An Enhanced Lightweight Multi-scale CNN with Resonance-dominant Frequency-band Learning for Robust Bearing Fault Diagnosis

Duy An Huynh[1,3], Thanh Tung Luu[1,3], Thanh Hai Nguyen[2,3]

[1]*Department of Construction Machinery and Handling Equipment, Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, 700000, Vietnam*

[2]*Department of Material Processing Technology, Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, 700000, Vietnam*

[3] *Vietnam National University Ho Chi Minh City, Ho Chi Minh City, 700000, Vietnam*

hdan.sdh232@hcmut.edu.vn
ttluu@hcmut.edu.vn
haint@hcmut.edu.vn

ABSTRACT

Rolling bearings are essential components in rotating machinery, and reliable early fault detection is crucial for avoiding unexpected failures and reducing maintenance costs. While deep learning techniques have achieved notable progress, most prior studies have emphasized feature enrichment without thoroughly examining vibration spectral characteristics or establishing mechanisms to select the most informative time–frequency patterns. To exploit this physical insight, this study proposes an enhanced lightweight multi-scale convolutional neural network (CNN) with adaptive frequency-band learning for robust bearing fault diagnosis. Vibration signals are first band-limited to the resonance-dominant range and transformed using the Continuous Wavelet Transform with the Morlet wavelet, improving signal-to-noise ratio while preserving impulsive fault components. The proposed architecture integrates depthwise separable convolutions, multi-scale parallel branches, Squeeze-and-Excitation channel recalibration, and Atrous Spatial Pyramid Pooling to capture discriminative resonance-sensitive features with reduced computational complexity. Grad-CAM is employed to visualize the resonance bands contributing to classification decisions. Experiments on the CWRU dataset achieve accuracy above 99% with only 7.69 M parameters. Robustness under additive noise and cross-domain validation on the HUST dataset confirm strong generalization and domain invariance.

Keywords: Rolling bearing; Pattern recognition; Fault diagnosis; Multi-scale convolution; Enhanced lightweight model.

Huynh et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

https://doi.org/10.36001/IJPHM.2026.v17i2.4757

1. INTRODUCTION

Among the various components of rotating machinery, rolling-element bearings are particularly crucial because they support rotating shafts and enable smooth mechanical motion under complex operating conditions. By reducing friction and mechanical wear between interacting components, bearings contribute significantly to improving energy efficiency and maintaining the overall reliability of mechanical systems.

Once bearing defects occur, the resulting vibration and dynamic responses can propagate through the entire mechanical structure, potentially leading to severe damage, unexpected machine shutdowns, and costly production losses. For this reason, bearing fault diagnosis has attracted extensive attention from both academia and industry in recent years (Chen, Cen, & Xiong, 2020; Zheng, Fu, Lu, & Zhu, 2021).

Existing approaches to bearing fault diagnosis can be broadly categorized into three groups, each presenting distinct advantages and limitations. The first group encompasses traditional signal processing methods, including envelope analysis, Short-Time Fourier Transform (STFT), Empirical Mode Decomposition (EMD), and Variational Mode Decomposition (VMD), often combined with handcrafted statistical features such as root mean square, kurtosis, and crest factor (Lei, Lin, He, & Zuo, 2013; Randall & Antoni, 2011). These techniques offer strong physical interpretability and remain effective under limited data conditions; however, their diagnostic performance is heavily dependent on domain expertise and prior knowledge of fault characteristics, making them difficult to scale across heterogeneous mechanical systems. The second group comprises classical machine learning methods, such as Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors (kNN), which operate on manually engineered feature sets extracted

from vibration signals (Cerrada et al., 2016; Widodo & Yang, 2007). While these methods perform reliably on small datasets, the labor-intensive feature engineering process and limited generalization capacity across varying operating conditions constrain their practical applicability. The third group, represented by deep learning architectures including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models, has demonstrated remarkable capability in automatically learning discriminative representations directly from raw or minimally preprocessed signals (Chen et al., 2022; Zhao et al., 2019). Despite achieving superior classification accuracy, these models often suffer from high computational complexity, large parameter counts, and limited interpretability — challenges that hinder their deployment in real-time or resource-constrained industrial environments.

Among these three paradigms, deep learning — and CNN-based architectures in particular — has emerged as the dominant approach due to its ability to extract hierarchical fault-related features without manual intervention. Nevertheless, the persistent trade-off between model complexity, diagnostic interpretability, and deployability in edge computing scenarios remains an open and practically significant challenge, motivating the development of lightweight yet explainable diagnostic frameworks.

Nowadays, deep learning techniques have emerged as effective solutions for intelligent fault diagnosis. In particular, convolutional neural networks (known as CNNs) have shown strong capability in automatically learning discriminative patterns from raw vibration signals or their frequency-domain spectral representations. A number of studies have explored different CNN-based frameworks, including 1D-CNN models designed for time-domain vibration analysis (Eren, Ince, & Kiranyaz, 2019; Zhang, Li, Peng, Chen, & Zhang, 2018), 2D-CNN approaches that utilize time–frequency images generated by spectrograms or wavelet transforms (Toma et al., 2022; Wang, Feng, Huo, & Kang, 2022), as well as hybrid CNN architectures that integrate signal transformations or feature fusion strategies (Deveci et al., 2024; Jia, Lei, Lin, Zhou, & Lu, 2016). Although 1D-CNNs provide relatively low computational cost, they may not sufficiently capture frequency-related characteristics that are important for accurate fault recognition. By contrast, 2D-CNNs trained on time–frequency representations, such as those derived from continuous wavelet transform (CWT) or short-time Fourier transform (STFT), can encode richer diagnostic information. Nevertheless, these networks typically employ convolutional kernels with fixed receptive fields, which may limit their adaptability to faults exhibiting different frequency distributions. To address this limitation, multi-scale CNN structures have been proposed to enhance feature extraction capability by employing parallel convolutional branches with different kernel sizes. For example, Zhao, Zhang, and Xue (2024) developed a multi-scale and multi-channel CNN combined with spatial attention for hyperspectral image classification, achieving improved performance compared with conventional CNN models. Other related studies have investigated attention-based CNN frameworks (Wang, Hua, Xu, & Zhao, 2023), CNN–LSTM hybrid networks (Zhou, Ma, & Chen, 2022), and approaches that integrate deep learning with signal decomposition methods such as EEMD (Rabah & Abdelhafid, 2018) or MSDARN (Liang, Cao, & Zhao, 2022). Despite these advances, many of these models tend to increase architectural complexity while simultaneously reducing interpretability.

To address these limitations, Depthwise Separable Convolutional Neural Networks (DS-CNNs) are one of the most promising alternatives for reducing computational complexity while preserving classification performance. Unlike conventional convolutions, DS-CNNs factorize the convolution operation into depthwise and pointwise steps, significantly lowering the number of trainable parameters and enabling efficient feature extraction. Liang, Chen, and Chung (2024) successfully implemented a 1D DS-CNN for bearing fault diagnosis on FPGA, demonstrating its suitability for real-time and resource-constrained applications. Meanwhile, Xu, Xiao, Sun, and Cui (2024) introduced an improved DS-CNN structure tailored for transfer learning in fault diagnosis, achieving better generalization and robustness when only limited training data is available. These advancements motivate the integration of DS-CNN into multi-branch diagnostic models, particularly for scenarios requiring both high interpretability and lightweight design. Among popular attention mechanisms, the Squeeze-and-Excitation (SE) block has been shown to effectively enhance channel feature representations in bearing fault diagnosis. By integrating SE into TCN, Wu et al. (2025) highlighted key frequency-domain features and improved SVM classification. Gu, Tian, Li, Wei, and Li (2024) applied SE-ResNet to acoustic–vibration fusion, where SE optimized channel selection for composite fault recognition. Wang et al. (2025) further incorporated SE into KAN convolution with dual-branch attention, achieving robust accuracy under variable conditions. Together, these studies demonstrate that SE improves both diagnostic accuracy and model adaptability. However, although earlier works have employed multi-scale convolutional kernels to enrich feature extraction, they have not provided a clear explanation of how fault-related frequency components appear within the extracted representations. Moreover, these studies have not proposed any mechanism for determining which feature maps contribute most effectively to distinguishing individual bearing fault types.

In addition to classification accuracy, model transparency and interpretability are essential for real-world fault diagnosis. Explainable AI (known as XAI) methods based Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al., 2017), allow users to visualize the regions in the input that influence model decisions. Studies like Jiang et al. (2025)

have demonstrated the effectiveness of Grad-CAM in analyzing bearing faults, especially in revealing resonance bands critical for classification.

This study introduces a novel Multi-Scale Channels Deep Neural Network (MSCNN) architecture enhanced with Squeeze-and-Excitation (SE) blocks, an Atrous Spatial Pyramid Pooling (ASPP) module, and Grad-CAM. This integrated framework—hereafter referred to as the Lightweight Hybrid-CNN—is designed to strengthen feature extraction while simultaneously improving interpretability in bearing fault diagnosis. By integrating multi-scale convolutional branches, channel-wise recalibration, and contextual feature aggregation within a unified CNN architecture, the proposed framework effectively captures fault-related frequency components that are intrinsically governed by bearing geometry and rotational speed. Unlike system-dependent characteristics—such as load variations or structural resonance effects—these intrinsic fault-frequency patterns exhibit temporal stability and physical consistency. The collaborative interaction among the embedded modules promotes the extraction of invariant spectral representations, reducing sensitivity to operating-condition fluctuations. Consequently, the constructed hybrid CNN structure enhances adaptability and supports transferable learning across heterogeneous mechanical systems. This capability is consistent with the primary objective of advanced fault diagnosis frameworks, which aim to achieve robust generalization and domain-invariant feature learning under varying industrial environments.

Compared to existing approaches, the proposed model offers the following contributions:

1. Lightweight multi-scale architecture: A novel hybrid CNN combining depthwise separable convolutions, multi-kernel branches, SE attention, and ASPP to efficiently capture fault features across multiple frequency scales with only 7.69M parameters.
2. Integrated interpretability: Grad-CAM is applied to CWT spectrograms, providing intuitive visualization of discriminative time–frequency regions and confirming physically meaningful resonance-band attention.
3. High accuracy with reduced complexity: The proposed model attains >99% accuracy on the CWRU benchmark while remaining suitable for real-time or embedded deployment, outperforming heavier baselines on the accuracy-complexity trade-off.
4. Cross-domain generalization: Transfer evaluation on the independent HUST dataset demonstrates domain-invariant feature learning, achieving over 99% accuracy including on compound fault classes rarely addressed in existing work.

It is worth emphasizing that the primary contribution of this work is of an engineering nature rather than a theoretical one. The novelty lies in the principled integration of resonance-aware preprocessing, multi-scale depthwise separable convolution, channel-wise recalibration, and contextual feature aggregation within a unified diagnostic framework — a combination guided by the physical understanding of bearing vibration mechanics that, to the best of the authors' knowledge, has not been previously explored in this form.

The remainder of this paper includes: Section 2 reviews foundational concepts including wavelet transformation, Multi-scale channel Deep CNN, Depthwise Separable Convolutional Neural Networks, SE block, ASPP and Grad-CAM. Section 3 details the proposed architecture and diagnostic procedure. Section 4 presents experimental results and comparisons. Section 5 validates cross-domain deep transfer diagnosis. Section 6 concludes the paper and suggests directions for future research development.

2. SEVERAL BASIC THEORIES

2.1. Characteristic Frequencies of Bearing Faults

Faults occurring at specific bearing components produce localized impact events. During operation, defective rolling elements repeatedly interact with damaged regions along the raceways, giving rise to periodic impulsive excitations (Aziz et al., 2025; Liefstingh, Taal, Restrepo, & Azarfar, 2021). These impulses subsequently stimulate high-frequency resonant responses within the bearing assembly and measurement system.

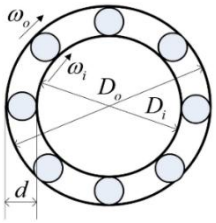


Figure 1. The structure of a typical rolling bearing.

Figure 1. illustrates the geometric configuration of a rolling bearing. Let D denote the pitch diameter (mid-diameter), d the roller diameter, and α the contact angle. The outer raceway diameter D_o and inner raceway diameter D_i can be expressed respectively as:

$$D_o = D + d\cos(\alpha) \quad (1)$$

$$D_i = D - d\cos(\alpha) \quad (2)$$

The linear circumferential velocity of the outer ring is related to its angular velocity ω_o by:

$$V_o = \frac{1}{2} D_o \omega \quad (3)$$

$$V_i = \frac{1}{2} D_i \omega_i \quad (4)$$

The cage velocity is determined by averaging the linear velocities of the inner and outer rings:

$$V_c = \frac{1}{2}(V_o + V_i) \tag{5}$$

Accordingly, the cage rotational frequency can be derived as:

$$f_c = \frac{1}{2} f_o (1 + \frac{d}{D}\cos(\alpha)) + \frac{1}{2} f_o (1 + \frac{d}{D}\cos(\alpha)) \tag{6}$$

The relative rotational frequency of the cage with respect to the outer ring becomes:

$$f_{ro} = \frac{1}{2}(f_o - f_i)\left(1 - \frac{d}{D}\cos(\alpha)\right) \tag{7}$$

Similarly, relative to the inner ring:

$$f_{ri} = \frac{1}{2}(f_o - f_i)\left(1 + \frac{d}{D}\cos(\alpha)\right) \tag{8}$$

The roller spin frequency is obtained from:

$$f_{spin} = \frac{D}{2d}(f_o - f_i)\left(1 - \left(\frac{d}{D}\cos(\alpha)\right)^2\right) \tag{9}$$

Assuming the bearing contains n rolling elements, the characteristic fault frequencies corresponding to the outer and inner raceways are expressed as:

$$f_{BPFO} = \frac{n}{2}(f_o - f_i)\left(1 - \frac{d}{D}\cos(\alpha)\right) \tag{10}$$

$$f_{BPFI} = \frac{n}{2}(f_o - f_i)\left(1 + \frac{d}{D}\cos(\alpha)\right) \tag{11}$$

Under the common condition where the outer ring is stationary and the inner ring rotates at frequency f_r, the cage rotational frequency simplifies to:

$$f_c = \frac{f_r}{2}\left(1 - \frac{d}{D}\cos(\alpha)\right) \tag{12}$$

The corresponding ball spin frequency and defect frequencies are then derived accordingly. These formulations are established under the assumption of pure rolling contact without slip between rolling elements and raceways. In practical measurements, slight deviations may occur due to frictional effects and structural dynamics.

2.2. Wavelet Transformation and Bearing Fault Characteristic Frequencies

The wavelet transform (WT) decomposes a signal into a series of basis functions referred to as wavelets (Peng & Chu, 2004). These functions are derived from a fundamental waveform known as the mother wavelet. By applying scaling and translation operations to the mother wavelet, a set of wavelet functions with different resolutions can be generated. This property enables the WT to analyze signals at multiple time–frequency scales. Formally, the WT transformation of a signal *x(t)* is expressed as:

$$W_x(a,b) = \frac{1}{\sqrt{|a|}}\int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{13}$$

In this formulation, $W_x(a,b)$ represents the wavelet coefficient corresponding to the input signal, $\psi(t)$ denotes the mother wavelet used to generate the wavelet family correspond to the scaling and shifting factors that control the dilation and temporal localization of the wavelet function, and $\psi^*(t)$ denotes the complex conjugate of the mother wavelet.

Vibration signals generated by rolling bearings are non-stationary and contain transient impulses associated with fault characteristic frequencies (FCFs). To extract these components while suppressing background interference, this work adopts the Continuous Wavelet Transform (CWT) for time–frequency decomposition. The CWT highlights short impulsive events while suppressing low-frequency structural noise and high-frequency sensor disturbances. The Morlet wavelet ("morl") is adopted as the mother wavelet owing to its excellent joint time–frequency localization for non-stationary vibration responses.

Although FCFs can be derived analytically from the bearing geometry and shaft rotational speed, as described in Section 2.1, defect-induced impulses are rarely observed as sharp spectral peaks at their nominal FCF values. Owing to structural transfer paths, damping, and sensor coupling, the fault energy is redistributed and amplified within the resonance regions of the housing–shaft–bearing assembly. A direct frequency matching strategy based on idealized physical models may therefore fail to capture the most informative diagnostic signatures.

This redistribution can be described as an amplitude-modulation process. Each impact between a rolling element and a localized defect excites the natural frequencies of the bearing and its housing, which decay quickly because of structural damping. As the impacts repeat at the FCF, the measured signal becomes a high-frequency resonance carrier whose amplitude is modulated at the FCF, so the fault information appears mainly as sidebands around the resonance frequencies rather than as a strong peak at the FCF itself. In the low-frequency region, these weak components are further masked by shaft harmonics, misalignment, and gear meshing, whereas the resonance region offers a higher fault-to-noise ratio.

To mitigate this limitation, a resonance-oriented frequency selection is introduced before feature extraction: the vibration signals are filtered within the range of 500 Hz to 10 kHz prior to performing the CWT. This resonance-dominant band contains the most informative fault-related responses generated by structural resonances of the bearing system, while low-frequency drift components and high-frequency sensor disturbances are suppressed. The range is also compatible with the sampling configurations of widely used bearing datasets, including CWRU, IMS, and HUST, preserving the principal resonance responses while preventing aliasing artifacts.

Focusing on this resonance-sensitive interval improves both the signal-to-noise ratio and the domain invariance of the extracted spectral representations. By emphasizing physically meaningful resonance responses rather than nominal FCF peaks, the resulting CWT-based time–frequency maps provide stable and transferable features, which supports generalization across varying operating conditions and heterogeneous bearing systems.

The scale parameter a in Eq. (13) is directly linked to frequency: a small scale compresses the wavelet to capture high-frequency content, while a large scale stretches it to capture low-frequency content. For the Morlet wavelet, the pseudo-frequency of each scale is computed from the center frequency of the wavelet and the sampling rate. Only the scales covering the band from 500 Hz to 10 kHz are computed, which reduces the size of the coefficient matrix and the computation time.

The magnitude of the CWT coefficients forms a scalogram, with time on the horizontal axis and frequency on the vertical axis. For a healthy bearing, the energy is spread smoothly over time; for a faulty bearing, short bright stripes appear at regular intervals, caused by impacts between the rolling elements and the defect. The time gap between two stripes is close to the inverse of the FCF, so the scalogram keeps a clear link between the image pattern and the physical fault mechanism. Before training, each scalogram is normalized to a fixed value range and resized to the input size of the network, giving the same input scale for all samples and datasets.

The Morlet wavelet is chosen because its shape resembles the decaying oscillation produced by a single fault impact, so it responds strongly when an impact occurs; alternatives such as the Mexican hat or Daubechies wavelets often match the impulse shape of bearing signals less well. The CWT is also preferred over the short-time Fourier transform, whose fixed window length cannot provide good time and frequency resolution at the same time. By using short windows at high frequencies and long windows at low frequencies, the CWT suits bearing signals in which short impacts excite high-frequency resonances. Unlike empirical mode decomposition, it also avoids mode mixing and end effects. The same CWT settings are used for the CWRU and HUST datasets, with only the scale range adjusted to the sampling rate of each dataset, which keeps the preprocessing consistent and the cross-domain test in Section 5 fair.

Several practical factors affect the quality of the scalogram. First, each signal segment must be long enough to include several fault impacts but short enough to limit computation and avoid mixing operating states; in this study, each segment covers several shaft revolutions. Second, the number of scales trades frequency resolution against image size, so a balanced number is used to preserve narrow resonance bands while maintaining speed. Third, strong noise can hide weak fault impacts, which the band-pass filter mitigates by removing the noise energy outside the resonance band. Fourth, changes in speed and load shift the fault frequencies, whereas the resonance band of the structure stays almost the same, making the resonance-based approach less sensitive to operating conditions. Together, these choices keep the input images closely linked to the physical behavior of the bearing, helping the network focus on real fault features instead of random patterns.

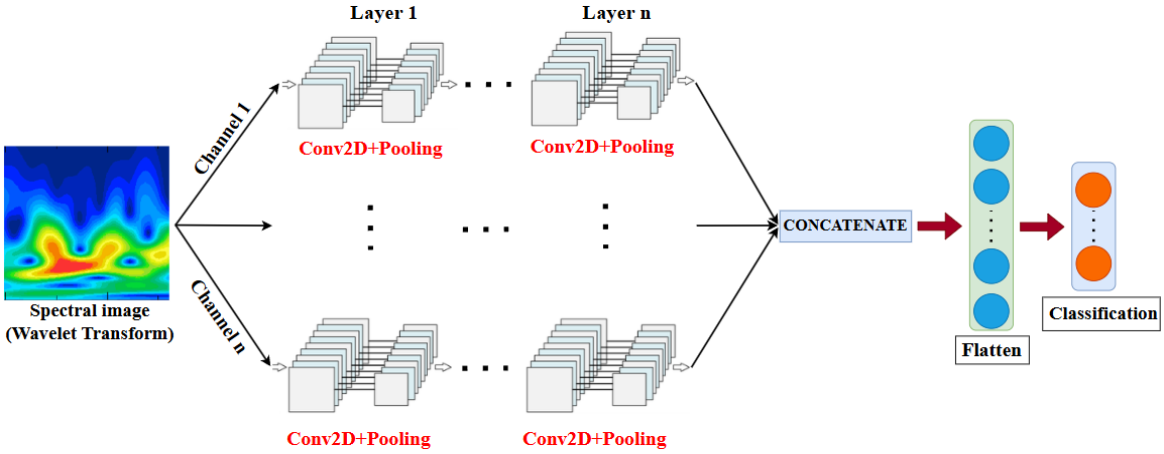


Figure 2. Multi-channel convolution neural network.

2.3. Multi-Scale Channel Deep CNN

One of the most recent advancements in CNN architecture is the introduction of multi-scaled channels, which allows the network to process input data at different scales simultaneously. This approach enhances the model's ability to capture features across various resolutions, making it particularly effective for complex and non-stationary data, such as vibration signals. For instance, Liefstingh et al. (2021) introduced two single-channel CNN architectures to develop a fault diagnosis system for the CWRU dataset. These architectures extract features from 2D spectrum images derived from applying Fast Fourier Transform (FFT) to the vibration signals. The best classification accuracy is 97.7%, with a standard deviation of 3.91%. On the fixed train-test split, the model attained a classification accuracy of 96.9%. Other researchers (Huang et al., 2024; Pandhare, Singh, & Lee, 2019; Peng & Chu, 2004; Wang et al., 2023) also have a similar or lower accuracy proportion with the simple structure of 2D-CNN or 1D-CNN. In order to reach higher accuracy, several researchers proposed another method such as getting more channels from several same-function sensors (Li, Zhao, Liu, & Fei, 2022). However, these methods lead to complexity for the model. Instead, Multi-channels Deep CNN (Figure 2) has raised awareness of the effectiveness of flexible feature extraction mechanisms for objects with both temporal and spatial relationships with better accuracy. By employing multiple parallel convolutional channels with different receptive fields, the network can capture complementary features at various scales and characterize complex patterns more comprehensively.

Mathematically, given an input feature map F (or image) with dimensions $H \times W$ (Height, Width), a multi-scaled convolutional layer applies several

convolutions with varying parallel kernel sizes $k_1 \times k_1, k_2 \times k_2, ..., k_n \times k_n$ .The output feature maps from these different convolutions are then concatenated or summed to form a combined feature map, which is passed to the next layer. For a convolution operation using a kernel of size k, the transformation can be formulated as:

$$F'_k = \mathrm{ReLU}\left(\mathrm{Conv}(F, W_k) + b_k\right) \tag{14}$$

Symbol	Description
F'_k	Feature map produced by convolution with kernel size k
W_k	Trainable filter weights for kernel size k
b_k	Bias parameter for kernel size k
ReLU	Rectified Linear Unit activation function

Table 1. Notation summary for Eq. (14)

The final representation generated by the multi-scale convolutional layer is obtained by combining the outputs produced by all convolution kernels with different sizes:

$$F' = Concat(F'_{k_1}, F'_{k_2}, \ldots, F'_{k_n}) \tag{15}$$

This allows the network to retain rich multi-scale information, improving its ability to detect and classify features that may vary in size or frequency.

2.4. Depthwise Separable CNN

DS-CNNs (Liang et al., 2024; Xu et al., 2024) are an efficient variant of traditional convolutional networks, designed to reduce the model's complexity without significantly sacrificing performance. The architecture of DS-CNNs includes two separate stages: depthwise convolution and pointwise convolution, as illustrated in Figure 3.

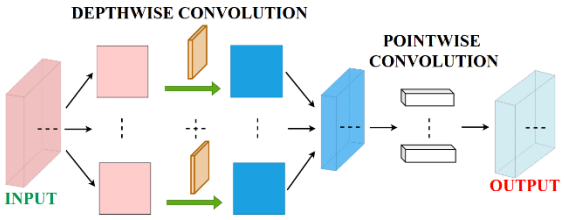


Figure 3. Depthwise Separable Convolutional Neural Networks.

In a standard convolutional layer, each filter functions across all input channels, simultaneously consolidating spatial and cross-channel information. This leads to an extensive array of parameters and significant computational complexity, particularly when addressing deep systems. Conversely, depthwise separable convolution divides this process into two steps:

Depthwise Convolution: This operation applies a singular spatial filter independently to each input channel, rather than utilizing a multi-channel filter across the entire input volume. For an input with D_{in} channels, D_{in} depthwise filters are employed, each tasked with capturing spatial features within an individual channel. This significantly reduces the number of multiplications compared to standard convolutions

Pointwise Convolution: After the depthwise convolution stage, a 1×1 convolution—commonly referred to as pointwise convolution—is employed to aggregate the feature maps generated by the depthwise operation across different channels. This step enables the model to capture dependencies among channels and construct more informative feature representations.

From a mathematical perspective, suppose the input feature map has dimensions $H \times W \times D_{in}$, and the objective is to generate an output with D_{out} channels using $K \times K$ filters. In a conventional convolutional layer, the total number of parameters can be expressed as:

$$\text{Standard Conv}: K \times K \times D_{in} \times D_{out} \quad (16)$$

Whereas in depthwise separable convolution, the number of parameters becomes:

$$\text{Depthwise: } K \times K \times D_{\text{in}} \quad (17)$$

$$\text{Pointwise: } 1 \times 1 \times D_{\text{in}} \times D_{\text{out}} \quad (18)$$

$$\text{Total: } K^2 \cdot D_{in} + D_{out} \cdot D_{out} \quad (19)$$

2.5. Squeeze-and-Excitation (SE) Block

Convolutional Neural Networks (CNNs) have been the cornerstone of numerous computer vision tasks due to their ability to extract local spatial and channel-wise features. While extensive research has explored the enhancement of spatial encodings, the explicit modeling of inter-channel relationships has remained relatively underexplored.

To address this gap, the Squeeze-and-Excitation (SE) Block architecture (Figure 4) was proposed to improve the representational capacity of CNNs through a lightweight dynamic channel-wise recalibration mechanism.

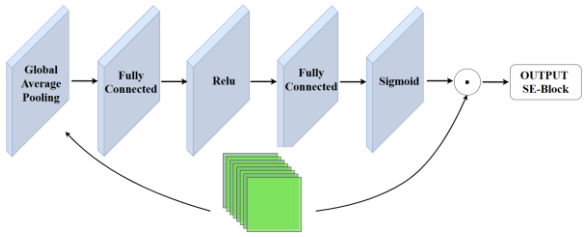


Figure 4. Squeeze-and-Excitation (SE) Block

The SE block operates in two main stages: Squeeze and Excitation.

The first step, the squeeze step, aims to embed global spatial information into a compact channel descriptor. This is achieved via global average pooling, which condenses the spatial dimensions of each channel

$U \in \mathbb{R}^{H \times W \times C}$ into a scalar statistic, resulting in a descriptor vector $z \in \mathbb{R}^{C}$:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \quad (20)$$

This vector captures the global context of each channel, enabling the network to leverage long-range dependencies beyond local receptive fields.

Subsequently, the excitation phase models nonlinear inter-channel dependencies based on the aggregated global context. This is implemented via a gating mechanism comprising two fully connected (FC) layers with a bottleneck structure:

$$s = \sigma\left(W_2 \cdot \delta\left(W_1 z\right)\right) \quad (21)$$

Symbol	Description
δ	ReLU activation function
σ	Sigmoid activation function
W_1, W_2	Learnable weight matrices of FC layers
z	Channel descriptor vector from squeeze step

Table 2. Notation summary for Eq. (21)

2.6. Atrous Spatial Pyramid Pooling (ASPP)

ASPP extends the concept of spatial pyramid pooling (SPP) by incorporating atrous (dilated) convolutions with varying dilation rates. Atrous convolution introduces zeros between kernel weights, effectively expanding the receptive field without increasing the number of parameters. This mechanism allows the model to sample features at multiple spatial scales in parallel, capturing both fine details and global context.

As illustrated in Figure 5, ASPP applies parallel atrous convolutions with different dilation rates (e.g., 6, 9, 12), all operating on the same input feature map. Each branch captures features at a different effective field of view. These multiscale features are then concatenated and fused to enhance the representational power of the network.

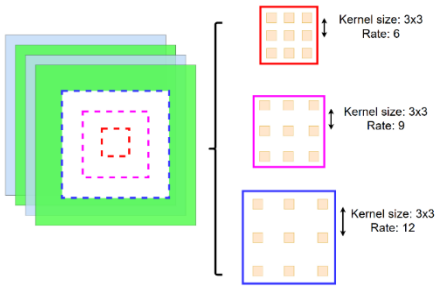


Figure 5. Atrous Spatial Pyramid Pooling mechanism

The larger the dilation rate, the wider the context the branch can capture. This multiscale context is especially beneficial in handling objects of varying sizes in dense prediction tasks.

Mathematically, an atrous convolution operation with dilation rate r and kernel size k on a 1D signal x is defined as:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] \cdot \omega[k] \quad (22)$$

where ω denotes the filter weights. In 2D, the operation extends similarly to handle spatial dimensions, enabling receptive fields to grow without loss of resolution.

Compared to traditional SPP which uses pooling at different scales, ASPP preserves spatial resolution and is more suitable for dense prediction tasks such as semantic segmentation, fault diagnosis, and medical imaging. Moreover, ASPP can be seamlessly integrated into fully convolutional architectures and adapted to lightweight backbones for real-time applications.

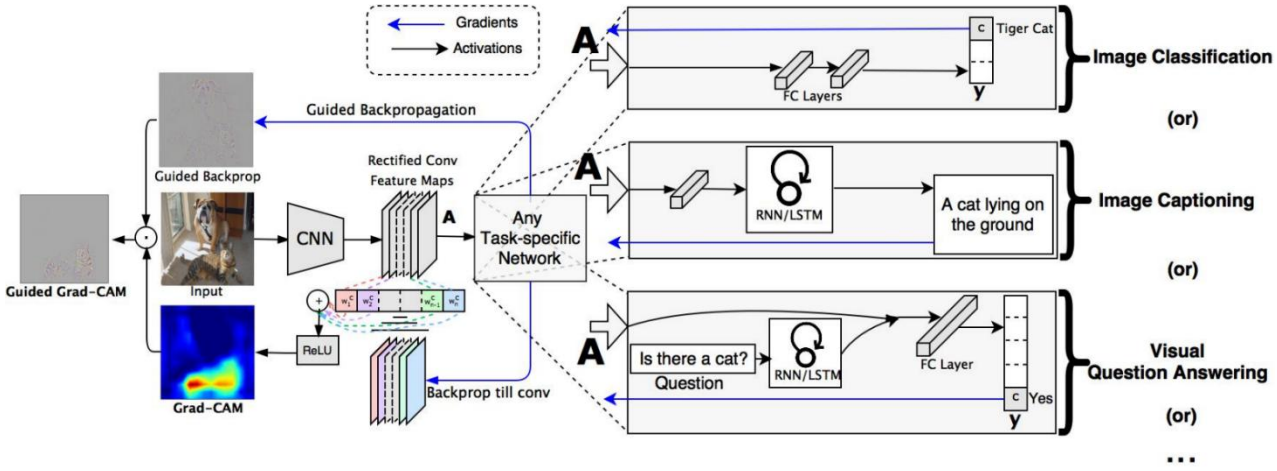


Figure 6. General integrated Grad-CAM and CNN architecture (Selvaraju et al., 2017)

2.7. Attention Mechanism: Grad-CAM

Grad-CAM (Selvaraju et al., 2017) provides a visual explanation that highlights the regions in the input feature map that are most influential in the model's decision (Figure 6). Unlike traditional attention mechanisms that explicitly modify the architecture of the network, Grad-CAM operates *post hoc*, allowing it to be applied to any model architecture without modifications.

When applied to CNNs, Grad-CAM generates heatmaps that highlight the areas in the input data - whether images, spectrograms, or time-series signals - that the network considers important for classification. Grad-CAM can be effectively applied to vibration data analysis, particularly in the context of fault diagnosis in rolling bearings and other rotating machinery.

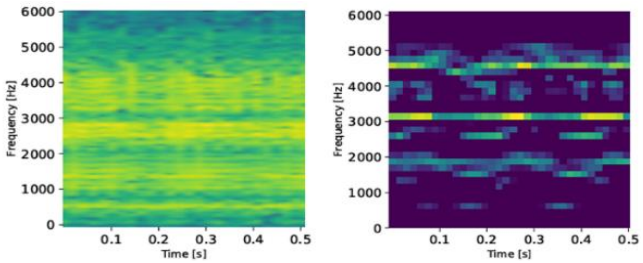


Figure 7. Abnormal diagnostic in a bearing inner race fault with Grad-CAM attention map

Figure 7 displays examples of two spectrograms along with their corresponding Grad-CAM activation maps (Liefstingh et al., 2021). The activation maps suggest the potential significance of the system's resonances in influencing the classification results. These resonances, marked by high energy concentrations (represented by yellow) in specific frequency bands, are primarily related to the system's transfer function parameters rather than being exclusively linked to the excitation signals generated by bearing defects.

3. PROPOSED METHOD FOR FAULT DIAGNOSTIC WITH LIGHTWEIGHT HYBRID-CNN AND GRAD-CAM

3.1. Proposed Model Architecture

To enhance diagnostic performance while reducing computational cost, the proposed architecture replaces standard CNN layers with DS-CNN layers. This combination significantly reduces the number of trainable parameters without sacrificing accuracy. The proposed Lightweight Hybrid-CNN structure is illustrated in Figure 8.

The above architecture uses multi-channel spectral images, derived from continuous wavelet transforms, as input. Each channel is processed through a multi-scale convolutional branch with different kernel sizes (5×5, and 11×11), which are investigated through experiments to capture effectively all failure features on one rotational cycle of the bearing, to capture diverse fault features. In contrast to traditional CNNs, each branch in our model uses depthwise separable 2D convolutional layers (DS-Conv2D), dropout, and max pooling.

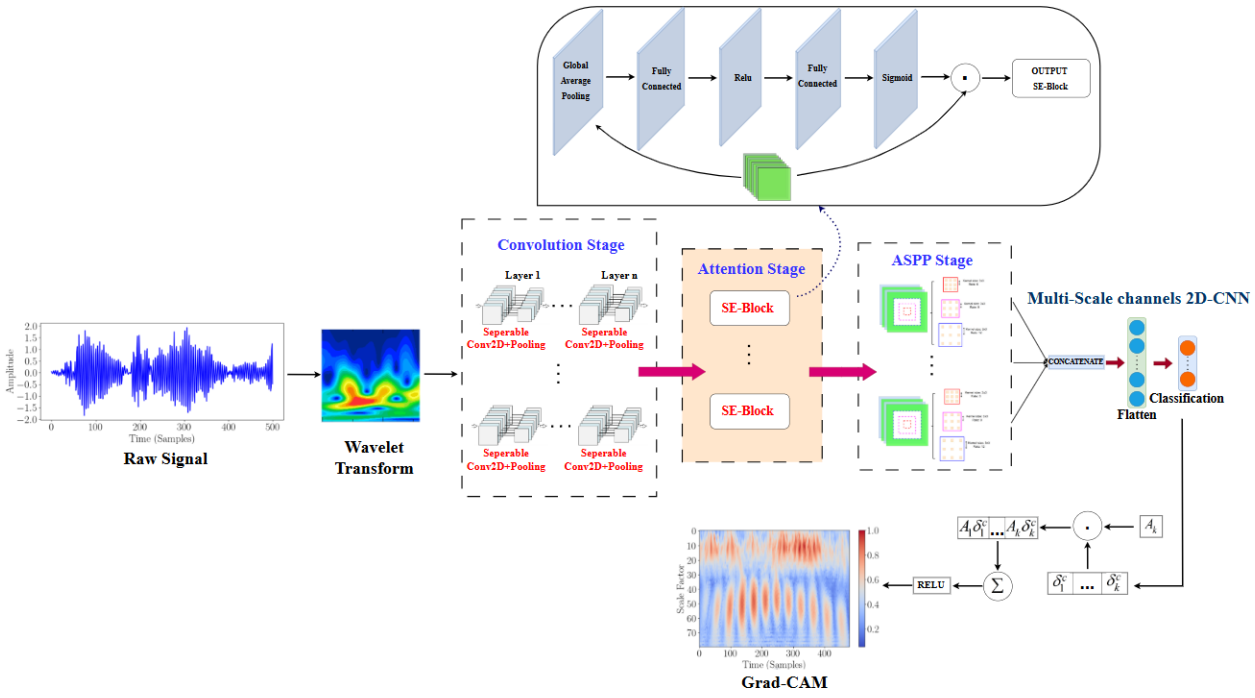


Figure 8. Design of the Lightweight Hybrid-CNN

These parallel branches efficiently extract rich features across different scales. In next step, SE block is employed to adaptively recalibrate channel-wise feature responses. By performing global average pooling followed by fully connected transformations and a sigmoid activation, the SE block emphasizes informative channels while suppressing less relevant ones, thereby enhancing discriminative feature representation. Meanwhile, ASPP module captures multi-scale contextual information by applying parallel atrous convolutions with different dilation rates. This design enlarges the receptive field without increasing computational cost, allowing the network to effectively integrate both local and global features. The joint utilization of SE and ASPP enables the model to focus on salient features while simultaneously exploiting multi-scale context, leading to improved accuracy and robustness in bearing fault diagnosis.

The extracted feature maps are then flattened and concatenated before being passed to fully connected layers for classification. Grad-CAM mechanism in Figure 8 helps visualize which parts of the input image (spectral image in this case) the model is focusing on when making a prediction. The Grad-CAM output, as shown in the lower-left part of the image, can be used to visually inspect which regions of the spectral image are most responsible for the prediction of the model. Table 3 illustrates the detail of the Lightweight Hybrid-CNN structure in this study.

Ordinal	*Layer Type*	*Kernel size/others*
1	input_layer_1	(100, 500, 1)
2	input_layer_3	(100, 500, 1)
4	conv2d_1	(5, 5)
5	conv2d_2	(11, 11)
7	dropout_1	0.2
8	dropout_2	0.2
10	DS conv_last1	(3, 3)
11	DS conv_last2	(7, 7)

13	max_pooling2d_1	(3, 3)
14	max_pooling2d_2	(3, 3)
15	SE Block_1	Reduction: 8
16	SE Block_2	Reduction: 8
17	ASPP Layer_1	Dilation rate: 1, 10
18	ASPP Layer_2	Dilation rate: 1, 10
19	flatten_1	/
20	flatten_2	/
21	concatenate	/
22	dense	/
23	dense_1	/

Table 3. Detail of the proposed structure

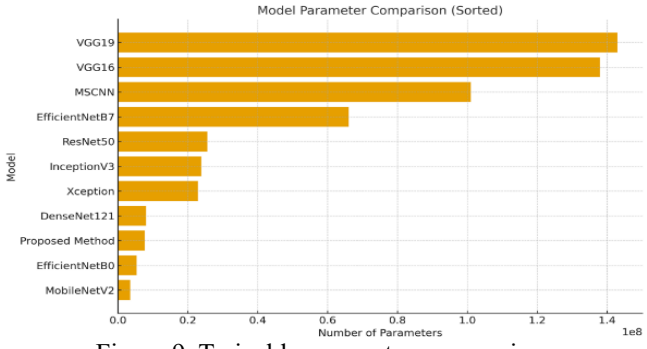


Figure 9. Trainable parameters comparison

In contrast to ResNet-based architectures, such as SE-ResNet152, which contain more than 60 million parameters, the proposed Lightweight Hybrid-CNN, with approximately 7.69 million parameters, achieves a more balanced compromise between computational efficiency and predictive performance. Additionally, the model's architecture avoids the use of heavy attention modules or transformer blocks, which further reduces memory and compute requirements. The Grad-CAM visualization step is computed post-hoc and only when interpretability is needed, thereby not affecting the model's online inference speed. Figure 9 presents a comparison of the model's size between the proposed method and several commonly used pretrained deep learning architectures. Notably, the parameter count of the proposed Lightweight Hybrid-CNN is comparable to compact networks while capture effectively fault-related features across different receptive fields and improve generalization without substantially increasing model complexity.

3.2. Faults Diagnostic Procedure Based on the Proposed Method

The complete fault diagnosis procedure of the proposed framework is organized into four sequential stages, as illustrated in Figure 10. In the first stage, raw vibration signals acquired from the bearing housing are divided into fixed-length segments and amplitude-normalized to suppress operating-condition variability. In the second stage, each segment is mapped to a two-dimensional time–frequency representation through the continuous wavelet transform, yielding spectral images that jointly preserve transient and stationary fault signatures. In the third stage, these images are propagated through the trained DS-MSCNN, whose parallel multi-scale branches extract complementary features across distinct receptive fields and classify the sample into one of the predefined health states. In the final stage, Grad-CAM produces class-discriminative activation maps that localize the time–frequency regions responsible for each prediction.

As shown in Figure 10, the distinguishing characteristic of the proposed framework is its explainability. Rather than treating the network as a black box, Grad-CAM superimposes attention maps onto the wavelet-transformed inputs, revealing the discriminative regions that drive each classification decision. Coupling accurate CNN-based feature extraction with this transparent reasoning process enables the diagnostic outcome to be cross-validated against established quantitative metrics—training loss, classification accuracy, and the confusion matrix—while simultaneously providing physical insight into the learned representations.

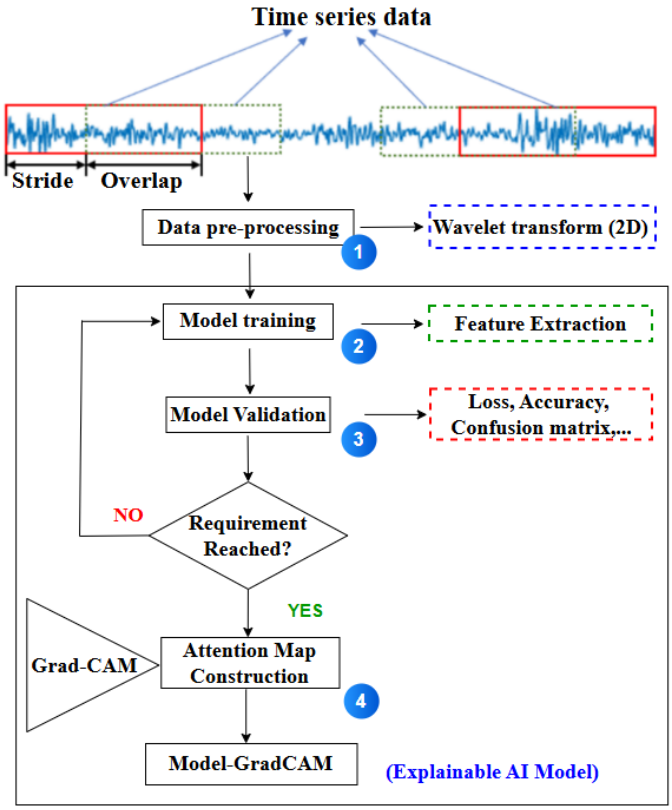


Figure 10. Design of the Lightweight Hybrid-CNN diagnostic procedure

4. EXPERIMENTAL RESULTS

4.1. Overview of the CWRU Bearing Dataset

The Case Western Reserve University (CWRU) Bearing Dataset is widely recognized as a standard benchmark in research on rotating machinery fault diagnosis. It is frequently used to evaluate machine learning and signal processing approaches for bearing fault detection and classification. The dataset was collected under various operating conditions and fault types. Due to its diversity and reliability, it has become an important dataset for both academic research and industrial investigation. The data was collected using a specialized experimental bearing (Figure 11) and test rig (Figure 12) designed to simulate real-world conditions in rotating machinery.

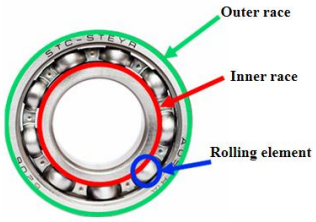


Figure 11. Rolling bearing test platform

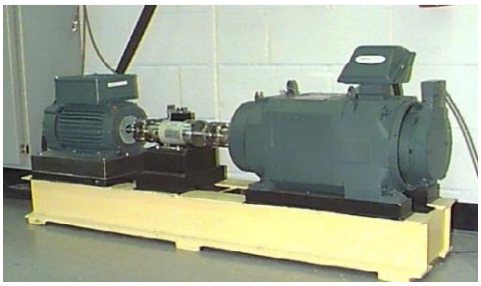


Figure 12. Rolling bearing test platform

4.2. Data Acquisition and Sampling Rates

The CWRU Bearing Dataset encompasses a range of bearing conditions, including normal (healthy) states and various fault conditions. The faults were artificially induced using electro-discharge machining (EDM), a precise method that ensures consistent fault creation. The dataset includes:

Inner Race Faults: Damage located on the inner race of the bearing.

Outer Race Faults: Defects present on the outer race.

Ball Faults: Imperfections or damage on the rolling elements themselves.

Each fault type was introduced at three different severity levels, determined by the fault diameter: 0.007 inches, 0.014 inches, and 0.021 inches. This gradation in severity allows for the assessment of diagnostic algorithms across a spectrum of fault intensities.

Vibration data in the CWRU dataset was recorded at sampling rates 12 kHz. This high-resolution data capture is crucial for accurately detecting and analyzing the often subtle and complex vibration patterns associated with bearing faults. The dataset includes recordings taken under varying motor loads (0 to 3 horsepower) and at different motor speeds, simulating a wide range of real-world operational conditions.

Accelerometers were strategically placed to capture the vibration signals in both the horizontal and vertical directions relative to the motor shaft. This multi-axis data collection provides a more comprehensive view of the operational health of the bearing, capturing nuances that might be missed with single-axis measurements.

4.3. Computational Environment and Training Efficiency

Experiments in this study were deployed to Google Colab Pro with the NVIDIA L4 (16GB VRAM). The training framework was implemented using Python, the TensorFlow library, and CUDA acceleration. The entire training duration was approximately 0.9 to 1.2 hours to achieve convergence, which is reasonable compared to similar deep learning methods reported in the literature.

To ensure reproducibility and consistent convergence behavior, all training settings and hyperparameters were strictly controlled and are summarized in Table 4. The proposed Lightweight Hybrid-CNN was implemented in TensorFlow with CUDA acceleration and trained using the settings below.

Hyperparameter	Value
Framework	TensorFlow + CUDA
Optimizer	Adam
Learning rate	1×10^{-3}
Loss function	Categorical cross-entropy
Batch size	64
Training epochs	10
Dropout rate	0.2
SE reduction ratio	8
ASPP dilation rates	1, 10
Branch 1 kernel sizes	5×5 (Conv) → 3×3 (DS-Conv)
Branch 2 kernel sizes	11×11 (Conv) → 7×7 (DS-Conv)
Max pooling size	3×3
Data split	70% train / 15% validation / 15% test
Hardware	NVIDIA L4 (16GB VRAM)

Table 4. Training hyperparameters and experimental configuration

It is noted that while The three-way train/validation/test split adopted on the CWRU dataset is a methodological safeguard against overfitting and a basis for model selection; it does not, by itself, constitute evidence of generalization to unseen operating conditions. Evidence of generalization is instead provided by the cross-domain evaluation on the entirely independent HUST dataset (Section 5), collected under different hardware and operating conditions, on which the proposed model achieves over 99% accuracy—including on the compound inner–outer fault class rarely addressed in prior work.

4.4. Bearing Fault Diagnostic Results

The experimental results on the CWRU dataset demonstrate the effectiveness of the proposed architecture for bearing fault diagnosis. As illustrated in Figure 13, the Lightweight Hybrid-CNN rapidly achieves accuracy above 99% within the first few training epochs and maintains stable performance during the remaining training process.

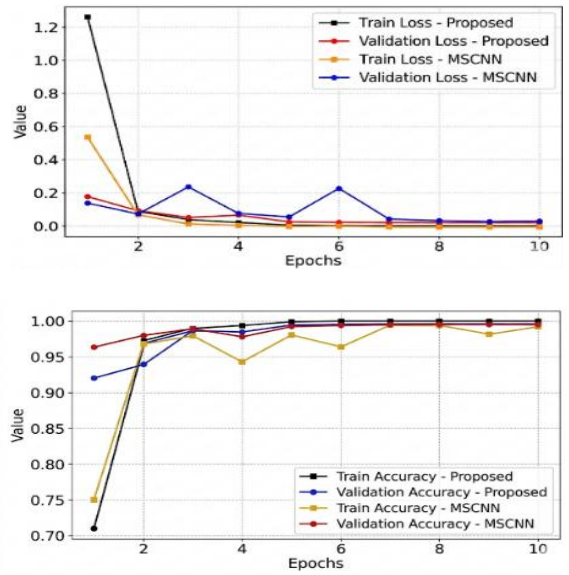


Figure 13. Training convergence comparison: (above) Proposed Lightweight Hybrid-CNN; (bottom) Baseline MSCNN

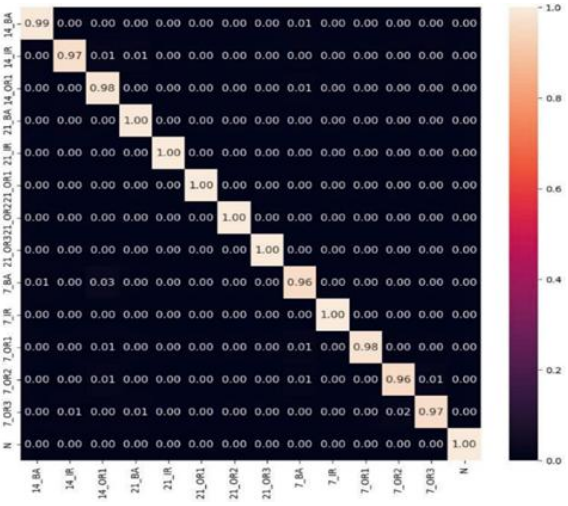


Figure 14. Classification confusion matrix

Figure 14 illustrates the normalized confusion matrix of the proposed method, while Figure 15 corresponds to the baseline MSCNN.

In contrast, the MSCNN model shows slightly lower performance, as several diagonal values fall below 1.00 (e.g., 0.97, 0.98), and there are non-zero off-diagonal entries, particularly in fault categories with similar characteristics. This suggests the presence of minor misclassifications. Such degradation in class-wise accuracy may be attributed to the higher model complexity and the lack of efficient regularization in MSCNN, making it more prone to overfitting or confusion between closely related fault classes.

Overall, the comparison reinforces the effectiveness of the depthwise separable architecture in the Lightweight Hybrid-CNN, not only in reducing parameters but also in enhancing class separability and robustness in fault diagnosis tasks.

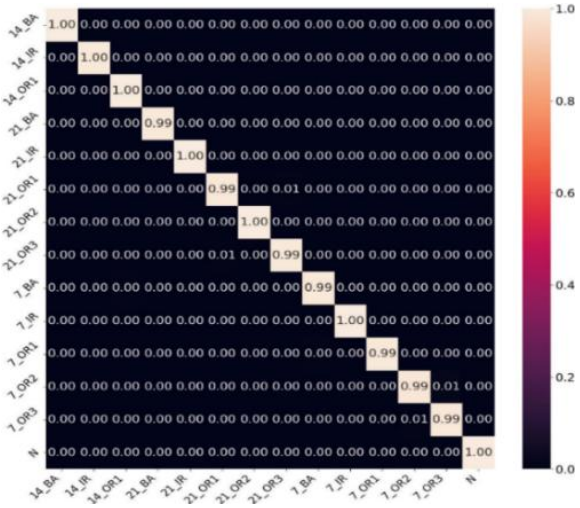


Figure 15. Confusion matrix of the MSCNN

Method	Accuracy (%)
1D-CNN (Zhang et al., 2018)	99.5
MSCNN	98.4
GC-CNFN (Liao, Wu, & Wu, 2022)	99.8
DKDL-Net (Petrosian et al., 2024)	99.4
DC-ResNet (Yu, Li, & Liu, 2023)	99.7
MQCNN+BiGRU+CSAFF (Liu et al., 2024)	99.7
Lite CNN (Yao, Zhou, & Jiang, 2023)	**99.8**
FaultNet (Khan, Asad, Kudelina, Vaimann, & Kallaste, 2022)	99.3
Proposed	**99.6**

Table 5. Comparative evaluation on CWRU dataset

As shown in Table 5, the proposed Lightweight Hybrid-CNN achieves a highly competitive accuracy of 99.6% on the CWRU dataset, outperforming several representative baselines including 1D-CNN (99.5%), MSCNN (98.4%), DKDL-Net (99.4%), and FaultNet (99.3%). Although GC-CNFN and Lite CNN report marginally higher accuracy (both 99.8%), and DC-ResNet and MQCNN+BiGRU+CSAFF reach 99.7%, these architectures generally rely on deeper

backbones, additional recurrent components, or knowledge-distillation pipelines that substantially increase model complexity. In contrast, the proposed model attains comparable accuracy with a markedly more compact architecture, demonstrating a clear efficiency advantage. Overall, the proposed Lightweight Hybrid-CNN offers a favorable trade-off between diagnostic accuracy and model compactness — characteristics that are particularly critical for PHM applications on resource-constrained embedded platforms.

In contrast, the proposed Lightweight Hybrid-CNN maintains a compact multi-scale convolutional architecture while incorporating Grad-CAM for model interpretability. This design provides a favorable balance between diagnostic performance, computational efficiency, and explainability. Furthermore, compared with conventional 1D-CNN approaches (Zhang et al., 2018) and Fault-Net (Khan et al., 2022), the proposed framework demonstrates clear improvements, highlighting the advantage of spectral-domain feature extraction combined with spatial attention-based visualization.

4.5. Ablation Study

To investigate the impact of each component in the proposed architecture, four model variants were evaluated: (1) DSCNN - SE - ASPP (the proposed Lightweight Hybrid-CNN model), (2) CNN - SE - ASPP, (3) DSCNN- ASPP, and (4) DSCNN-SE.

The box-plot results in Figure 16 highlight the strong performance of the proposed DSCNN-SE-ASPP model, which achieves the highest mean accuracy (~0.996) with minimal variance, indicating stable and reliable predictions. In comparison, CNN-SE-ASPP attains a slightly lower mean accuracy (~0.984), suggesting that depthwise separable convolution can reduce model parameters while preserving competitive accuracy.

When the SE module is removed, DSCNN-ASPP shows reduced accuracy (~0.972) and increased dispersion, indicating the importance of channel-wise feature selection. Similarly, DSCNN-SE yields the lowest mean accuracy (~0.97) and the highest variance, emphasizing the role of ASPP in capturing multi-scale frequency information and improving robustness under varying operating conditions.

Overall, these results demonstrate that each module contributes to performance improvement: SE enhances feature selectivity, ASPP strengthens frequency-band sensitivity, and DSCNN reduces computational complexity. Their integration provides an effective balance between accuracy, generalization, and efficiency for reliable and interpretable bearing fault diagnosis.

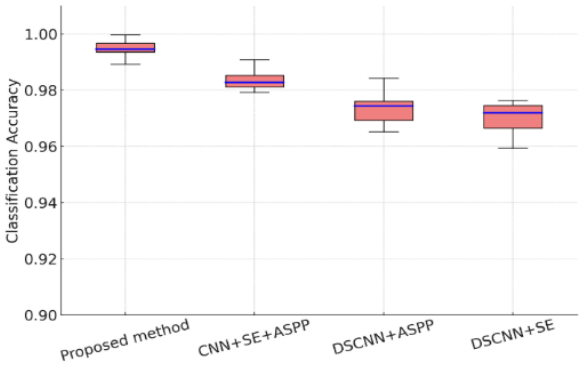


Figure 16. Box-Plot Analysis of Ablation Study

4.6. Interpretability for Diagnostic Method

Figure 17 to Figure 20 below illustrate the Grad-CAM attention mechanism and how the model's decisions are evaluated. In normal conditions, the raw vibration signal (Figure 17 – top left) shows relatively consistent patterns. The normal case wavelet transform (Figure 17 – top right) exhibits a balanced distribution of frequencies over time, without any significant spikes or abnormalities. This indicates a smooth operating condition without defects. The frequency components are relatively uniform, and there is no high-energy concentration at specific time points, no particular areas dominate the time-frequency representation, and the Lightweight Hybrid-CNN can easily classify this as a healthy condition, as confirmed by minimal attention in the Grad-CAM heat maps (*Figure 17- bottom*).

To ensure that the attention maps are physically meaningful rather than artifacts of the visualization procedure, the Grad-CAM responses are extracted from the final ASPP feature block, where the receptive fields of the parallel dilated branches already span the entire 500 Hz–10 kHz analysis band. The gradient of the predicted class score with respect to each feature map is globally averaged to obtain channel importance weights, which are then combined with the corresponding activation maps and rectified so that only positive contributions are retained. The resulting one-dimensional saliency profile is resampled to the original signal length and superimposed on both the time-domain waveform and its continuous wavelet transform, so that every highlighted region can be traced back to a specific time instant and frequency band. This mapping allows the diagnostic evidence used by the network to be compared directly with the theoretical fault-characteristic frequencies of the tested bearing.

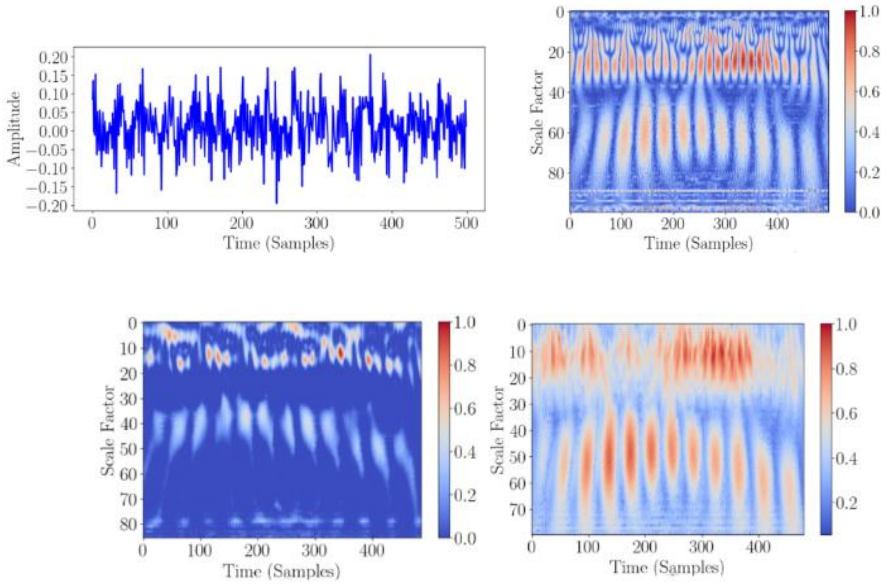


Figure 17. Grad-CAM results for normal case

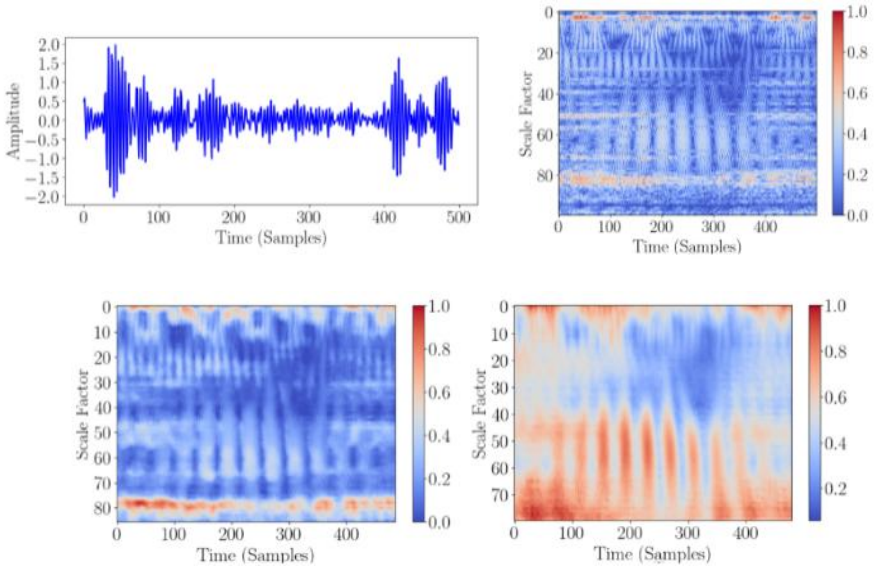


Figure 18. Grad-CAM results for ball element fault case

For the ball element fault case (Figure 18), the raw vibration signal exhibits irregular patterns, and the wavelet transformation has noticeable energy concentration at specific time intervals, which correspond to the occurrences of ball element faults. The frequency characteristics of the signal change dramatically, with more pronounced energy at specific scales, indicating localized damage. The changes in time-frequency representation allow the proposed model to focus on these areas. The Grad-CAM responses for the ball element fault (Figure 18) reveal distributed activations across both high-frequency (approximately 5–10 kHz) and low-to-mid frequency regions (approximately 500 Hz–2 kHz). The high-frequency emphasis corresponds to short-duration impulsive excitations generated when the defective rolling element repeatedly contacts the inner and outer raceways, thereby exciting structural resonance modes of the bearing–housing system. Simultaneously, the model highlights periodic components in the lower frequency band associated with the ball spin frequency (BSF) and its modulation effects. This dual-band activation reflects the rolling–sliding contact mechanism characteristic of ball defects, which inherently produces both impulsive resonance excitation and modulated periodic vibration patterns. The Grad-CAM distribution confirms that the proposed multi-scale CNN captures this composite resonance behavior within the validated 500 Hz–10 kHz band, rather than relying on isolated statistical features. Similar results are also gained for the inner race and outer race fault cases (Figure 19 and Figure 20)

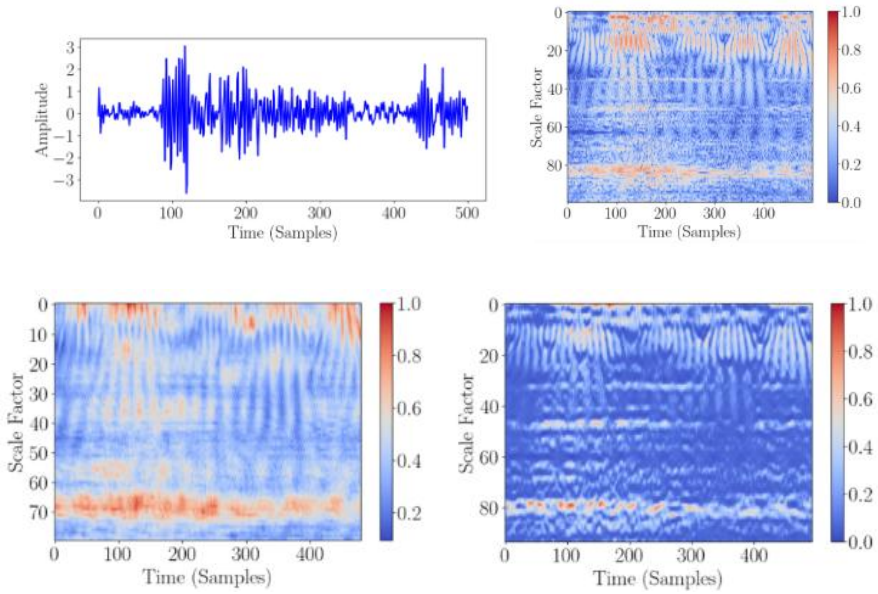


Figure 19. Grad-CAM results for the inner race fault case

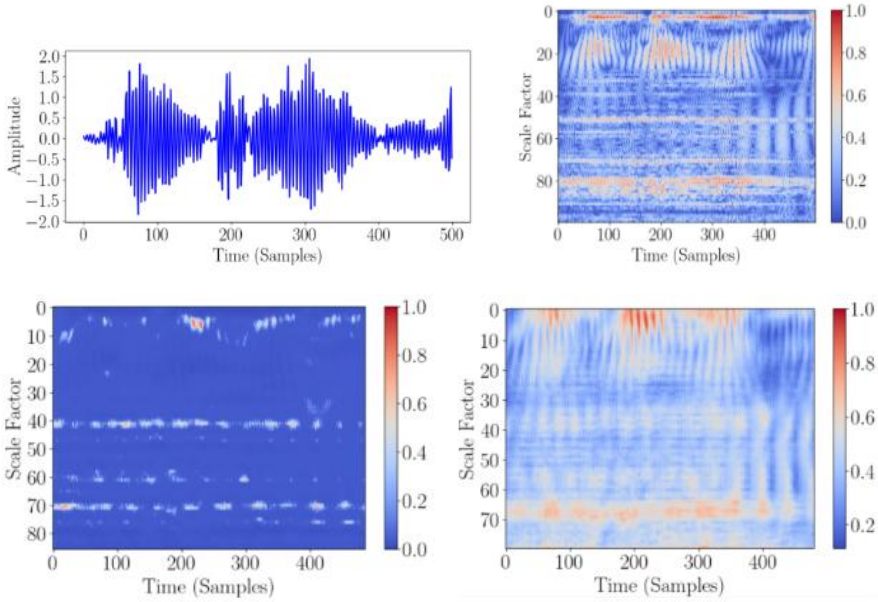


Figure 20. Grad-CAM results for outer race fault case

For the inner race fault case (Figure 19), the Grad-CAM maps predominantly emphasize the low-to-mid frequency region (approximately 500 Hz–2 kHz), while showing comparatively weaker activation in the higher resonance band. This behavior is physically consistent with the rotating nature of inner race defects. Because the defect travels relative to the fixed sensor location, the resulting impacts are strongly modulated by shaft rotation and tend to distribute energy over harmonics of the ball pass frequency of the inner race (BPFI), rather than producing sustained high-frequency resonance bursts. The concentration of attention in the larger-scale (lower-frequency) region indicates that the network relies primarily on periodic fault-related components rather than transient high-frequency impulses. This selective activation demonstrates that the model differentiates inner race faults through their characteristic modulated resonance signatures within the 500 Hz–10 kHz analysis range. Similarly, for the outer race fault (Figure 20), Grad-CAM reveals strong and stable activations concentrated in the 500 Hz–2 kHz band, with additional localized responses in the higher resonance region. Unlike inner race defects, an outer race defect remains spatially fixed relative to the sensor. Each rolling element impact occurs at nearly the same angular position, generating a quasi-periodic excitation source. This

spatial consistency produces repetitive impact forces with nearly constant transfer paths, leading to stronger and more coherent excitation of structural modes. As a result, the vibration spectrum exhibits clearer concentration around the ball pass frequency of the outer race (BPFO) and its harmonics. The persistent Grad-CAM activation in the low-to-mid frequency band reflects this structurally constrained excitation mechanism. Localized high-frequency activation further indicates resonance amplification caused by repeated impacts. Compared to the inner race case, the outer race defect produces more stable resonance-driven responses, which the CNN captures through concentrated attention patterns.

The Grad-CAM analyses in Figs. 18–20 further clarify the necessity of the multi-kernel design. Small kernels predominantly capture localized high-frequency resonance bursts generated by short-duration impact forces, whereas large kernels emphasize spatially extended periodic patterns associated with characteristic fault frequencies. Importantly, the relative contribution of each kernel branch varies depending on the fault type: ball defects activate both branches due to rolling–sliding contact dynamics, inner race faults rely primarily on large-kernel periodic modulation patterns, and outer race faults exhibit stable low-frequency resonance responses with localized high-frequency excitation. This fault-dependent activation behavior demonstrates that the multi-scale architecture is not redundant but physically justified, as it enables the separation and integration of distinct vibration mechanisms within the validated 500 Hz–10 kHz band.

Therefore, the proposed multi-kernel configuration is structurally motivated by bearing vibration physics rather than empirical tuning, ensuring that impulsive resonance phenomena and periodic modulation mechanisms are simultaneously represented in the learned feature space.

4.7. Evaluation of Diagnosis Mechanism in a Noisy Environment

In numerous investigations, Additive White Gaussian Noise (AWGN) is intentionally injected into vibration measurements to replicate the random disturbances typically present in real bearing operation. This type of noise exhibits a power spectral density whose magnitude conforms to a Gaussian probability distribution across frequencies, making it an effective representation of broadband sensor and environmental interference. In the present study, the Signal-to-Noise Ratio (SNR) is adopted as the primary indicator for quantifying the level of noise superimposed on the vibration signals, as illustrated in Figure 23. The SNR reflects the ratio between the energy of the informative vibration component and the additive noise component, and its value is determined according to the following formulation:

$$SNR = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{23}$$

where P_{signal} and P_{noise} denote the average power of the original vibration signal and of the injected Gaussian noise, respectively. A high SNR therefore indicates that the fault-related impulses remain clearly distinguishable, whereas a low SNR corresponds to a severely contaminated measurement in which the characteristic impact components are largely masked by broadband disturbance. For each target SNR level, the variance of the generated noise sequence is scaled from the measured power of the clean signal so that the prescribed ratio is obtained exactly, and the noise is superimposed sample by sample on the raw time-domain waveform before any normalization or segmentation is applied. This procedure guarantees that the degradation affects the same spectral content that the multi-scale convolutional branches are expected to exploit.

Four noise levels, namely 0, 5, 10, and 15 dB, are considered in order to span conditions ranging from extremely harsh to moderately noisy industrial environments. In all cases the network is trained exclusively on the clean training subset and evaluated on the noise-corrupted test subset, so that the reported results reflect genuine robustness rather than adaptation to a particular noise realization. Each experiment is repeated over several independent trials with different random noise seeds, and the averaged accuracies are reported to eliminate the influence of individual realizations. The visual effect of this contamination on the raw vibration waveform is compared in Figure 21, while the corresponding quantitative results are summarized in Table 6.

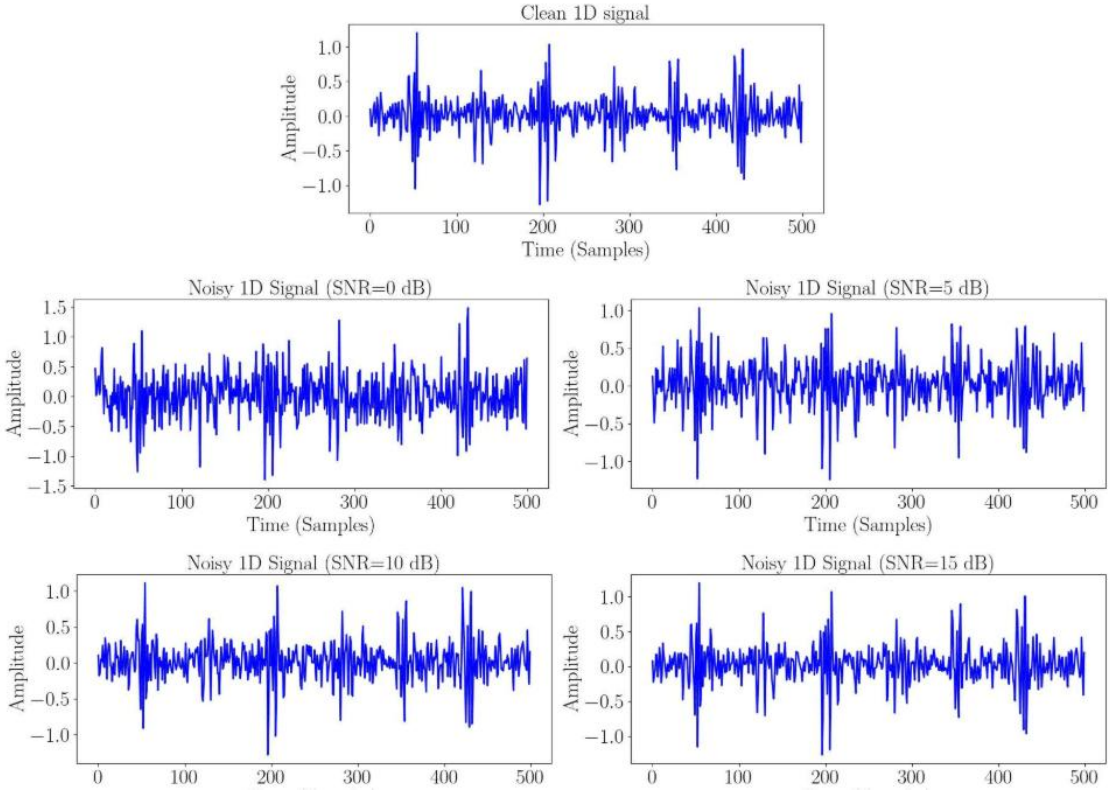


Figure 21. Comparison of Additive White Gaussian Noise in raw data.

SNR (dB)	MSCNN (%)	Proposed Method (%)	Improvement (%)
0	~85.5	~87.6	+2.1
5	~89.7	~91	+1.3
10	~92.8	~94.7	+1.9
15	~96.4	~98.2	+1.8

Table 6. Classification accuracy (%) under different SNR levels

Table 6 summarizes the average classification accuracy of the proposed model and the MSCNN baseline under four levels of additive white Gaussian noise, providing a quantitative basis for robustness comparison.

As shown in Table 6, the proposed method consistently outperforms MSCNN across all noise conditions, with accuracy gains ranging from +1.3% at SNR = 5 dB to +2.1% at SNR = 0 dB. The performance advantage is most pronounced at SNR = 0 dB (87.6% vs. 85.5%), where the resonance-focused CWT preprocessing and SE-based channel recalibration effectively preserve fault-related spectral components even under severe noise contamination. At higher SNR levels (10 dB and 15 dB), the proposed method reaches 94.7% and 98.2% respectively, compared with 92.8% and 96.4% for MSCNN. Detailed confusion matrices for each SNR level are provided in Figure 22.

The confusion matrices in Figure 24 illustrate how the proposed framework behaves under four different noise intensities (0 dB, 5 dB, 10 dB, and 15 dB). Overall, the model maintains a high level of robustness against injected noise, yielding average accuracies ranging from roughly 87.6% to 98.2%. Even in the most challenging scenario, SNR = 0 dB—where the noise power is equivalent to the signal power—the classifier retains reasonable predictive capability, achieving 87.6% accuracy. Some misclassification appears in fault categories with closely overlapping spectral characteristics, which is expected under such heavily degraded conditions. As the SNR improves to 5 dB and 10 dB, the diagnostic accuracy rises sharply, surpassing the 95% threshold. This

trend indicates that the multi-scale depth-wise separable convolutions in conjunction with the SE/ASPP module are highly effective at capturing discriminative spectral cues despite substantial interference. At SNR = 15 dB, the model attains near-perfect recognition across all fault types, reaching approximately 98% accuracy.

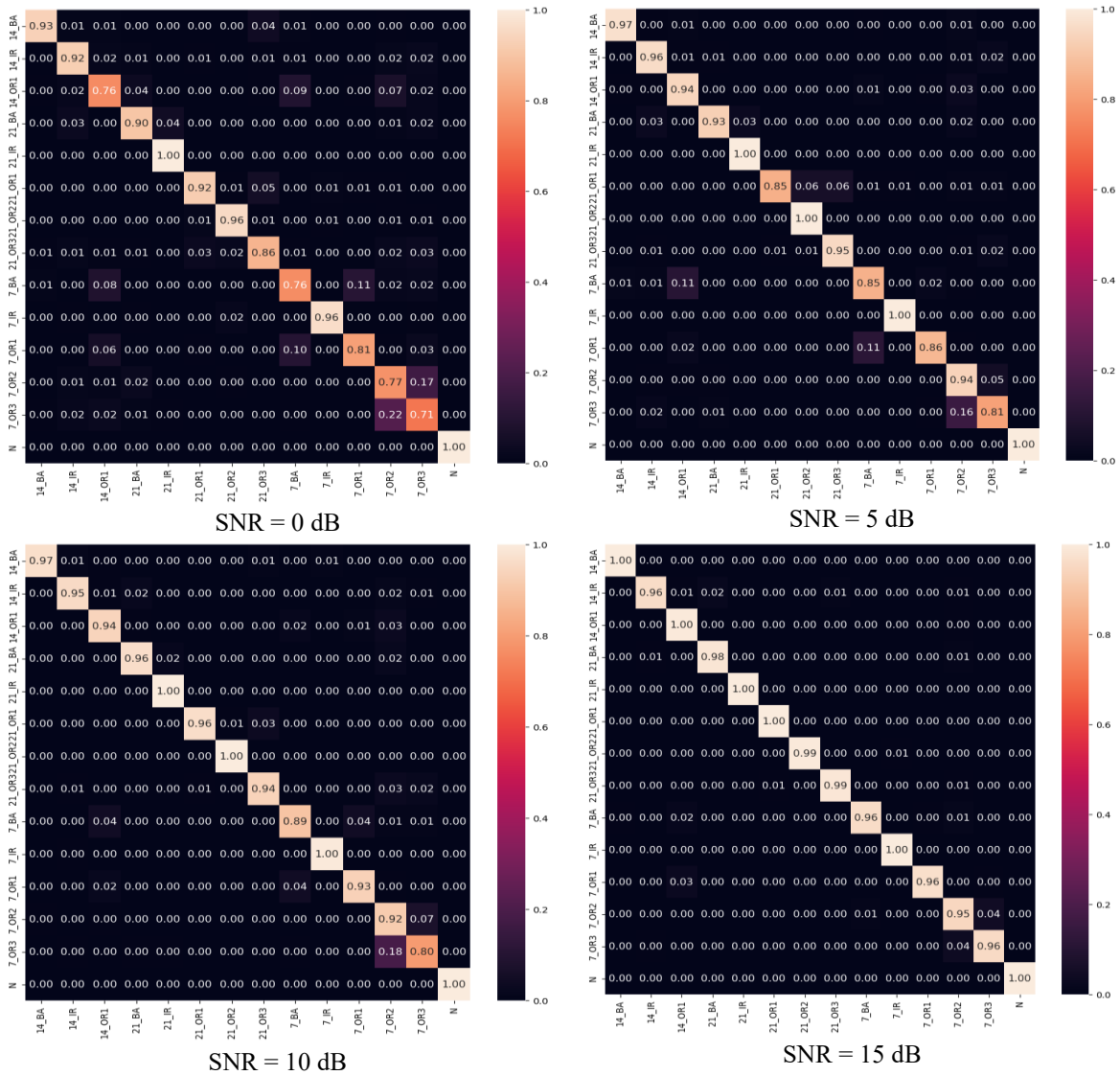


Figure 22. The confusion matrices of the proposed model with different SNR levels

5. CROSS-DOMAIN DEEP TRANSFER DIAGNOSIS

5.1. Overview of the HUST Dataset

The HUST dataset, developed by Nguyen and Hoang (2023), serves as a representative and application-oriented benchmark for evaluating ball bearing fault-diagnosis methods. In this dataset, faults were intentionally introduced by employing a wire-cutting process to create micro-notches approximately 0.2 mm in width (Figure 23), effectively mimicking the early stages of defect formation. Vibration measurements were then acquired from an electric motor running under light-load conditions, using a high sampling rate of 51,200 Hz to ensure adequate capture of dynamic responses.



Figure 23. Fault location of HUST Dataset (a-inner race fault, b-outer race fault, c-ball fault, and d-combined inner-outer fault)

An important characteristic of this dataset lies in its reflection of the coupled mechanical behavior among different bearing faults. Damage occurring in one structural element frequently influences other components, leading to compound fault patterns. To represent these interactions, the dataset provides 90 raw vibration sequences distributed across six diagnostic categories, including single defects—inner race, outer race, and ball faults—as well as their pairwise combinations. These signals were collected from five individual bearings under three distinct operating conditions, offering a diverse and realistic testbed for assessing robust fault-classification algorithms as shown in Table 7.

Fault label	Fault category	Sample size
0	Normal	500
1	Inner race fault	500
2	Outer race fault	500
3	Ball fault	500
4	Inner and outer fault	500

Table 7. Details of the HUST dataset

In practical industrial environments, the occurrence of concurrent defects affecting both the raceway and the rolling elements—such as inner–ball or outer–ball crack combinations—is relatively rare. These conditions typically emerge only after extended periods of severe deterioration, making them challenging to reproduce consistently in laboratory settings. As a result, the available data for such fault types are insufficient for conducting a rigorous and statistically meaningful assessment. For this reason, the present study concentrates on the fault categories that appear more frequently and are more representative of real operating conditions, namely single defects (inner race crack – I, outer race crack – O, and ball crack – B) and the combined inner–outer race defect (IO). The exclusion of inner–ball and outer–ball combinations ensures that the evaluation remains both practical and methodologically sound. A summary of the dataset composition is provided in Table 7.

5.2. Bearing Fault Diagnosis and Evaluation

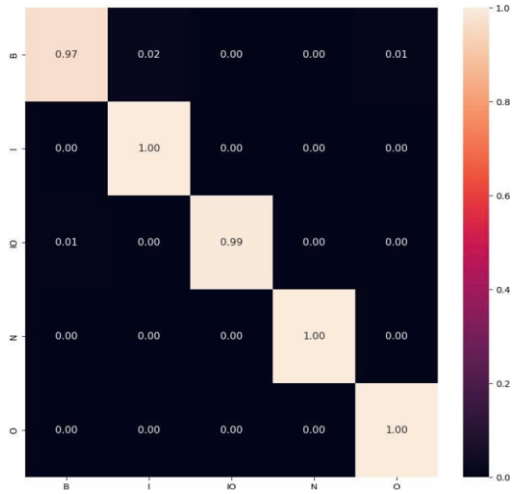


Figure 24. Confusion matrix of proposed method

The classification results in Figure 24 present a strong diagonal dominance, indicating highly discriminative features across all fault types. The per-class accuracies achieved along the diagonal are as follows: B = 0.97, I = 1.00, IO = 0.99, N = 1.00, and O = 1.00, resulting in a macro-average accuracy of approximately 0.992 (99.2%). Of particular interest is the performance on the compound fault class (IO), which reaches 0.99. This high score indicates that the network effectively recognizes the combined inner–outer race defect patterns instead of misidentifying them as individual single-fault categories.

These findings highlight the effectiveness of SE-block in refining channel feature emphasis and the contribution of ASPP in maintaining the integrity of multi-scale spectral structures. Together, these modules enable the network to achieve near-complete class separability, even for the IO class—a fault type that many existing approaches either exclude or misclassify. Overall, the confusion matrix demonstrates excellent precision and recall across all categories, negligible influence of class imbalance, and strong robustness with respect to the operating variations present in the HUST dataset.

To provide a comparative reference, the MSCNN baseline was also evaluated on the HUST dataset under identical

training conditions. Table 8 presents the per-class and overall accuracy of both models.

Method	Ball (%)	Inner (%)	IO (%)	Normal (%)	Outer (%)	Overall (%)
MSCNN	89	98	97	99	99	98
Proposed	97	100	99	100	100	99

Table 8. Per-class accuracy comparison on HUST dataset

The proposed method demonstrates consistent superiority across all fault categories. Notably, on the compound fault class (IO) — which represents simultaneous inner and outer race damage — the proposed model achieves 99% accuracy compared to 98% for MSCNN, highlighting the benefit of SE-based channel recalibration and ASPP multi-scale contextual learning in separating overlapping fault signatures.

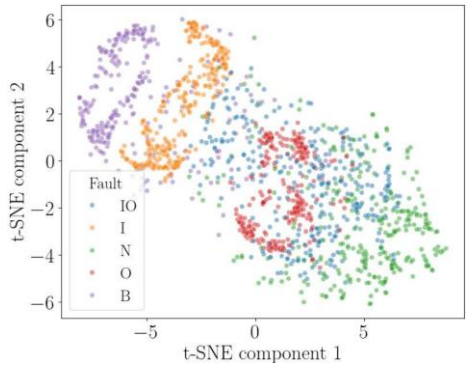


Figure 25. t-SNE visualization of the raw HUST dataset

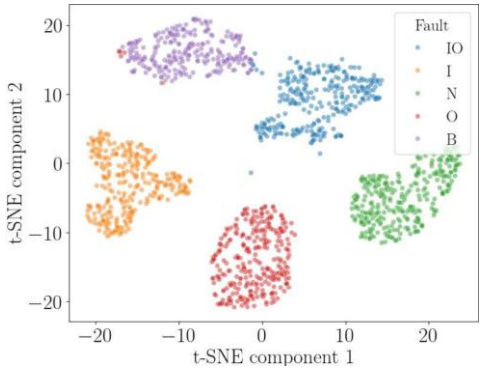


Figure 26 . t-SNE visualization of the raw dataset following proposed method

Figure 25 illustrates the t-SNE projection of the original HUST vibration signals. The visualization shows extensive mixing among the fault categories, with no clear boundaries separating them. Notably, the IO compound-fault samples occupy regions that heavily overlap with the single-fault groups (I, O, and B), demonstrating that the raw signal domain lacks sufficient discriminative structure to support reliable classification. In contrast, the t-SNE plot in Figure 26 displays the feature distribution obtained from the final fully connected layer of the proposed Efficient Lightweight Multi-Scale CNN enhanced with an improved Attention Mechanism. Here, each fault class forms a compact and well-defined cluster, and even the IO category—typically difficult to isolate—appears distinctly separated from the others. The pronounced inter-class margins indicate that the model effectively learns robust, fault-specific, and domain-invariant representations. These observations validate the strong discriminative capability of the proposed framework. By converting heavily overlapping raw data into clearly partitioned feature clusters, the method provides a direct explanation for the near-perfect classification scores achieved on the HUST dataset. Moreover, the improved separability highlights the model's ability to handle both simple faults and more complex compound defect patterns with high reliability.

6. CONCLUSION

This work presents an engineering contribution at the system integration level, demonstrating that physically motivated design choices, when systematically combined, yield a diagnostic architecture that is simultaneously lightweight, accurate, and interpretable — qualities rarely optimized together in existing approaches. In terms of model complexity, the proposed Lightweight Hybrid-CNN requires only approximately 7.69 million parameters, compared with more than 60 million for ResNet-based baselines such as SE-ResNet152—an almost 8× reduction in parameter count. The proposed Lightweight Hybrid-CNN focuses on the resonance-dominant frequency range (500 Hz–10 kHz), where bearing defects generate periodic impact forces that excite structural resonance modes. CWT with Morlet wavelet effectively captures this resonance–modulation structure in the time–frequency domain, while multi-kernel depth-wise separable convolution simultaneously characterizes impulsive resonance bursts and periodic modulation patterns across ball, inner race, and outer race fault types. The integration of SE and ASPP modules further enhances channel-wise recalibration and multi-scale contextual learning, and Grad-CAM visualization confirms that the network consistently attends to physically meaningful resonance regions beyond black-box classification. Experimental results demonstrate classification accuracy exceeding 99% on the CWRU dataset and 99.2% on the

HUST dataset, confirming strong discriminative capability, noise robustness, and cross-domain generalization. Future work will focus on hardware deployment and extended cross-domain validation for real-time predictive maintenance applications.

ACKNOWLEDGMENTS

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

REFERENCES

Aziz, S., Khan, M. U., Usman, A., Faraz, M., Ghadi, Y. Y., & Montes, G. A. (2025). Bearing faults classification using novel log energy-based empirical mode decomposition and machine Mel-frequency cepstral coefficients. *Digital Signal Processing*, 156, 104776.

Cerrada, M., et al. (2016). Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mechanical Systems and Signal Processing*, 70–71, 87–103.

Chen, Z., Cen, J., & Xiong, J. (2020). Rolling bearing fault diagnosis using time-frequency analysis and deep transfer convolutional neural network. *IEEE Access*, 8, 150248–150261.

Chen, Z., et al. (2022). A transformer-based fault diagnosis approach for rotating machinery. *IEEE Transactions on Industrial Informatics*, 18(6), 4086–4096.

Deveci, B. U., et al. (2024). Transfer learning enabled bearing fault detection methods based on image representations of single-dimensional signals. *Information Systems Frontiers*, 26, 1345–1397.

Eren, L., Ince, T., & Kiranyaz, S. (2019). A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier. *Journal of Signal Processing Systems*, 91(2), 179–189.

Gu, X., Tian, Y., Li, C., Wei, Y., & Li, D. (2024). Improved SE-ResNet acoustic–vibration fusion for rolling bearing composite fault diagnosis. *Applied Sciences*, 14(5), 2182.

Huang, K., Zhu, L., Ren, Z., Lin, T., Zeng, L., Wan, J., & Zhu, Y. (2024). An improved fault diagnosis method for rolling bearings based on 1D_CNN considering noise and working condition interference. *Machines*, 12(6), 383.

Jia, F., Lei, Y., Lin, J., Zhou, X., & Lu, N. (2016). Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mechanical Systems and Signal Processing*, 72, 303–315.

Jiang, K., Yang, Z., Jin, T., Chen, C., Liu, Z., & Zhang, B. (2025). CNN-based rolling bearing fault diagnosis method with quantifiable interpretability. *IEEE Transactions on Instrumentation and Measurement*.

Khan, M. A., Asad, B., Kudelina, K., Vaimann, T., & Kallaste, A. (2022). The bearing faults detection methods for electrical machines—the state of the art. *Energies*, 16(1), 296.

Lei, Y., Lin, J., He, Z., & Zuo, M. J. (2013). A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, 35(1–2), 108–126.

Li, L., Zhao, X., Liu, X., & Fei, J. (2022, June). Rolling bearing fault diagnosis based on convolutional neural network and multi-sensor information fusion. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (Vol. 10, pp. 1396–1401). IEEE.

Liang, H., Cao, J., & Zhao, X. (2022). Multi-scale dynamic adaptive residual network for fault diagnosis. *Measurement*, 188, 110397.

Liang, Y.-P., Chen, H., & Chung, C.-C. (2024). A one-dimensional depthwise separable convolutional neural network for bearing fault diagnosis implemented on FPGA. *Sensors*, 24(23), 7831.

Liao, S., Wu, Y., & Wu, Y. (2022). Bearing fault diagnosis using a Grad-CAM-based convolutional neuro-fuzzy network. *Mathematics*, 10(13), 2075.

Liefstingh, M., Taal, C., Restrepo, S. E., & Azarfar, A. (2021, November). Interpretation of deep learning models in bearing fault diagnosis. *Annual Conference of the PHM Society*, 13(1).

Liu, H., Zhang, F., Tan, Y., Huang, L., Li, Y., Huang, G., Luo, S., & Zeng, A. (2024). Multi-scale quaternion CNN and BiGRU with cross self-attention feature fusion for fault diagnosis of bearing. *Measurement Science and Technology*, 35(8), 086138.

Nguyen, D. T., & Hoang, S. H. (2023). HUST bearing: A practical dataset for ball bearing fault diagnosis. *BMC Research Notes*, 16, 138.

Pandhare, V., Singh, J., & Lee, J. (2019, May). Convolutional neural network based rolling-element bearing fault diagnosis for naturally occurring and progressing defects using time-frequency domain features. In *2019 Prognostics and System Health Management Conference (PHM-Paris)* (pp. 320–326). IEEE.

Peng, Z. K., & Chu, F. L. (2004). Application of the wavelet transform in machine condition monitoring and fault diagnostics: A review with bibliography. *Mechanical Systems and Signal Processing*, 18(2), 199–221.

Petrosian, O., Pengyi, L., Yulong, H., Jiarui, L., Zhaoruikun, S., Guofeng, F., & Liping, M. (2024). DKDL-Net: A lightweight bearing fault detection model via decoupled knowledge distillation and low-rank adaptation fine-tuning. *IEEE Transactions on Industrial Informatics*, 20(3), 4567–4578.

Rabah, A., & Abdelhafid, K. (2018). Rolling bearing fault diagnosis based on improved complete ensemble empirical mode decomposition with adaptive noise combined with minimum entropy deconvolution. *Journal of Vibroengineering*, 20(1), 240–257.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics — A tutorial. *Mechanical Systems and Signal Processing*, 25(2), 485–520.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

Toma, R. N., et al. (2022). A bearing fault classification framework based on image encoding techniques and a convolutional neural network under different operating conditions. *Sensors*, 22(13), 4881.

Wang, B., Feng, G., Huo, D., & Kang, Y. (2022). A bearing fault diagnosis method based on spectrum map information fusion and convolutional neural network. *Processes*, 10(7), 1426.

Wang, Q., Yin, C., She, K., Tong, Q., Lu, G., Zhang, H., & Lu, J. (2025). Bearing fault diagnosis for variable operating conditions based on KAN convolution and dual branch fusion attention. *Scientific Reports*, 15(1), 21442.

Wang, X., Hua, T., Xu, S., & Zhao, X. (2023). A novel rolling bearing fault diagnosis method based on BLS and CNN with attention mechanism. *Machines*, 11(2), 279.

Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, 21(6), 2560–2574.

Wu, Y., Dai, J., Yang, X., Shao, F., Gong, J., Zhang, P., & Liu, S. (2025). The fault diagnosis of rolling bearings based on FFT-SE-TCN-SVM. *Actuators*, 14(3), 152.

Xu, H., Xiao, Y., Sun, K., & Cui, L. (2024). Improved depthwise separable convolution for transfer learning in fault diagnosis. *IEEE Sensors Journal*, 24(20), 33606–33613.

Yao, C., Zhou, L., & Jiang, W. (2023). Lite and efficient deep learning model for bearing fault diagnosis using the CWRU dataset. *Sensors*, 23(6), 3157.

Yu, B., Li, S., & Liu, J. (2023). Fault diagnosis of printing press bearing based on deformable convolutional ResNet. *AIMS Mathematics*, 8(3), 5273–5291.

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., et al. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115, 213–237.

Zhao, R., Zhang, C., & Xue, D. (2024). A multi-scale multi-channel CNN introducing a channel-spatial attention mechanism for hyperspectral remote sensing image classification. *European Journal of Remote Sensing*, 57(1), 1–14.

Zheng, Z., Fu, J., Lu, C., & Zhu, Y. (2021). Research on rolling bearing fault diagnosis of small dataset based on a new optimal transfer learning network. *Measurement*, 177, 109285.

Zhou, J., Ma, R., & Chen, J. (2022, January). Bearing fault diagnosis based on 1D-CNN-LSTM-TL. In *7th International Conference on Electronic Technology and Information Science (ICETIS)* (pp. 1–6). IEEE.

BIOGRAPHIES



Duy An Huynh is a Ph.D. candidate at Ho Chi Minh City University of Technology (Ho Chi Minh city, Vietnam). He received the B.S. degree in Mechatronic Engineering Technology from Industrial University of Ho Chi Minh, Ho Chi Minh city, Viet Nam, in 2019 and the M.S. degree in Control Engineering and Automation from Ho Chi Minh City University of Technology, Ho Chi Minh city, Vietnam, in 2022. His research interests include fault diagnosis and predictive maintenance in industrial systems and artificial neural networks applications.



Thanh Tung Luu is an Associate Professor, Doctor. He has been teaching at the Ho Chi Minh City University of Technology from 2003 to the present. Currently, he is the Head of Department of Construction Machinery and Handling Equipment. The main research direction is the oscillation of the Heavy lifting equipment and specialized equipment.



Thanh Hai Nguyen is an Associate Professor, Doctor. He has been teaching at the Ho Chi Minh City University of Technology since 2000. His main research directions are ultrasonic technology, welding technology, advanced forming and machining processes, as well as the design, simulation, and manufacturing of mechanical devices and equipment.