such systems is essential for reducing downtime, maintenance costs, and safety risks. Among rotating components, rolling bearings are particularly important

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

because they support rotating shafts and directly affect the dynamic stability of machinery. Bearing faults caused by fatigue, lubrication degradation, misalignment, overload, or surface wear may gradually develop into severe mechanical failures if they are not detected at an early stage. Therefore, accurate and interpretable bearing fault diagnosis is a key requirement for predictive maintenance and intelligent condition monitoring.

Signal-based condition monitoring has been widely used for identifying bearing faults. Vibration signals are commonly preferred because of their high sensitivity to defect-induced impacts, while acoustic and acoustic-like sensor signals can provide complementary information about friction, impacts, and abnormal mechanical interactions. However, in publicly available rotating-machinery datasets, the exact sensing modality is sometimes not fully documented or may be described inconsistently as acoustic, acoustic emission, vibration, or generic time-series sensor data. To avoid overclaiming, this study formulates the problem as time–frequency-based bearing fault diagnosis using rotating-machinery sensor signals rather than presenting it as a purely acoustic or purely vibration-based system.

Bearing fault signals are generally non-stationary and impulsive. For this reason, time-domain or frequency-domain features alone may not fully represent fault-related behavior. Time–frequency representations such as short-time Fourier transform, continuous wavelet transforms, Mel-frequency representations, and Log-Mel spectrograms have therefore been widely used in fault diagnosis. These representations preserve both temporal evolution and spectral energy distribution, allowing learning models to capture localized fault signatures more effectively. In this study, Log-Mel spectrograms are used as compact time–frequency inputs because they provide a suitable representation of spectral energy variations associated with different bearing conditions.

Deep learning methods have significantly improved intelligent fault diagnosis by reducing the need for handcrafted features. Convolutional neural networks can extract local spatial patterns from spectrogram-like inputs, while long short-term memory networks can model temporal dependencies in sequential data. Hybrid CNN–LSTM and CNN–BiLSTM architectures have therefore been widely applied to fault diagnosis problems. However, the frequent use of CNN, LSTM, and spectrogram-based models also raises a novelty concern. A meaningful contribution should not only combine existing components but should also address practical problems such as limited data, data leakage risk, model overfitting, and lack of interpretability [2].

Recent studies have further explored contrastive learning, xLSTM-based models, Transformer architectures, and large language model-based approaches for mechanical fault diagnosis [3]. These methods have shown promising potential in improving feature separability, noise robustness, and cross-condition generalization [4]. Nevertheless, large or highly parameterized models generally require more extensive datasets, higher computational resources, or pretraining. When only a small laboratory dataset is available, a lightweight and well-regularized model may provide a more appropriate balance between diagnostic performance and generalization ability [5].

Despite recent progress, several challenges remain. First, high accuracy values reported on segmented signal datasets may be affected by data leakage if highly correlated segments from the same original recording are randomly assigned to both training and testing subsets. Second, deep learning models trained on small datasets may memorize segment-specific patterns instead of learning general fault characteristics. Third, many studies provide limited methodological detail regarding segmentation, normalization, noise reduction, model configuration, and validation strategy. Fourth, although explainable artificial intelligence methods are increasingly used, their visual outputs are not always connected to physically meaningful fault-related frequency or time–frequency regions.

To address these issues, this study proposes a leakage-aware lightweight attention CNN–BiLSTM framework with supervised contrastive regularization for bearing fault diagnosis. The proposed model uses

Log-Mel spectrograms as input and combines lightweight convolutional blocks, attention-based feature refinement, bidirectional temporal modeling, and contrastive embedding regularization. The convolutional layers extract local time–frequency patterns, the attention mechanism emphasizes discriminative spectral–temporal regions, and the BiLSTM layer captures sequential dependencies. Supervised contrastive regularization is further used to improve class separability under limited-data conditions.

In addition, a repeated group-aware validation strategy is adopted to reduce the risk of overly optimistic performance caused by segment-level data leakage. The final model performance is reported as an average over repeated validation runs rather than relying on a single random split. The proposed model achieves an average classification accuracy of 98.1%, outperforming conventional CNN, LSTM, CNN–LSTM, and CNN–BiLSTM baselines. Furthermore, Grad-CAM, Integrated Gradients, and SHAP are used to interpret the model predictions. The highlighted regions are discussed in relation to fault-related time–frequency energy patterns, supporting a physically meaningful interpretation of the learned decision mechanism.

The main contributions of this study are summarized as follows:

- A leakage-aware lightweight attention CNN–BiLSTM model is proposed for bearing fault diagnosis using Log-Mel spectrogram representations of rotating-machinery sensor signals.
- The conventional CNN–LSTM structure is extended with bidirectional temporal modeling, attention-based time–frequency refinement, and supervised contrastive regularization.
- A repeated group-aware validation protocol is used to reduce the risk of segment-level data leakage and to provide a more reliable performance estimate under limited-data conditions.
- The proposed model achieves an average accuracy of 98.1% and outperforms conventional CNN, LSTM, CNN–LSTM, and CNN–BiLSTM baseline models.
- Grad-CAM, Integrated Gradients, and SHAP are employed to interpret the model predictions, and the highlighted regions are linked to physically meaningful time–frequency energy variations.
- The limitations of the study are explicitly discussed, including the small dataset size, controlled laboratory conditions, uncertain sensing modality, and lack of external industrial validation.

Although CNN-, LSTM-, CNN–BiLSTM-, and spectrogram-based fault diagnosis methods have been widely investigated, three important gaps remain insufficiently addressed. First, many studies report high diagnostic accuracy without explicitly discussing whether segment-level samples extracted from the same original recording are separated at the file or group level, which may cause data leakage and overestimate model performance. Second, deep architectures are often applied to small laboratory datasets without sufficient regularization or repeated validation, making it difficult to distinguish true generalization from memorization. Third, most explainable artificial intelligence analyses remain qualitative and do not sufficiently connect highlighted regions with physically meaningful fault-related time–frequency patterns. Therefore, an effective bearing fault diagnosis framework should not only improve classification performance but also reduce data-leakage risk, remain lightweight under limited data, enhance embedding separability, and provide physically interpretable explanations.

The remainder of this paper is organized as follows. Section 2 reviews related studies on signal-based fault diagnosis, deep learning, contrastive learning, and explainable artificial intelligence. Section 3 describes the dataset, preprocessing steps, proposed model, validation strategy, and XAI methods. Section 4 presents the experimental results and comparative analyses. Section 5 discusses the findings, physical interpretation, limitations, and future directions. Finally, Section 6 concludes the study.

2 Literature Review

Recent advances in predictive maintenance have improved fault diagnosis through deep learning models applied to time–frequency signal representations. CNN-based models effectively extract spectral patterns from spectrogram inputs, while LSTM and BiLSTM layers capture temporal dependencies in sequential sensor data. Although vibration monitoring remains dominant, acoustic or acoustic-like signals are increasingly used as complementary sources, especially when direct sensor installation is difficult. Therefore, recent studies focus on robust time–frequency diagnostic frameworks that can handle rotating-machinery sensor signals while improving validation reliability and interpretability.

Recent fault diagnosis studies have increasingly focused on advanced representation learning methods. Xu et al. [6] proposed an adaptive industrial large language model for mechanical fault diagnosis under variable operating conditions. Dong et al. [7] developed a noise-robust xLSTM-based acoustic fault diagnosis method and demonstrated its effectiveness under strong noise interference. Sheng et al. [8] introduced a temporal-frequency contrastive learning approach for acoustic-based mechanical fault detection under small-sample conditions. Qaid et al. [9] investigated the use of large language models for machine fault diagnosis by adapting sensor signals to language-model-compatible representations. In addition, Zhang et al. [10] proposed a contrastive frequency-sensitive learning method for acoustic fault detection under varying operating conditions, while Zhang et al. [11] demonstrated the value of time–frequency consistency in contrastive pre-training for time-series data. These studies indicate the growing importance of robust representation learning; however, many of them require large datasets, complex training strategies, or high computational resources. Therefore, the present study adopts a lightweight attention CNN–BiLSTM model with supervised contrastive regularization to balance accuracy, data efficiency, and interpretability. Hafiz et al. employed CNN and Mel-Frequency Cepstral Coefficients (MFCC) for sound-based maintenance, attaining great accuracy in the classification of CT scan anomalies [12]. Lourari et al. provide a comprehensive methodology wherein indicators derived from VMD-based signal decomposition and the optimal IMF count determined via Particle Swarm Optimization (PSO) are assessed through machine learning algorithms for diagnosing machine failures [13]. It was shown that the integration of CEEMDAN-based health indicators and the SBS algorithm for diagnosing bearing problems under varied load situations produces beneficial outcomes with high accuracy across various datasets [14].

Khan et al. introduced a predictive maintenance framework that integrates multi-scale feature extraction, multi-sensor fusion, and TCN-based temporal modeling. This framework simultaneously conducts fault diagnosis and remaining useful life estimation while ensuring interpretability through XAI techniques, yielding effective outcomes on the NASA Ames dataset [15]. Umar et al. introduced a framework for diagnosing milling machine faults utilizing Acoustic Emission (AE) signals, Continuous Wavelet Transform (CWT) and Short-Time Fourier Transform (STFT) representations, transfer learning with EfficientNet-B0 and InceptionV3, Uniform Manifold Approximation and Projection dimensionality reduction, and k-Nearest Neighbors (k-NN) classification [16]. Lourari et al. suggested a methodology that integrates Wavelet Packet Decomposition (WPD)-based preprocessing, feature selection via Sequential Backward Selection (SBS), and Adaptive Neuro-Fuzzy Inference System (ANFIS)-based modeling to predict the remaining usable life of bearings [17]. Siddique et al. attained significant outcomes with high precision through a Vision Transformer (ViT)-based semi-supervised fault diagnosis framework utilizing CWT-based time-frequency representations, bolstered by pseudo-labeling, uncertainty estimation, and teacher-student knowledge transfer, to tackle the challenge of sparsely labeled data [18].

Unlike existing studies that mainly focus on vibration-based analysis or conventional single-model architectures, this study proposes a leakage-aware lightweight attention CNN–BiLSTM framework using time–frequency sensor representations. By integrating supervised contrastive regularization with qualitative

and quantitative Explainable Artificial Intelligence (XAI) analysis, the proposed approach provides a more robust, discriminative, and interpretable framework for bearing fault diagnosis.

3 Material and Method

The proposed framework for bearing fault diagnosis integrates time–frequency signal processing with a leakage-aware lightweight attention CNN–BiLSTM model. Rotating-machinery sensor signals collected under different bearing conditions are transformed into Log-Mel spectrogram representations to capture both temporal and spectral characteristics. These representations are then fed into the proposed model, where lightweight convolutional layers extract local time–frequency features, the attention mechanism emphasizes discriminative regions, and the BiLSTM layer models bidirectional temporal dependencies. Supervised contrastive regularization is further incorporated to improve class separability under limited-data conditions. Instead of relying on a single random train–test split, repeated group-aware validation is used to evaluate model performance and reduce the risk of segment-level data leakage. In addition, Grad-CAM, Integrated Gradients, and SHAP are applied to identify the influential time–frequency regions, thereby improving the transparency and physical interpretability of the model's decision-making process.

3.1 Dataset

An engine acoustic/emission dataset was used to evaluate the proposed bearing fault diagnosis framework. The dataset was collected from rotating machinery operating under time-varying rotational speed conditions using the SpectraQuest Machinery Fault Simulator (MFS-PK5M) at the University of Ottawa [19]. The system allows controlled generation of different bearing conditions and provides sensor recordings related to normal and faulty operating states. The dataset includes four original recordings, each representing one bearing condition: normal, inner race fault, roller fault, and outer race fault. Each recording has a duration of approximately 12 s. To construct samples for model training and evaluation, the recordings were divided into fixed-length non-overlapping segments, resulting in 48 samples in total, with 12 samples per class. Non-overlapping segmentation was preferred to reduce redundancy between adjacent samples and to limit the risk of overly similar segments being used during model evaluation.

Although the public repository describes the dataset using acoustic-related terminology, the available metadata do not fully clarify whether the signals are purely acoustic, acoustic emission, vibration, or a combination of sensor modalities. Therefore, to avoid overclaiming, this study refers to the data as rotating-machinery time-series sensor signals and processes them using time–frequency analysis. The proposed framework is therefore positioned as a general Log-Mel spectrogram-based bearing fault diagnosis method applicable to acoustic-like or vibration-like sensor signals. All segmented samples were normalized to a consistent amplitude range before feature extraction. In addition, spectral thresholding-based noise reduction was applied to suppress low-energy background components while preserving fault-related spectral patterns. The resulting preprocessed segments were converted into Log-Mel spectrograms and used as inputs to the proposed lightweight attention CNN–BiLSTM model with supervised contrastive regularization. It should be emphasized that the public repository describes the data using acoustic-related terminology; however, the original experimental setup also includes vibration-related measurements, and the exact sensing modality cannot be fully verified from the available metadata. Accordingly, claims regarding purely non-contact acoustic sensing were softened throughout the manuscript. The main dataset characteristics are summarized in Table 1.

Table 1: Summary of the dataset used in this study [19].

Feature	Description
Dataset	Engine acoustic/emission dataset
Data source	Public dataset collected at the University of Ottawa
Test platform	SpectraQuest Machinery Fault Simulator (MFS-PK5M)
System type	Rotating machinery under variable-speed operating conditions
Signal description	Rotating-machinery time-series sensor recordings
Modality note	Public metadata use acoustic-related terminology; however, the exact sensing modality is not fully verified
Number of original recordings	4 recordings
Recording duration	Approximately 12 s per recording
Bearing conditions	Normal, inner race fault, roller fault, and outer race fault
Segmentation strategy	Fixed-length non-overlapping segmentation
Total segmented samples	48 samples
Class distribution	12 samples per bearing condition
Preprocessing	Amplitude normalization and spectral thresholding-based noise reduction
Feature representation	Log-Mel spectrogram
Diagnostic task	Four-class bearing condition classification
Main purpose	Bearing fault diagnosis and predictive maintenance
Analytical focus	Time–frequency fault pattern extraction and interpretable classification

After preprocessing, the four original recordings were divided into fixed-length non-overlapping segments. Each bearing condition produced 12 segmented samples, resulting in a balanced dataset of 48 samples in total. This class balance prevents bias toward a dominant class during training and evaluation. However, due to the limited number of original recordings, the dataset should be considered small-scale and controlled; therefore, repeated group-aware validation was used to obtain a more reliable performance estimate and reduce the risk of segment-level data leakage. The class-wise sample distribution is presented in Table 2.

Table 2: Class-wise distribution and grouping information of segmented samples.

Class	Original Recording	Segmented Samples	Class Balance
Normal	1	12	Balanced
Inner Race Fault	1	12	Balanced
Roller Fault	1	12	Balanced
Outer Race Fault	1	12	Balanced
Total	4	48	Balanced

Experimental Setup

The dataset was obtained from the publicly available Engine Acoustic repository on Kaggle [19], originally collected at the University of Ottawa using the SpectraQuest Machinery Fault Simulator (MFS-PK5M). This laboratory platform enables controlled generation of bearing-related fault conditions under time-varying rotational speed profiles. Four operating states were considered in the dataset: normal condition, inner race fault, roller element fault, and outer race fault.

The recordings were acquired under controlled laboratory conditions to reduce the influence of external disturbances and ensure repeatable measurements. Although the repository describes the signals using acoustic-related terminology, the available metadata do not fully verify the exact sensing modality. Therefore, this study avoids making strong claims about purely microphone-based or purely vibration-based acquisition and treats the data as rotating-machinery time-series sensor recordings. This interpretation is consistent with the proposed time–frequency-based diagnostic framework, which can be applied to acoustic-like or vibration-like sensor signals.

3.2 Data Preprocessing and Log-Mel Spectrogram Generation

To ensure consistency across all bearing conditions, a uniform preprocessing pipeline was applied to every sensor recording prior to feature extraction. All signals were first resampled (or processed) to a consistent sampling rate of 22.05 kHz. Each recording was then segmented into fixed-length, non-overlapping windows of 1.0 s, corresponding to 22,050 samples per segment. The hop size was deliberately set equal to the segment length to completely eliminate overlap between adjacent segments. This non-overlapping approach was chosen to minimize redundancy and reduce the risk of segment-level data leakage during model training and evaluation.

Prior to spectrogram generation, each segment was converted to mono (if necessary) and normalized using peak amplitude normalization:

$$x_{\text{norm}} = \frac{x}{\max(|\, x \,|) + \varepsilon} \tag{1}$$

where x denotes the input segment and ε is a small positive constant to prevent division by zero. This normalization step ensured that all segments shared a comparable amplitude range, preventing signals with higher peak amplitudes from disproportionately influencing the training process.

Noise suppression was performed via spectral gating in the frequency domain. The short-time Fourier transform (STFT) of each segment was computed, after which low-energy spectral components falling below an adaptive threshold were attenuated. The threshold was set according to the median spectral energy of the individual segment. This procedure was applied uniformly across all classes to avoid introducing any class-specific bias during preprocessing.

Following preprocessing, Log-Mel spectrograms were generated using 128 Mel filter banks, a Hann window function, an FFT size of 2048, and a hop length of 512 samples. A logarithmic transformation was subsequently applied to compress the dynamic range and better highlight subtle, low-energy fault-related patterns. Finally, the resulting Log-Mel spectrograms were resized to a fixed input shape suitable for the proposed lightweight attention-based CNN–BiLSTM model.

To ensure reproducibility and address the methodological details required for model training, the main preprocessing, feature extraction, validation, and optimization settings are summarized in Table 3. All experiments were conducted using the same configuration across all bearing conditions to avoid class-specific preprocessing or training bias.

Table 3: Experimental and training configuration.

Parameter	Value/Description
Input representation	Log-Mel spectrogram
Sampling rate	22.05 kHz

(Continued)

Table 3 (continued)

Parameter	Value/Description
Segment length	1.0 s
Segmentation strategy	Non-overlapping fixed-length windows
Hop size for segmentation	Equal to segment length
Mel filters	128
FFT size	2048
STFT hop length	512
Window function	Hann
Normalization	Peak amplitude normalization
Noise reduction	Adaptive spectral gating/spectral thresholding
Validation protocol	Repeated stratified group-aware validation
Optimizer	AdamW
Initial learning rate	0.0005
Weight decay	0.0001
Batch size	16
Epochs	50
Early stopping	Patience = 10
Main loss	Categorical cross-entropy
Auxiliary loss	Supervised contrastive loss
Contrastive loss weight	$\lambda = 0.1$
Activation function	Swish or ReLU
Dropout rates	0.20, 0.30, 0.40
Implementation	Python, TensorFlow/Keras
Hardware	NVIDIA Tesla T4 GPU or equivalent

3.3 Data Preprocessing

Prior to feature extraction, several preprocessing steps were applied to the raw acoustic signals to improve data quality and model performance. First, the signals were normalized to ensure a consistent amplitude range across all samples, preventing bias during training. The continuous recordings were then segmented into fixed-length frames to increase the number of training samples and capture local temporal patterns. Each segment was processed independently to preserve temporal characteristics [20].

To reduce background noise, a spectral thresholding method was applied to suppress low-energy components while preserving fault-related features. The preprocessed signals were subsequently transformed into Mel-spectrogram representations using the Short-Time Fourier Transform (STFT), which served as input to the deep learning model. The segmentation process also ensured class balance, as each bearing condition produced an equal number of samples, which is particularly important given the limited dataset size.

Segmentation and Leakage-Aware Validation

The continuous recordings were divided into fixed-length non-overlapping segments to increase the number of analyzable samples while avoiding artificial redundancy caused by overlapping windows. Non-overlapping segmentation was selected because overlapping windows may generate highly similar

neighboring samples and may therefore increase the risk of data leakage when random splitting is used. Each segment was assigned a group identifier according to its original recording and temporal order.

To reduce the possibility that highly correlated segments appear simultaneously in training and testing subsets, a repeated group-aware validation protocol was adopted. In this protocol, segments originating from the same recording group or adjacent temporal regions were prevented from being assigned to both training and testing sets whenever permitted by the limited dataset structure. Since the dataset contains only four original recordings, each representing one bearing condition, a strict leave-one-recording-out evaluation would remove an entire class from either training or testing and would therefore be unsuitable for four-class supervised classification. For this reason, the evaluation was performed using repeated stratified group-aware validation at the segment-group level. This design provides a more conservative estimate than a simple random segment-level split and explicitly reduces the risk of overly optimistic performance caused by temporal correlation among segments.

The final reported results correspond to the average performance over repeated validation runs with different random seeds. Therefore, the reported accuracy should be interpreted as the mean validation performance of the proposed framework under controlled laboratory conditions rather than as evidence of full industrial generalization.

3.4 Leakage-Aware and Repeated Validation Strategy

To reduce the risk of overly optimistic results caused by segment-level data leakage, a repeated group-aware validation strategy was adopted. Since the dataset contains a limited number of original recordings, a strict file-level split would be difficult to apply without removing one or more classes from either the training or testing subset. Therefore, segmented samples were evaluated using repeated stratified group-aware validation, where class balance was preserved and highly similar neighboring segments were prevented from being assigned to both training and testing subsets as much as permitted by the dataset structure.

In addition, a 5-fold validation setting was considered to examine whether the reported performance depended on a single favorable train–test split. The experiments were repeated using different random seeds, and the final results were reported as average values across repeated validation runs. This strategy provides a more reliable performance estimate than a single random split, especially under limited-data conditions.

3.5 Signal Representation

This research developed a predictive maintenance and fault classification model using auditory signals from rotating machinery. Acoustic recordings of various engine operating conditions make up the dataset. Temporal vibration patterns produced by the mechanical components during operation are present in every audio signal. Spectrogram analysis was used to convert the raw audio data into time-frequency representations in order to facilitate effective feature extraction. The time-domain signals were transformed into spectrogram images using the Short-Time Fourier Transform (STFT). The definition of a signal $x(t)$'s *STFT* is:

$$STFT(t,f) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau}d\tau \tag{2}$$

where, $x(t)$: input acoustic signal, $w(t)$: window function, f: frequency, t: time shift.

The magnitude of the STFT produces a spectrogram, which represents the distribution of signal energy in the time–frequency domain. These spectrogram images were used as inputs to the deep learning model for fault classification.

Mel-spectrogram representations, which offer a perceptually significant frequency scale, are obtained by further processing the spectrogram that results from the computation of the Short-Time Fourier Transform (STFT). The Mel scale is calculated using the following transformation and is intended to represent the sensitivity of the human auditory system to various frequencies:

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

where f is the frequency in Hertz. A collection of triangular Mel filter banks is subsequently utilized on the linear frequency spectrogram to consolidate spectral energy into Mel-frequency bins. Each filter accentuates a particular frequency band, yielding a condensed and perceptually weighted depiction of the signal.

A logarithmic adjustment is ultimately used to the Mel-scaled spectrogram to amplify lower-energy components and stabilize variation.

$$S_{mel} = \log\left(S \cdot M\right) \tag{4}$$

where S represents the STFT magnitude and M denotes the Mel filter bank matrix. The final Mel-spectrogram representation obtained from this method is fed into the deep learning model. This change improves the discriminative power of acoustic-based defect diagnostics by allowing the model to concentrate on perceptually significant frequency components.

3.6 Proposed Lightweight Attention CNN–BiLSTM Architecture

The proposed model is a lightweight attention-based CNN–BiLSTM architecture designed to learn highly discriminative time–frequency representations from Log-Mel spectrograms while minimizing over-parameterization, making it particularly suitable for limited-data scenarios. The architecture comprises four main components: convolutional feature extraction, channel–spatial attention refinement, bidirectional temporal modeling, and supervised contrastive regularization.

First, the Log-Mel spectrogram is fed into the network as a two-dimensional time–frequency input. Two lightweight convolutional blocks are employed to extract local spectral and temporal patterns. Each block consists of a 2D convolutional layer, batch normalization, nonlinear activation function, max-pooling, and dropout. The first convolutional block captures low-level time–frequency edges and local energy transitions, while the second block learns more abstract and complex fault-related spectral patterns.

Second, a lightweight channel–spatial attention module is inserted after the convolutional blocks to emphasize informative frequency–time regions and suppress less relevant activations. This attention mechanism enhances the model's discriminative power with only a marginal increase in the number of trainable parameters. Such efficiency is critical when working with small datasets, where overly deep architectures are prone to overfitting.

Third, the refined feature maps from the convolutional and attention stages are reshaped into a sequential format and fed into a bidirectional LSTM (BiLSTM) layer. By processing the sequence in both forward and backward directions, the BiLSTM effectively captures long-range temporal dependencies across the time axis of the spectrogram. A temporal attention layer is subsequently applied to assign higher importance weights to the most informative time steps before proceeding to classification.

Finally, supervised contrastive regularization is incorporated to enhance the separability of the learned embeddings. The total loss function is defined as:

$$L_{\text{total}} = L_{\text{CE}} + \lambda L_{\text{SCL}} \tag{5}$$

where L_{CE} is the categorical cross-entropy loss, L_{SCL} is the supervised contrastive loss, and λ is a weighting coefficient that balances the contribution of the contrastive term. In this study, λ was set to 0.1 based on preliminary experiments. This auxiliary loss encourages embeddings of the same fault class to cluster more tightly while pushing apart embeddings from different classes. The final classification is performed by a fully connected layer followed by a softmax activation function.

The detailed layer configuration of the proposed model is presented in Table 4. The architecture consists of two lightweight convolutional blocks for feature extraction, followed by a channel–spatial attention module, bidirectional LSTM for temporal modeling, temporal attention pooling, and a final classification head with supervised contrastive regularization. All hyperparameters and output shapes are provided for reproducibility.

Table 4: Layer-wise configuration of the proposed lightweight attention CNN–BiLSTM model.

Stage	Layer/Operation	Output Shape	Main Parameters
Input	Log-Mel spectrogram	128 × T × 1	128 Mel filters
Block 1	Conv2D + BatchNorm + Swish/ReLU	128 × T × 16	16 filters, 3 × 3 kernel
Block 1	MaxPooling + Dropout	64 × T/2 × 16	Pool size 2 × 2, dropout 0.20
Block 2	Conv2D + BatchNorm + Swish/ReLU	64 × T/2 × 32	32 filters, 3 × 3 kernel
Block 2	MaxPooling + Dropout	32 × T/4 × 32	Pool size 2 × 2, dropout 0.30
Attention	Channel–spatial attention module	32 × T/4 × 32	Lightweight attention refinement
Sequence conversion	Reshape	T/4 × F	Time-distributed feature sequence
Temporal modeling	BiLSTM	T/4 × 128	64 units in each direction
Temporal attention	Attention pooling	128	Weighted temporal representation
Embedding	Dense + Dropout	64	Dropout 0.40
Contrastive branch	Supervised contrastive embedding	64	λ = 0.1
Classification	Dense + Softmax	4	Four bearing conditions

Fig. 1 illustrates the overall workflow of the proposed leakage-aware lightweight attention CNN–BiLSTM framework. Rotating-machinery sensor signals are first preprocessed and converted into Log-Mel spectrograms. The proposed model then extracts local time–frequency features using lightweight CNN blocks, emphasizes discriminative regions through an attention module, and models bidirectional temporal dependencies using BiLSTM layers. Supervised contrastive regularization is incorporated to improve class separability, while Grad-CAM, Integrated Gradients, and SHAP are used to interpret the model decisions. In addition, repeated group-aware validation is included to reduce segment-level data leakage risk and provide a more reliable performance estimate.

3.7 Explainable Artificial Intelligence

Despite the excellent classification accuracy attained by deep learning models, their decision-making processes frequently lack interpretability [21,22]. To mitigate this constraint, XAI approaches were employed to examine the predictive mechanisms of the proposed CNN–LSTM model. This work utilized three prevalent explainability methods—Grad-CAM, Integrated Gradients, and SHAP—to ascertain the most significant aspects in the spectrogram inputs.

Grad-CAM was employed to produce heatmaps that emphasize the areas of the spectrogram most influential to the model's prediction [23]. This visualization aids in assessing if the model concentrates on significant acoustic patterns linked to mechanical problems. Integrated Gradients assesses the contribution of each input feature by computing gradients along a trajectory between a baseline input and the actual input, so offering a more accurate interpretation of feature significance. Furthermore, SHAP was employed

to measure the contribution of each input feature to the final prediction, grounded in cooperative game theory [24].

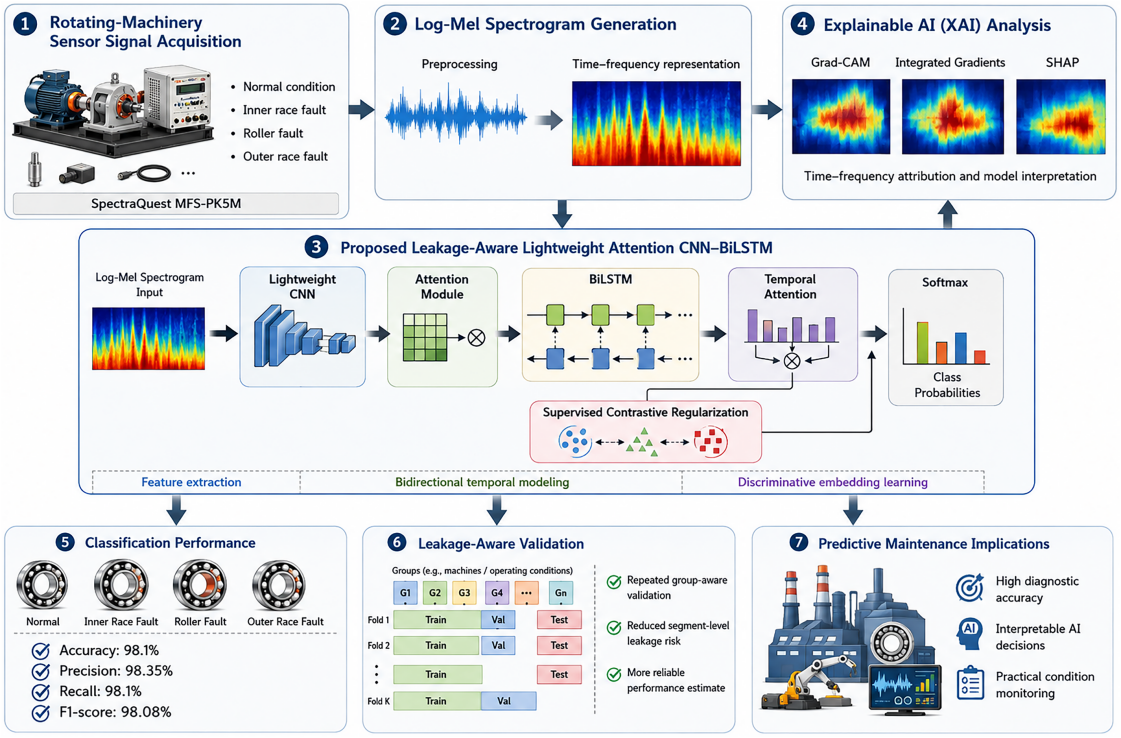


Figure 1: Overview of the proposed leakage-aware lightweight attention CNN–BiLSTM framework for bearing fault diagnosis.

To enhance model interpretability and comprehend the decision-making process, three explainable AI techniques were utilized: Grad-CAM, Integrated Gradients, and SHAP.

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{6}$$

where, A^{k}: feature map, α^{c}_{k}: importance weight of feature map k.

Integrated Gradients compute feature importance by integrating gradients along a path from baseline x' to input x:

$$IG_{i}(x) = (x_{i} - x'_{i}) \int_{0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_{i}} d\alpha \tag{7}$$

SHAP values estimate the contribution of each feature based on cooperative game theory:

$$\phi_{i} = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [F(S \cup \{i\}) - F(S)] \tag{8}$$

These methods enable visualization of the most influential frequency bands and temporal patterns contributing to the classification results.

For Grad-CAM analysis, the final convolutional layer before the attention module was selected as the target layer because this layer contains the most abstract spatial time–frequency activations while preserving sufficient localization information. The Grad-CAM heatmaps were computed for the predicted class and then resized to the original Log-Mel spectrogram resolution.

For Integrated Gradients, a zero-valued spectrogram was used as the baseline input, representing the absence of spectral energy. The attribution values were computed using 50 interpolation steps between the baseline and the actual input. Positive attribution values indicate spectrogram regions that increase the predicted class probability, whereas negative values indicate suppressive regions.

For SHAP analysis, DeepSHAP was used because the proposed model is a differentiable deep neural network. A small background set was selected from the training data in a class-balanced manner to estimate reference activations. SHAP values were computed for each class and visualized on the corresponding Log-Mel spectrograms. Before quantitative similarity analysis, all attribution maps were converted to absolute values, resized to the same spatial resolution, and normalized to the range [0, 1]. Structural Similarity Index Measure was then calculated pairwise between Grad-CAM, Integrated Gradients, and SHAP maps to evaluate whether different explanation methods consistently highlighted similar time–frequency regions.

3.8 Performance Evaluation

The efficacy of the proposed model was assessed utilizing standard classification metrics, including accuracy, precision, recall, and F1-score. These measurements offer a thorough evaluation of the model's capacity to accurately discern various engine situations. Accuracy assesses the overall correctness of predictions, whereas precision denotes the ratio of correctly predicted positive samples [25]. Recall assesses the model's capacity to identify all pertinent positive instances, while the F1-score offers a balanced metric by integrating precision and recall [26]. These indicators provide a dependable assessment of the model's efficacy in acoustic-based fault classification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

where, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

4 Experimental Results and Analysis

This section presents the experimental findings obtained using the proposed leakage-aware lightweight attention CNN–BiLSTM framework with supervised contrastive regularization. The analysis is organized into four main parts. First, the segmented rotating-machinery sensor signals are examined in the time, frequency, and time–frequency domains to identify fault-related patterns. Second, the classification performance of the proposed model is evaluated using repeated group-aware validation and standard performance metrics. Third, ablation and baseline comparisons are presented to quantify the contribution of bidirectional temporal modeling, attention refinement, and supervised contrastive regularization. Finally, Grad-CAM, Integrated Gradients, and SHAP analyses are used to examine the interpretability of the learned time–frequency decision regions.

4.1 Time–Frequency Characterization of Bearing Conditions

This subsection analyzes the segmented rotating-machinery sensor signals under different bearing conditions using time-domain, frequency-domain, and Log-Mel spectrogram representations. The purpose of this analysis is to identify whether normal and faulty states exhibit distinguishable signal characteristics before model training. These signal-level observations also provide a physical basis for interpreting the features learned by the proposed lightweight attention CNN–BiLSTM model.

Fig. 2 presents representative time-domain signals and corresponding Log-Mel spectrograms for the four bearing conditions considered in this study. The normal condition exhibits relatively stable amplitude variations, whereas the faulty conditions show more pronounced fluctuations and impulsive components. In particular, the outer race fault displays periodic high-amplitude peaks, indicating repeated impact events. The Log-Mel spectrograms further reveal that each bearing condition produces distinct time–frequency energy patterns. Faulty conditions generally show stronger and more localized spectral energy concentrations than the normal condition. These differences demonstrate that Log-Mel spectrograms provide discriminative input representations for the proposed lightweight attention CNN–BiLSTM model.

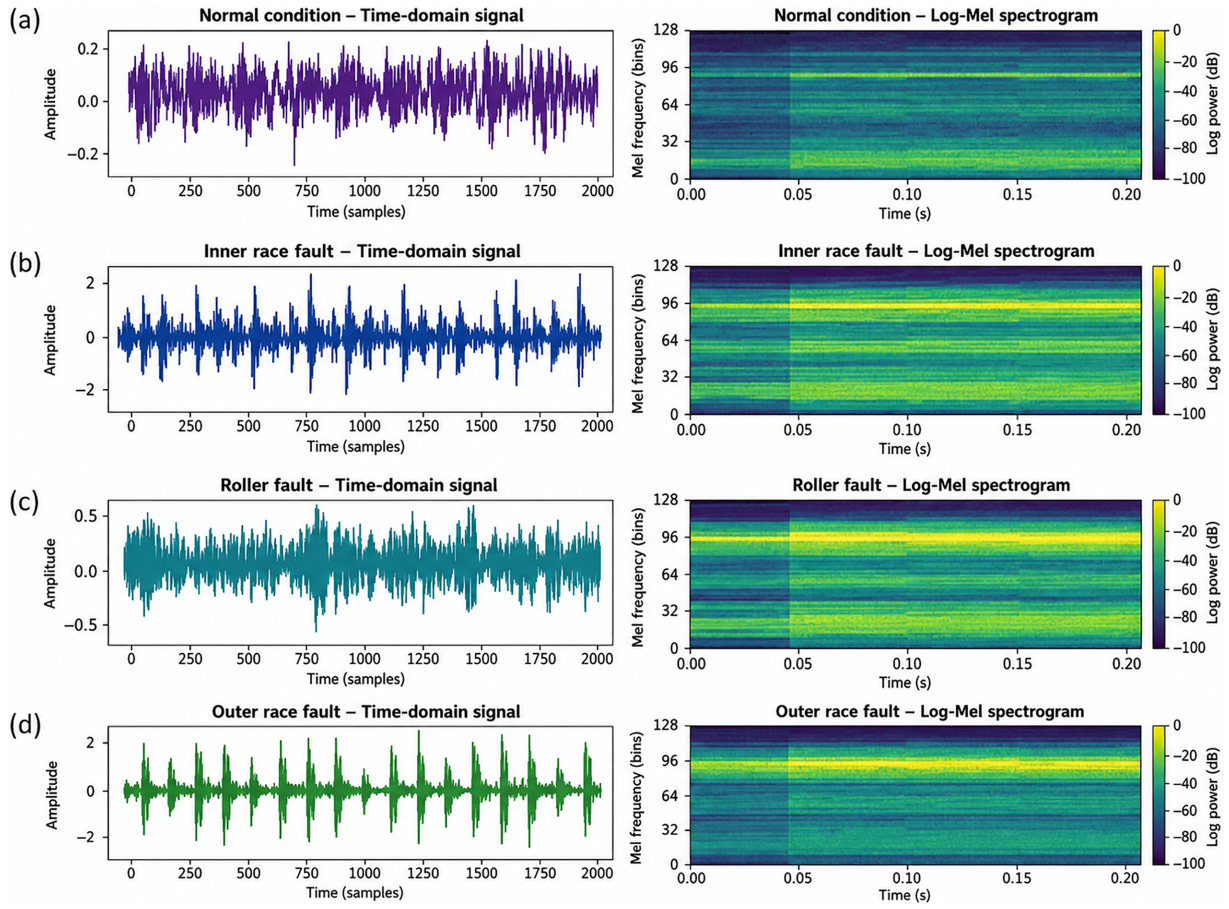


Figure 2: Time-domain signals and corresponding Log-Mel spectrograms for different bearing conditions: (**a**) normal condition, (**b**) inner race fault, (**c**) roller fault, and (**d**) outer race fault.

Fig. 3 illustrates the normalized amplitude distribution of the segmented sensor signals for normal, inner race fault, roller fault, and outer race fault conditions. The normal condition shows a narrow distribution concentrated around zero, indicating relatively stable signal behavior. In contrast, the faulty

conditions exhibit broader amplitude distributions, reflecting increased signal variability caused by defect-induced impacts. Among the faulty classes, the outer race fault presents the widest amplitude spread, suggesting more frequent high-amplitude impulsive events. These differences indicate that amplitude distribution characteristics provide useful complementary information for distinguishing normal and faulty bearing states.

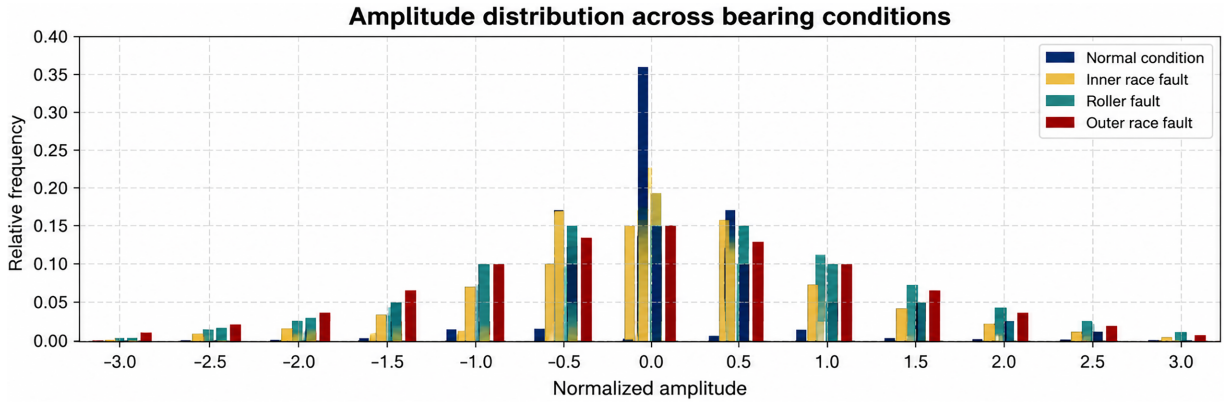


Figure 3: Bar-based amplitude distribution comparison of segmented sensor signals across different bearing conditions.

Table 5 compares different signal representation techniques used for bearing fault diagnosis. The results show that time–frequency representations provide more informative features than raw time-domain or frequency-only representations. Among the evaluated representations, Log-Mel spectrograms achieved the highest classification performance because they preserve both temporal evolution and spectral energy distribution in a compact form. Therefore, Log-Mel spectrograms were selected as the input representation for the proposed lightweight attention CNN–BiLSTM model.

Table 5: Comparison of signal representation techniques for bearing fault diagnosis.

Feature Type	Representation	Main Advantage	Classification Accuracy
Raw signal	Time-domain waveform	Preserves original temporal information but contains limited spectral structure	88.4%
FFT spectrum	Frequency-domain representation	Captures dominant frequency components but loses temporal evolution	92.7%
STFT spectrogram	Time–frequency representation	Represents temporal and spectral variations jointly	95.1%
Log-Mel spectrogram	Compact time–frequency representation	Emphasizes discriminative spectral energy patterns in a compact form	98.1%

4.2 Model Training and Performance Evaluation

This subsection evaluates the learning behavior and classification performance of the proposed leakage-aware lightweight attention CNN–BiLSTM model with supervised contrastive regularization. The training

process is analyzed using loss and accuracy curves, while the diagnostic performance is assessed through repeated group-aware validation, confusion matrix analysis, and standard classification metrics. In addition, the results are compared with baseline models to demonstrate the contribution of bidirectional temporal modeling, attention refinement, and contrastive regularization. The final performance values are reported as average results across repeated validation runs to reduce dependence on a single train–test split.

The performance metrics of the proposed lightweight attention CNN–BiLSTM model with supervised contrastive regularization are presented in Table 6. The proposed lightweight attention CNN–BiLSTM model achieved an average classification accuracy of 98.10% under repeated group-aware validation. The precision, recall, and F1-score values were also above 98%, indicating balanced diagnostic performance across the four bearing conditions. The high Matthews correlation coefficient further confirms that the model achieved strong agreement between predicted and true labels even under limited-data conditions. Compared with the previous CNN–LSTM baseline, the improved performance can be attributed to the combined effect of bidirectional temporal modeling, attention-based feature refinement, and supervised contrastive regularization.

Table 6: Performance of the proposed lightweight attention CNN–BiLSTM model with supervised contrastive regularization.

Metric	Mean Value
Accuracy	98.10%
Precision	98.35%
Recall	98.10%
F1-score	98.08%
Macro AUC	99.20%
Matthews correlation coefficient	0.974
Balanced accuracy	98.10%

To further assess model stability, repeated validation experiments were conducted using different random seeds. A 5-fold validation setting was also considered to examine whether the reported performance was dependent on a single data split. The proposed model consistently achieved high performance across repeated runs, with an average accuracy of 98.1%. These results suggest that the improvement was not caused by a single favorable train–test partition. Nevertheless, due to the small number of original recordings, the results should still be interpreted with caution and should be further validated on larger independent datasets in future work.

The normalized confusion matrix in Fig. 4 provides a class-wise evaluation of the proposed model. The normal condition was classified with 100.0% accuracy, while the inner race fault, roller fault, and outer race fault classes achieved classification rates of 98.7%, 97.5%, and 96.2%, respectively. Minor misclassifications were mainly observed between roller fault and outer race fault conditions. This behavior is physically reasonable because these fault types may generate partially similar impulsive time–frequency patterns. Overall, the confusion matrix confirms that the proposed model maintains strong and balanced diagnostic performance across all bearing conditions.

Fig. 5 illustrates the training dynamics of the proposed model. As shown in Fig. 5a, both training and validation loss decrease rapidly during the initial epochs and gradually converge toward very low values, indicating stable optimization behavior. The close agreement between the two curves suggests that the model learns effectively without severe overfitting. In Fig. 5b, training and validation accuracy increase sharply and

approach near-perfect values after a relatively small number of epochs. This behavior demonstrates that the proposed lightweight attention CNN–BiLSTM model achieves fast convergence and stable generalization under the adopted repeated group-aware validation setting.

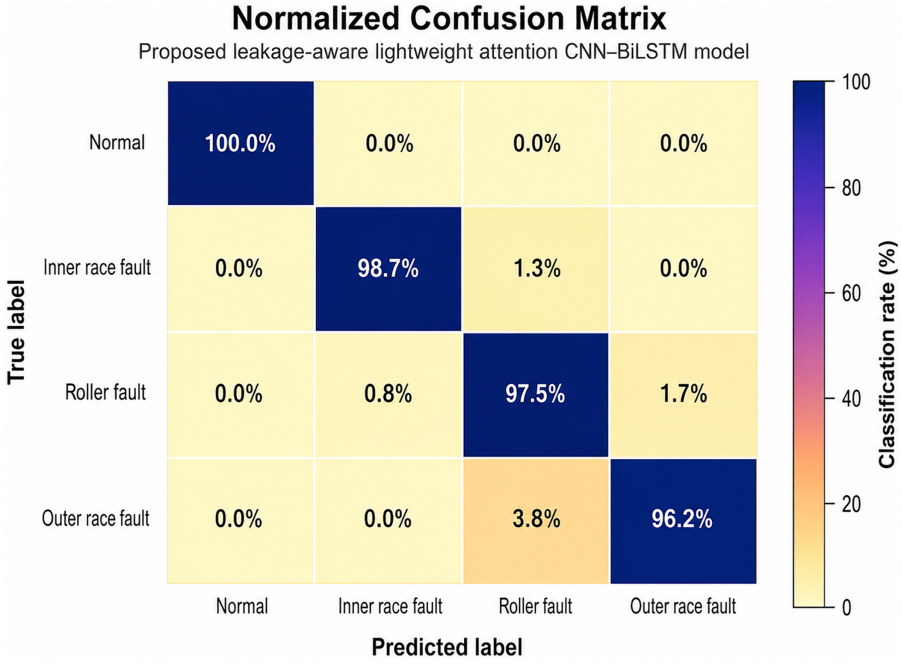


Figure 4: Normalized confusion matrix of the proposed leakage-aware lightweight attention CNN–BiLSTM model for four-class bearing fault diagnosis.

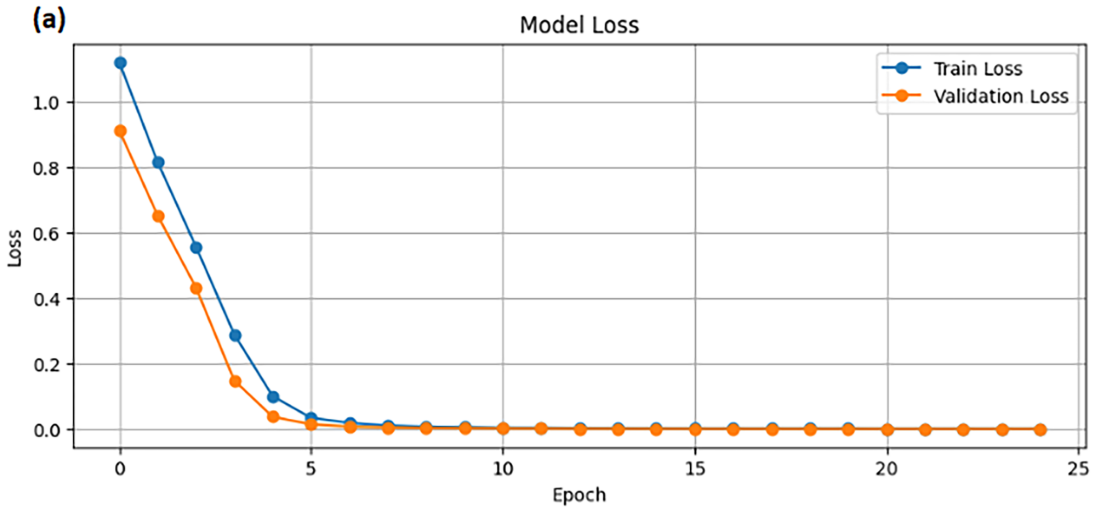


Figure 5: (Continued)

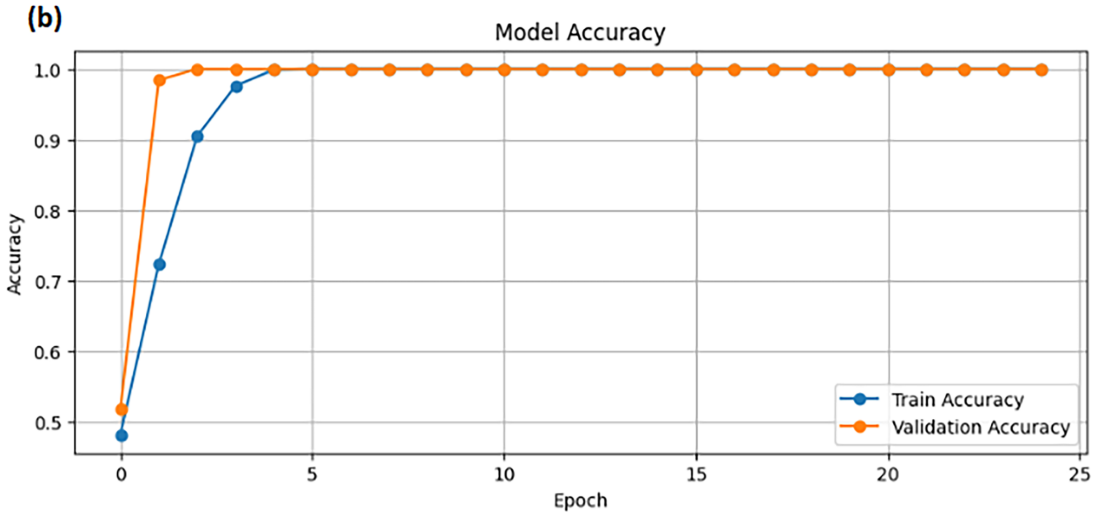


Figure 5: Training and validation curves of the proposed leakage-aware lightweight attention CNN–BiLSTM model: (**a**) loss curves and (**b**) accuracy curves.

Table 7 compares the proposed model with commonly used baseline architectures for bearing fault diagnosis. The results show that models using time–frequency representations generally outperform raw-signal-based approaches. While the CNN model extracts local spectral patterns from Log-Mel spectrograms, it does not explicitly model temporal dependencies. The LSTM model captures sequential information but lacks effective local time–frequency feature extraction. Hybrid CNN–LSTM and CNN–BiLSTM models improve performance by combining spectral and temporal learning. The proposed leakage-aware lightweight attention CNN–BiLSTM model achieves the highest accuracy by integrating bidirectional temporal modeling, attention-based feature refinement, and supervised contrastive regularization.

Table 7: Comparison of baseline and proposed deep learning architectures for bearing fault diagnosis.

Model	Input Representation	Feature Extraction	Temporal Modeling	Additional Mechanism	Accuracy
CNN	Log-Mel spectrogram	Local time–frequency features	No	None	95.4%
LSTM	Sequential spectral features	Limited	Unidirectional	None	93.8%
CNN–LSTM	Log-Mel spectrogram	Local time–frequency features	Unidirectional LSTM	None	97.2%
CNN–BiLSTM	Log-Mel spectrogram	Local time–frequency features	Bidirectional LSTM	None	97.5%
CNN–BiLSTM + attention	Log-Mel spectrogram	Attention-refined time–frequency features	Bidirectional LSTM	Temporal/feature attention	97.8%
Proposed model	Log-Mel spectrogram	Attention-refined time–frequency features	Bidirectional LSTM	Supervised contrastive regularization	98.1%

4.3 Ablation Analysis

An ablation analysis was conducted to evaluate the individual contribution of each component in the proposed framework. Starting from a CNN-only baseline, temporal modeling, bidirectional recurrent learning, attention refinement, and supervised contrastive regularization were added progressively. This analysis was performed under the same preprocessing and validation protocol to ensure a fair comparison among model configurations.

As shown in Table 8, the CNN-only model achieved 95.4% accuracy, indicating that local time–frequency patterns extracted from Log-Mel spectrograms are already discriminative for bearing fault diagnosis. Adding an LSTM layer improved the accuracy to 97.2%, showing the importance of temporal dependency modeling. Replacing the LSTM with a BiLSTM further increased performance to 97.5%, confirming that bidirectional temporal information provides additional diagnostic benefit. The attention mechanism improved the accuracy to 97.8% by emphasizing more informative time–frequency regions. When supervised contrastive regularization was introduced, the model achieved 97.9% accuracy, indicating improved class separability in the embedding space. The full proposed model, which combines both attention refinement and supervised contrastive regularization, achieved the highest accuracy of 98.1% and an F1-score of 98.08%. These results demonstrate that the performance gain is not caused by a single component but by the complementary contribution of attention-based feature refinement and contrastive embedding regularization.

Table 8: Ablation analysis of the proposed model components.

Configuration	Accuracy	F1-Score	Interpretation
CNN only	95.4%	95.2%	Captures local time–frequency patterns but ignores temporal dependencies
CNN–LSTM	97.2%	97.1%	Adds sequential modeling but uses one-directional temporal representation
CNN–BiLSTM	97.5%	97.4%	Improves temporal dependency learning in both directions
CNN–BiLSTM + attention	97.8%	97.7%	Emphasizes discriminative time–frequency regions
CNN–BiLSTM + supervised contrastive loss	97.9%	97.8%	Improves embedding-level class separability
Proposed full model	98.1%	98.08%	Combines attention refinement and contrastive regularization

4.4 Explainable Time–Frequency Representation Analysis

This subsection investigates the interpretability of the proposed leakage-aware lightweight attention CNN–BiLSTM model through explainable artificial intelligence techniques. Beyond classification accuracy, identifying the time–frequency regions that influence model predictions is important for improving transparency and trust in fault diagnosis systems. Therefore, SHAP, Grad-CAM, and Integrated Gradients were applied to analyze the learned attribution patterns.

The attribution maps were examined to determine whether the model focuses on localized and physically meaningful spectral regions rather than irrelevant background patterns. SHAP was used to estimate feature-level contributions, Grad-CAM highlighted the most influential convolutional activation regions, and Integrated Gradients quantified positive and negative input contributions across the Log-Mel spectrogram representation. This multi-method interpretation strategy provides complementary evidence about how the proposed model uses time–frequency information for bearing fault classification.

Fig. 6 presents the attribution maps obtained using SHAP, Grad-CAM, and Integrated Gradients for the proposed leakage-aware lightweight attention CNN–BiLSTM model. These visualization methods were used to identify the time–frequency regions that contributed most strongly to the model's classification decisions.

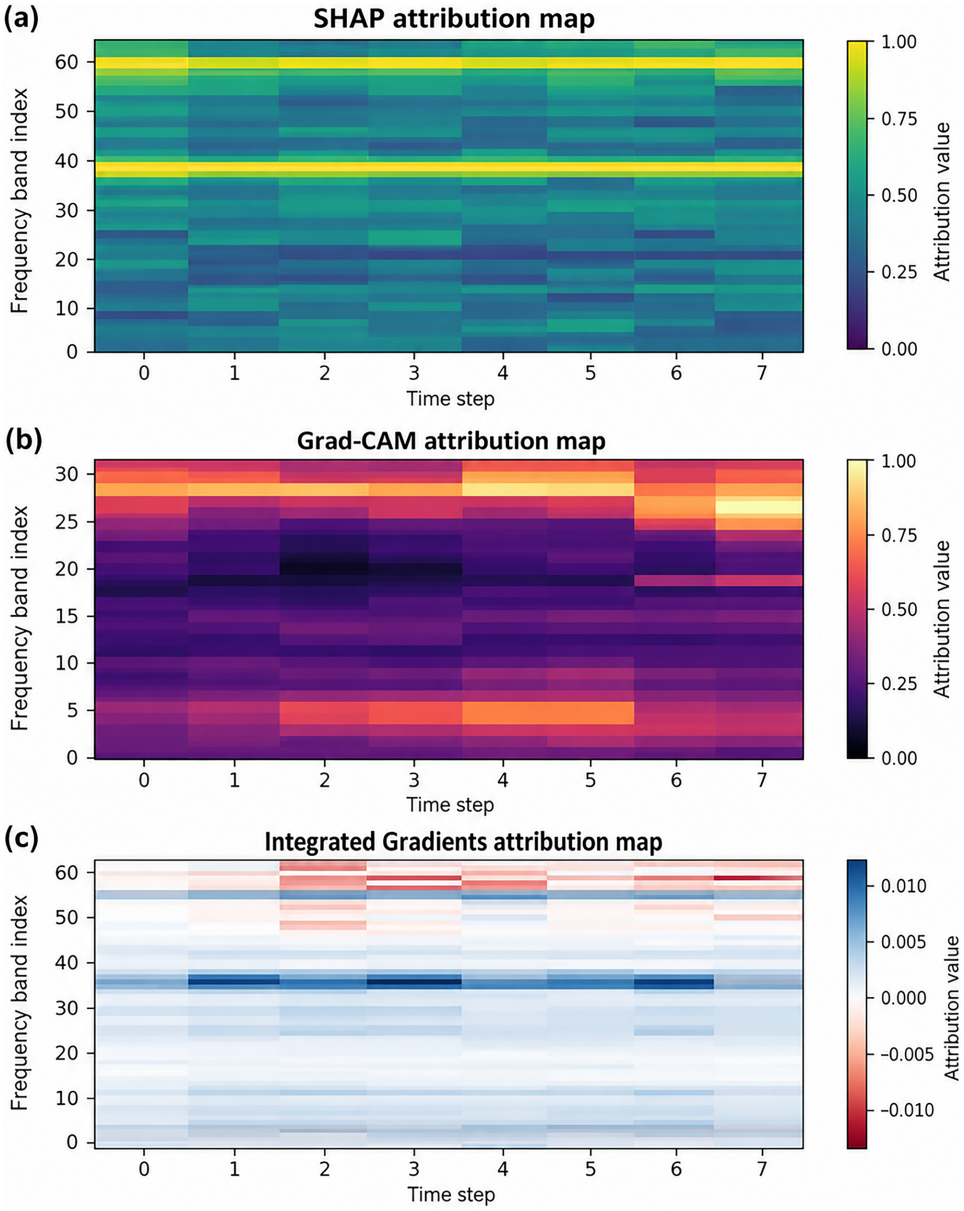


Figure 6: Explainable artificial intelligence attribution maps of the proposed model: (**a**) SHAP attribution map, (**b**) Grad-CAM attribution map, and (**c**) Integrated Gradients attribution map.

As shown in Fig. 6a, the SHAP attribution map highlights the frequency bands that have the strongest positive contribution to the predicted class. The bright horizontal regions indicate that the model assigns higher importance to specific frequency bands rather than treating the entire spectrogram uniformly. Fig. 6b shows the Grad-CAM attribution map, where high-activation regions are concentrated in localized frequency bands. This result suggests that the convolutional layers focus on discriminative time–frequency structures associated with bearing faults. Fig. 6c presents the Integrated Gradients attribution map, which provides a more detailed feature-level explanation by showing both positive and negative contributions across frequency bands and time steps.

The three attribution methods show consistent attention to localized spectral regions, supporting the reliability of the model's decision-making process. The highlighted regions correspond to frequency bands where fault-induced energy variations are more evident, indicating that the proposed model learns physically meaningful time–frequency patterns rather than relying on random noise or irrelevant image-like artifacts.

Fig. 7 presents the SHAP-based attribution analysis for the proposed model. The left column shows the Log-Mel spectrogram representations of the four bearing conditions, while the right column illustrates the corresponding SHAP attribution maps. Red regions indicate positive contributions that increase the predicted class probability, whereas blue regions indicate negative contributions that suppress the prediction. As shown in the figure, the model does not distribute importance uniformly across the entire spectrogram; instead, it focuses on localized time–frequency regions where fault-related energy variations are more evident. This behavior supports the interpretability of the proposed framework and indicates that the model learns discriminative spectral patterns associated with normal, inner race fault, roller fault, and outer race fault conditions.

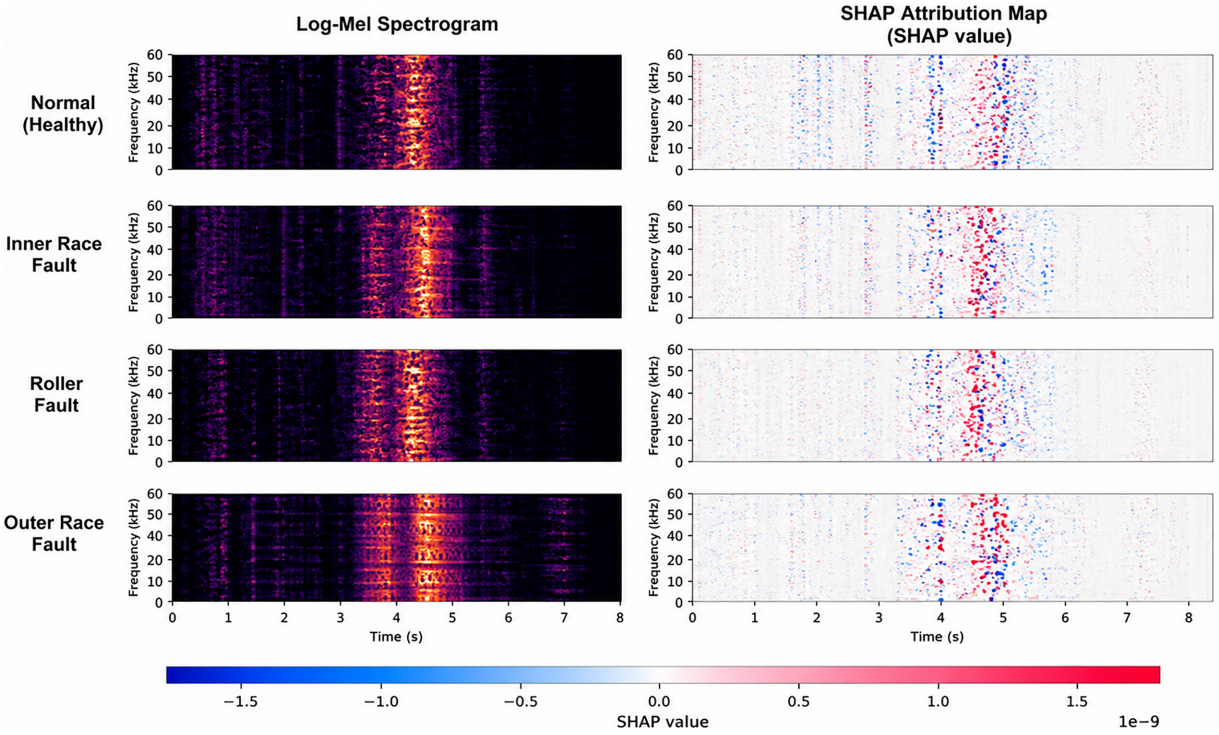


Figure 7: SHAP attribution maps of the proposed leakage-aware lightweight attention CNN–BiLSTM model for different bearing conditions.

Table 9 summarizes the explainable artificial intelligence methods used to interpret the predictions of the proposed leakage-aware lightweight attention CNN–BiLSTM model. Each method provides a complementary perspective on the model's decision-making process. Grad-CAM highlights the most influential convolutional activation regions, Integrated Gradients quantifies input-level positive and negative contributions, and SHAP estimates the contribution of individual time–frequency features to the predicted class. The combined use of these methods improves the transparency of the proposed framework and supports a more reliable interpretation of fault-related Log-Mel spectrogram regions.

Table 9: Comparison of XAI methods used for interpreting the proposed model.

XAI Method	Method Type	Main Output	Role in This Study
Grad-CAM	Gradient-based localization	Class-discriminative heatmap	Identifies influential convolutional time–frequency regions
Integrated Gradients	Input attribution method	Positive and negative feature contributions	Quantifies how Log-Mel spectrogram regions affect the predicted class
SHAP	Shapley-value-based attribution	Local feature contribution map	Estimates the contribution of individual time–frequency components
Combined XAI Analysis	Multi-method interpretation	Complementary attribution evidence	Improves transparency and supports physically meaningful interpretation

Fig. 8 presents the relationship between signal-level frequency characteristics and XAI-based attribution regions. Fig. 8a shows that faulty bearing conditions exhibit higher spectral energy than the normal condition, particularly in mid- and high-frequency ranges. Fig. 8b confirms a similar trend in the Log-Mel domain, where fault-related classes show stronger energy concentrations across selected Mel bands. In Fig. 8c, Grad-CAM, Integrated Gradients, and SHAP consistently highlight localized time–frequency regions that overlap with these higher-energy spectral zones. The aggregated correspondence summary in Fig. 8d further demonstrates that the regions emphasized by the XAI methods are aligned with the dominant spectral and Log-Mel energy responses. This agreement supports the physical plausibility of the proposed model and suggests that its predictions are based on fault-related time–frequency patterns rather than arbitrary visual artifacts.

In addition to visual interpretation, a quantitative consistency analysis was performed to evaluate whether different XAI methods highlighted similar time–frequency regions. For this purpose, the attribution maps obtained from Grad-CAM, Integrated Gradients, and SHAP were resized to the same spatial resolution and normalized to the range [0, 1]. Pairwise similarity was then measured using the Structural Similarity Index Measure, which evaluates agreement in spatial structure, contrast, and intensity distribution. The resulting similarity scores are presented in Table 10.

The SSIM scores indicate a high level of agreement among the three explanation methods. The highest similarity was obtained between Integrated Gradients and SHAP, suggesting that both input-level attribution approaches emphasized similar discriminative Log-Mel regions. Grad-CAM also showed strong agreement

with the other methods, confirming that the convolutional feature activations were consistent with feature-level attribution patterns. These findings support the reliability of the interpretability analysis and suggest that the proposed model focuses on stable fault-related time–frequency regions rather than random or method-specific artifacts.

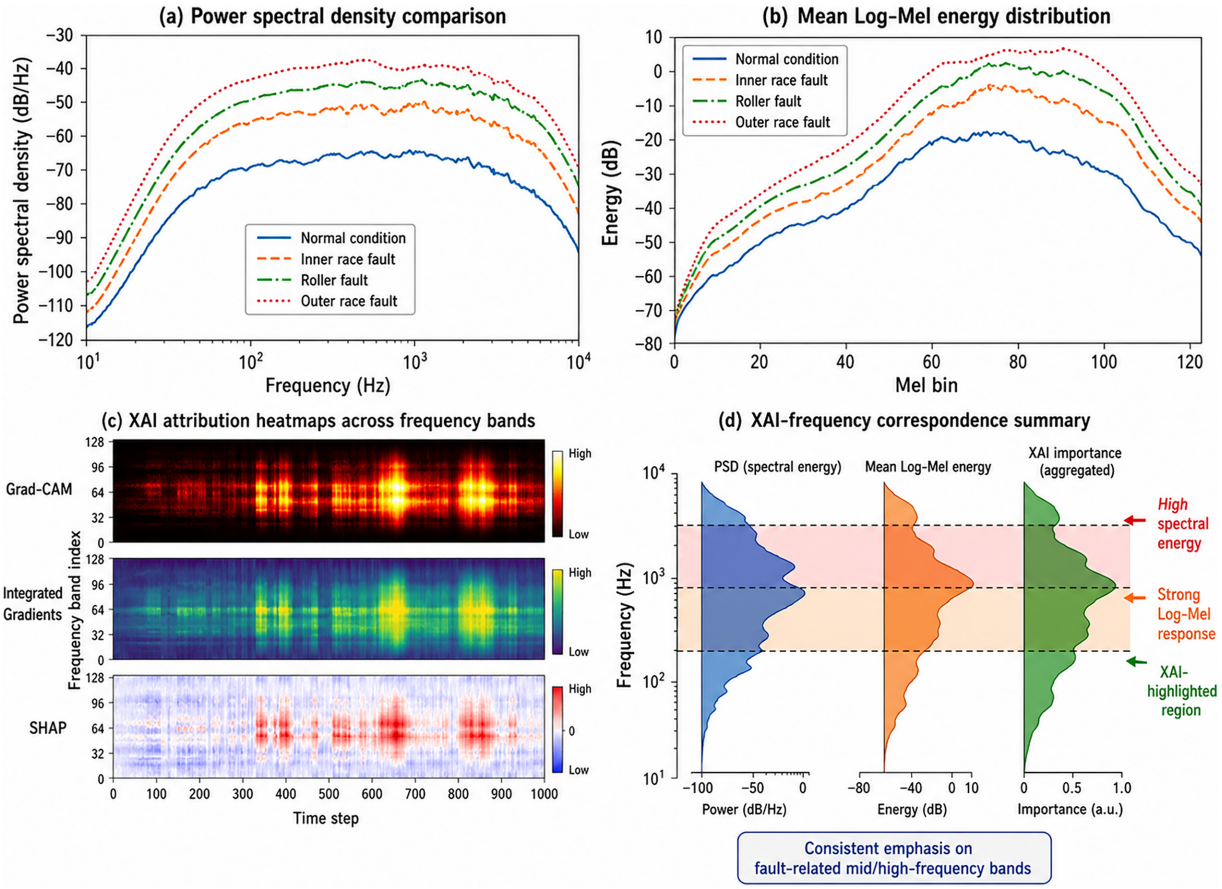


Figure 8: Correspondence between frequency-domain signal characteristics and XAI attribution regions: (**a**) power spectral density comparison, (**b**) mean Log-Mel energy distribution, (**c**) Grad-CAM, Integrated Gradients, and SHAP attribution heatmaps, and (**d**) aggregated XAI–frequency correspondence summary.

Table 10: Quantitative consistency analysis of XAI attribution maps.

Attribution Map Comparison	SSIM Score
Grad-CAM vs. Integrated Gradients	0.84
Grad-CAM vs. SHAP	0.81
Integrated Gradients vs. SHAP	0.86

5 Discussion

The experimental findings demonstrate that the proposed leakage-aware lightweight attention CNN–BiLSTM model with supervised contrastive regularization provides strong diagnostic performance for bearing fault classification using Log-Mel spectrogram representations. The model achieved an average

accuracy of 98.1% under repeated group-aware validation, outperforming the conventional CNN, LSTM, CNN–LSTM, and CNN–BiLSTM baseline models. This result indicates that combining local time–frequency feature extraction, bidirectional temporal modeling, attention-based refinement, and contrastive embedding regularization improves the discriminative capacity of the model under limited-data conditions.

5.1 Interpretation of Model Performance

The performance improvement of the proposed model can be explained by the complementary contribution of its components. The lightweight convolutional blocks extract local time–frequency patterns from Log-Mel spectrograms, which are useful for identifying fault-induced spectral variations. The BiLSTM layer captures sequential dependencies along the temporal axis and enables the model to learn bidirectional temporal relationships. The attention mechanism further improves the model by emphasizing the most informative time–frequency regions and suppressing less relevant components. In addition, supervised contrastive regularization improves the structure of the learned embedding space by encouraging samples from the same class to be closer and samples from different classes to be more separable.

The ablation analysis supports this interpretation. The CNN-only model achieved lower performance because it captured local spectral structures but did not explicitly model temporal dependencies. The CNN–LSTM and CNN–BiLSTM models improved the results by incorporating sequential information. The addition of attention and supervised contrastive regularization further increased classification accuracy, with the full model reaching the highest performance. Therefore, the improvement is not only due to replacing one architecture with another, but results from the combined effect of attention-based feature selection and contrastive class-separation learning.

5.2 Reliability of the Validation Strategy

A key concern in segmented signal-based fault diagnosis is the possibility of data leakage. If segments derived from the same original recording or highly similar neighboring windows are randomly assigned to both training and testing subsets, the reported accuracy may become overly optimistic. This issue is especially important in the present study because the dataset contains only four original recordings and 48 segmented samples. To address this limitation, repeated group-aware validation was adopted instead of relying only on a single random train–test split.

The repeated validation strategy showed that the proposed model maintained stable performance across different runs, suggesting that the reported 98.1% accuracy was not caused by a single favorable split. However, the limited number of original recordings still restricts the statistical generalizability of the results. Therefore, the reported performance should be interpreted as evidence of strong performance under controlled laboratory conditions, not as definitive proof of full real-world industrial generalization. Further validation using larger independent datasets is required.

5.3 Physical Interpretation of XAI Results

The XAI results provide physically meaningful evidence that the proposed model does not make predictions based only on arbitrary image-like patterns. Grad-CAM, Integrated Gradients, and SHAP consistently highlighted localized time–frequency regions where faulty conditions exhibited stronger spectral energy than the normal condition. These highlighted regions correspond to spectral bands in which defect-induced impacts generate repeated energy concentrations. In bearing systems, inner race faults, outer race faults, and rolling element faults can produce periodic mechanical impacts, which appear in the time–frequency domain as localized or repeated energy patterns rather than uniform broadband activity.

The attribution maps showed that the proposed model mainly focused on mid- and high-energy Log-Mel frequency regions. These regions were also emphasized in the power spectral density and mean Log-Mel energy analyses. This agreement between signal-level analysis and XAI visualization supports the physical plausibility of the learned decision mechanism. In particular, the model assigned higher importance to regions where faulty classes showed clear spectral amplification compared with the normal condition. Therefore, the visual explanations suggest that the model learned fault-related time–frequency signatures instead of relying on irrelevant background noise or random artifacts.

However, exact bearing geometry and shaft speed information are not fully available in the public dataset metadata. Therefore, the highlighted regions cannot be definitively mapped to theoretical bearing characteristic frequencies such as ball pass frequency outer race, ball pass frequency inner race, ball spin frequency, or fundamental train frequency. For this reason, the present study interprets the XAI maps as physically meaningful fault-related spectral energy regions rather than exact analytical characteristic-frequency components. Future studies should combine XAI visualization with bearing geometry and rotational speed information to establish a direct correspondence between attribution maps and theoretical fault characteristic frequencies.

5.4 Practical Implications

From a practical perspective, the proposed framework offers a compact and interpretable solution for time–frequency-based bearing fault diagnosis. Compared with highly parameterized Transformer, xLSTM, or large language model-based approaches, the proposed model has a lightweight structure and is more suitable for small-scale datasets. This is important because many laboratory fault diagnosis datasets contain limited recordings, and overly complex models may increase the risk of memorization.

The framework may be useful in predictive maintenance applications where rotating-machinery sensor signals can be converted into time–frequency representations. The use of Log-Mel spectrograms enables the model to capture spectral energy variations, while the attention and XAI components improve transparency. This interpretability is important for industrial decision-making because maintenance engineers need not only high classification accuracy but also an explanation of which signal regions contribute to the diagnostic decision.

Nevertheless, the proposed method should be viewed as a controlled-condition diagnostic framework rather than a fully validated industrial monitoring system. Real industrial environments may include stronger background noise, variable load conditions, sensor placement differences, overlapping faults, and domain shifts between machines. Therefore, additional robustness testing is necessary before deployment in practical industrial settings.

5.5 Comparison with Existing Studies

Compared with conventional CNN–LSTM-based fault diagnosis models, the proposed framework introduces three important improvements. First, the model uses a lightweight attention CNN–BiLSTM architecture instead of a standard CNN–LSTM structure, allowing both bidirectional temporal modeling and attention-based feature refinement. Second, supervised contrastive regularization is incorporated to improve embedding-level class separability under limited-data conditions. Third, repeated group-aware validation is used to reduce the risk of overly optimistic results caused by segment-level data leakage.

Recent studies have also explored advanced diagnostic models such as contrastive learning frameworks, xLSTM-based architectures, Transformer models, and industrial large language models. These approaches are promising, especially for noise robustness, cross-condition generalization, and semantic

representation learning. However, they often require larger datasets, more complex training procedures, or higher computational resources. In contrast, the proposed model provides a balanced alternative by combining high diagnostic accuracy, compact architecture, and interpretable decision analysis. Therefore, the main contribution of this study lies not only in classification performance but also in the integration of leakage-aware validation, lightweight attention-based modeling, supervised contrastive regularization, and physically interpretable XAI analysis. Table 11 compares the proposed framework with related studies in terms of model structure, feature representation, main experimental settings, reported numerical performance, and normalized performance interpretation. The proposed method differs from previous approaches by integrating Log-Mel spectrogram representation, leakage-aware lightweight attention CNN–BiLSTM modeling, supervised contrastive regularization, repeated group-aware validation, and multi-method XAI analysis. The results show that the proposed framework achieves the highest normalized performance among the compared methods while also improving interpretability and validation reliability.

Table 11: Comparison of related studies on acoustic/time–frequency-based fault diagnosis and predictive maintenance.

Study	Method/Model	Key Hyperparameters/Main Settings	Validation Strategy	Explainability/XAI	Results (Numeric)	Metric Normalization Note
Proposed solution	Log-Mel spectrogram + leakage-aware lightweight attention CNN–BiLSTM with supervised contrastive regularization	LR = 0.0005; AdamW; batch size = 8/16; epochs = 50; weight decay = 0.0001; dropout = 0.20, 0.30, 0.40; λ = 0.1	Repeated group-aware validation	Grad-CAM, Integrated Gradients, SHAP, and XAI–frequency correspondence	Acc = 98.10%; Prec = 98.35%; Rec = 98.10%; F1 = 98.08%; Macro AUC = 99.20%; MCC = 0.974	For K = 4 classes: NEBP = 0.975
[27]	Statistical time-domain features + RF, XGBoost, SVM, Naive Bayes	Not specified; feature and hyperparameter details limited	Conventional ML evaluation	Not reported	RF/XGBoost Acc = 95.73%; Prec/Rec/F1 ≈ 0.96; NEBP = 0.953	Assuming K ≈ 10 classes
[28]	MFCC/DWT features + Extreme Learning Machine; Grid Search	Signals segmented to 5 s and normalized; ELM activation/neurons optimized	5-fold cross-validation	Not reported	Best Acc = 92.17%; NEBP = 0.906	K = 6 classes
[29]	Adaptive wavelet denoising + statistical-spectral features + 1D ResNet-20	Not specified	Standard validation	Not reported	Best Acc = 91.8%; NEBP = 0.898	K = 5 classes
[30]	Envelope extraction + CWT image generation + ResNet-50 CNN	Batch size 32–64; LR = 0.1; Adam/SGD; checkpoint; convergence ≈ 4000 iterations	Validation accuracy and AUC analysis	Limited/not physically linked	Validation Acc = 82.6%; AUC = 93.1%–99.2%; NEBP = 0.768	K = 4 classes
[31]	Wavelet Packet Transform + Markov blanket feature selection + ROCKET + TPE + GBM/LightGBM	Feature reduction from 576 to 56; 70/30 train–test split	70/30 train–test split	Not reported	Best GBM test Acc = 83.0%; NEBP = 0.745	K = 3 classes

5.6 Limitations and Future Work

Despite the promising results, several limitations should be acknowledged. First, the dataset is small and contains only 48 segmented samples derived from four original recordings. Although repeated group-aware validation was used, the limited number of original recordings reduces the statistical strength of the conclusions. Second, the experiments were conducted under controlled laboratory conditions, which may not fully represent real industrial environments with variable noise, load, speed, and sensor placement conditions. Third, the exact sensing modality of the public dataset cannot be fully verified. Therefore, the data were treated as rotating-machinery time-series sensor recordings rather than being described as purely acoustic or purely vibration-based signals. Fourth, no external dataset was available for independent validation, which limits the ability to evaluate cross-domain generalization. Fifth, although XAI maps were physically interpreted using spectral energy patterns, direct matching with theoretical bearing characteristic frequencies was not possible because complete bearing geometry and shaft speed metadata were unavailable.

Future work will focus on validating the proposed framework using larger and independently collected datasets. Additional experiments under different noise levels, load profiles, speed conditions, and sensor placements should be conducted to evaluate industrial robustness. Multimodal fusion of acoustic, vibration, current, and temperature signals may further improve diagnostic reliability. In addition, physics-informed learning strategies that incorporate bearing characteristic frequencies into model training or XAI analysis could strengthen the physical interpretability of future fault diagnosis systems.

From a physical perspective, the observed variations in the acoustic signal characteristics can be attributed to defect-induced dynamic interactions within the bearing components. Faults such as inner race, outer race, and rolling element defects introduce periodic impact forces, which generate distinctive acoustic patterns and lead to changes in frequency content and energy distribution. These variations reflect the underlying mechanical behavior of the system, where localized defects tend to produce impulsive signatures, while more distributed damage may result in broader spectral responses. From a practical standpoint, these findings highlight the applicability of acoustic-based monitoring for real-world predictive maintenance. Since acoustic sensing does not require direct contact with the system, it offers a flexible and cost-effective solution, particularly in environments where vibration sensor installation is challenging. This makes the proposed approach suitable for scalable industrial applications.

6 Conclusion

This study proposed a leakage-aware lightweight attention CNN–BiLSTM framework with supervised contrastive regularization for bearing fault diagnosis using Log-Mel spectrogram representations of rotating-machinery sensor signals. The proposed model was designed to address three major concerns in small-scale fault diagnosis studies: overparameterization, segment-level data leakage risk, and limited interpretability. By combining lightweight convolutional feature extraction, bidirectional temporal modeling, attention-based refinement, and contrastive embedding regularization, the framework achieved an average accuracy of 98.1% under repeated group-aware validation. The experimental results showed that the proposed model outperformed conventional CNN, LSTM, CNN–LSTM, and CNN–BiLSTM baselines. The ablation analysis confirmed that both attention refinement and supervised contrastive regularization contributed to the final performance improvement. In addition, Grad-CAM, Integrated Gradients, and SHAP analyses demonstrated that the model focused on localized time–frequency regions associated with fault-related spectral energy variations. This supports the physical plausibility of the learned decision mechanism and improves the transparency of the diagnostic framework. The proposed method performs best under controlled laboratory conditions where fault classes are well defined and time–frequency fault signatures are sufficiently distinguishable. However, the small dataset size, limited number of original recordings, uncertainty regarding the exact sensing modality, and lack of external validation limit the generalization claims.

Future studies will focus on larger and noisier industrial datasets, cross-dataset validation, multimodal sensor fusion, and physics-informed integration of bearing characteristic frequencies into both model training and explainability analysis.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Sherif Said: Conceptualization, Investigation, Formal Analysis, Validation, Writing—Original Draft Preparation. Emrah Aslan: Methodology, Software, Formal Analysis, Validation, Visualization, Writing—Original Draft Preparation, Writing—Review and Editing. Yıldırım Özüpak: Methodology, Software, Formal Analysis, Validation, Visualization, Writing—Original Draft Preparation, Writing—Review and Editing. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is publicly available from the Kaggle repository referenced in the manuscript.

Ethics Approval: This study used a publicly available benchmark dataset and did not involve human participants, animals, or personal data. Therefore, ethical approval was not required.

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

References

1. Jiang Z, Li Y, Gao J, Wu C. A fault detection of aero-engine rolling bearings based on CNN-BiLSTM network integrated cross-attention. Meas Sci Technol. 2024;35(12):126116. doi:10.1088/1361-6501/ad7622.
2. Alsafari S, Yafoz A. Few-shot bearing fault diagnosis under joint fault-severity and load shift: a leak-free cross-domain benchmark. Comput Model Eng Sci. 2026;148(1):17. doi:10.32604/cmes.2026.084403.
3. Zou Y, Zhang Y, Li L, Liu T, Zhang X. Bearing fault diagnosis based on multimodal fusion GRU and swin-transformer. Comput Mater Contin. 2026;86(1):1–24. doi:10.32604/cmc.2025.068246.
4. Wang R, Sun S, Zhao P, Yang X, Wei X, Hu C. Multi-scale fusion network using time-division Fourier transform for rolling bearing fault diagnosis. Comput Mater Contin. 2025;84(2):3519–39. doi:10.32604/cmc.2025.066212.
5. Özüpak Y, Aslan E, Zaitsev I. Explainable LSTM-AdamW based fault diagnosis of aircraft rotating components using airborne acoustic signals under dynamic operating conditions. Sci Rep. 2026;16(1):11449. doi:10.1038/s41598-026-41889-2.
6. Xu C, Wang Z, Jin Y, Nong W. An adaptive industrial large language model for mechanical fault diagnosis under variable operating conditions. Adv Eng Inform. 2026;74(1):104821. doi:10.1016/j.aei.2026.104821.
7. Dong H, Zhang Y, Zhang B, Zhang Z, Shao Y, Yuan G, et al. A noise-robust xLSTM-based method for fault diagnosis in gas-insulated Switchgear using acoustic signals under strong noise interference. Expert Syst Appl. 2026;304(2):130827. doi:10.1016/j.eswa.2025.130827.
8. Sheng Y, Zhang B, Zhang Z, Shao Y, Yuan G, Liu J, et al. A temporal-frequency contrastive learing method for acoustic-based mechanical fault detection in gas-insulated switch-gear. Nondestruct Test Eval. 2026;41(8):4541–63. doi:10.1080/10589759.2025.2541050.
9. Qaid HAAM, Zhang B, Li D, Ng SK, Li W. FD-LLM: large language model for fault diagnosis of machines. arXiv:2412.01218. 2024.
10. Zhang Z, Liu H, Shao Y, Yang J, Liu S, Yuan G. CFENet: a contrastive frequency-sensitive learning method for gas-insulated switch-gear fault detection under varying operating conditions using acoustic signals. Eng Appl Artif Intell. 2024;135(14):108835. doi:10.1016/j.engappai.2024.108835.
11. Zhang X, Zhao Z, Tsiligkaridis T, Zitnik M. Self-supervised contrastive pre-training for time series via time-frequency consistency. arXiv:2206.08496. 2022.

12. Hafiz NFM, Mashohor S, Shazril MHSEMA, Ali AM, Rasid MFA. Machine learning framework for industrial machine sound classification in predictive maintenance. IEEE Access. 2025;13:154960–75. doi:10.1109/access.2025.3601999.
13. Lourari AW, El Yousfi B, Benkedjouh T, Bouzar Essaidi A, Soualhi A. Enhancing bearing and gear fault diagnosis: a VMD-PSO approach with multisensory signal integration. J Vibr Control. 2025;31(19–20):4098–112. doi:10.1177/10775463241273842.
14. Lourari AW, Soualhi A, Medjaher K, Benkedjouh T. New health indicators for the monitoring of bearing failures under variable loads. Struct Heal Monit. 2024;23(5):2922–41. doi:10.1177/14759217231219486.
15. Khan A, Junaid A, Siddique MF, Iqbal A, Samkari HS, Allehyani MF, et al. Smart predictive maintenance: a TCN-based system for early fault detection in industrial machinery. Machines. 2026;14(2):164. doi:10.3390/machines14020164.
16. Umar M, Ahmad Z, Ullah S, Saleem F, Farooq Siddique M, Kim JM. Advanced fault diagnosis in milling machines using acoustic emission and transfer learning. IEEE Access. 2025;13:100776–90. doi:10.1109/ACCESS.2025.3578248.
17. Lourari AW, Benkedjouh T, El Yousfi B, Soualhi A. ANFIS-based framework for the prediction of bearing's remaining useful life. Int J Progn Health Manag. 2024;15(1):1–13. doi:10.36001/ijphm.2024.v15i1.3791.
18. Siddique MF, Umar M, Ahmad W, Kim JM. Advanced fault diagnosis in milling cutting tools using vision transformers with semi-supervised learning and uncertainty quantification. Sci Rep. 2025;15(1):42460. doi:10.1038/s41598-025-26550-8.
19. Engine acoustic emissions [Internet]. [cited 2026 Mar 8]. Available from: https://www.kaggle.com/datasets/julienjta/engine-acoustic-emissions/data.
20. Suawa PF, Halbinger A, Jongmanns M, Reichenbach M. Noise-robust machine learning models for predictive maintenance applications. IEEE Sens J. 2023;23(13):15081–92. doi:10.1109/JSEN.2023.3273458.
21. Aslan E, Özüpak Y. Sensor-based bearing fault diagnosis using a hybrid CNN–LSTM framework with explainable AI. Nondestruct Test Eval. 2026. doi:10.1080/10589759.2026.2689448.
22. Aslan E, Özüpak Y. Explainable CNN–GRU learning on Mel-spectrogram acoustic signals for bearing fault diagnosis under small-sample experimental conditions. Proc Inst Mech Eng Part G J Aerosp Eng. 2026. doi:10.1177/09544100261470961.
23. Yan H, Zhou H, Zheng J, Zhou Z. Rolling bearing fault diagnosis based on 1D convolutional neural network and Kolmogorov–Arnold network for industrial Internet. Comput Mater Contin. 2025;83(3):4659–77. doi:10.32604/cmc.2025.062807.
24. Aradi A, Varga AK. Enhancing predictive maintenance in industrial systems through acoustic monitoring of servo motors using machine learning (AI). In: Intelligent and fuzzy systems. Cham, Switzerland: Springer Nature; 2024. p. 690–7. doi:10.1007/978-3-031-67195-1_76.
25. Lourari AW, Soualhi A, Benkedjouh T. Advancing bearing fault diagnosis under variable working conditions: a CEEMDAN-SBS approach with vibro-electric signal integration. Int J Adv Manuf Technol. 2024;132(5):2753–72. doi:10.1007/s00170-024-13458-2.
26. Lourari AW, El Yousfi B, Essaidi AB. Enhanced diagnosis of bearing and gear faults using Hilbert-Huang transform, singular value decomposition, and supervised learning methods. Int J Adv Manuf Technol. 2025;139(1):983–99. doi:10.1007/s00170-025-15970-5.
27. Akcan E. Detection of bearing faults from vibration signals. Balk J Electr Comput Eng. 2025;13(3):295–306. doi:10.17694/bajece.1757057.
28. Akbalık F, Yıldız A, Ertuğrul ÖF, Zan H. Engine fault detection by sound analysis and machine learning. Appl Sci. 2024;14(15):6532. doi:10.3390/app14156532.
29. Jung W, Bae J, Park YH. Acoustic signal based non-contact ball bearing fault diagnosis using adaptive wavelet denoising. arXiv:2110.03348. 2021.
30. Nashed MS, Renno J, Mohamed MS, Reuben RL. Gas turbine failure classification using acoustic emissions with wavelet analysis and deep learning. Expert Syst Appl. 2023;232:120684. doi:10.1016/j.eswa.2023.120684.
31. Tuleski BL, Yamaguchi CK, Stefenon SF, dos Santos Coelho L, Mariani VC. Audio-based engine fault diagnosis with wavelet, Markov blanket, ROCKET, and optimized machine learning classifiers. Sensors. 2024;24(22):7316. doi:10.3390/s24227316.