

## Article

# Transformer Attention-Guided Dual-Path Framework for Bearing Fault Diagnosis

Saif Ullah <sup>1</sup>, Wasim Zaman <sup>1</sup> and Jong-Myon Kim <sup>1,2,\*</sup>

<sup>1</sup> Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, Republic of Korea; saifuou@mail.ulsan.ac.kr (S.U.); wasim94@mail.ulsan.ac.kr (W.Z.)

<sup>2</sup> PD Technology Co., Ltd., Ulsan 44610, Republic of Korea

\* Correspondence: jmkim07@ulsan.ac.kr; Tel.: +82-52-259-2217

## Abstract

Reliable bearing fault diagnosis plays an important role in maintaining the safety and performance of rotating machinery in industrial systems. Although deep learning models have achieved remarkable success in this field, their dependence on a single feature-extraction approach often restricts the diversity of learned representations and limits diagnostic accuracy. To overcome this limitation, this study proposes an attention-guided dual-path framework that integrates spatial and time–frequency feature learning with transformer-based classification for precise fault identification. In the proposed framework, vibration signals collected from an experimental bearing test rig are simultaneously processed through two complementary pipelines: one converts the signals into two-dimensional matrix images to extract spatial features, while the other transforms them into continuous wavelet transform (CWT) scalograms to capture fine-grained temporal and spectral information. The extracted features are fused through a lightweight transformer encoder with an attention mechanism that dynamically emphasizes the most informative representations. This fusion enables the model to effectively capture cross-domain dependencies and enhance discriminative capability. Experimental validation on an industrial vibration dataset demonstrates that the proposed model achieves 99.87% classification accuracy, outperforming conventional CNN and transformer-based approaches. The results confirm that integrating multi-domain features with attention-driven fusion significantly improves the robustness and generalization of deep learning models for intelligent bearing fault diagnosis.

**Keywords:** bearing; deep learning; fault diagnosis; predictive maintenance; transformer



Academic Editors: Jérémie Schutz and Christophe Sauvey

Received: 22 October 2025

Revised: 18 November 2025

Accepted: 19 November 2025

Published: 23 November 2025

**Citation:** Ullah, S.; Zaman, W.; Kim, J.-M. Transformer Attention-Guided Dual-Path Framework for Bearing Fault Diagnosis. *Appl. Sci.* **2025**, *15*, 12431. <https://doi.org/10.3390/app152312431>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Bearings are fundamental components of electric motors and are used in power plants, industrial facilities, and diverse transportation systems, including automobiles, aircraft, marine vessels, and space technologies. These bearings must withstand severe and heavy pressures and high speeds [1,2]. Over time, such conditions can cause faults in bearings, which may eventually lead to system failure. Such faults are reportedly responsible for nearly 45% of failures in electric motors [3]. Because bearings play a key role in machine performance, their faults can result in major problems, such as machine damage, delays in production, and risks to human safety [4]. The advancement of robust fault diagnosis (FD) techniques for rolling bearings is essential for maintaining the reliability and safety of mechanical systems.

Bearing FD techniques are generally classified into three main categories: model-based, empirical, and data-driven approaches. Model-based approaches use mathematical models to simulate bearing behavior, enabling analysis, diagnosis, and prediction of operational data for accurate fault identification [5–9]. In contrast, empirical approaches depend on the expertise and practical knowledge of specialists to interpret data and identify faults. However, the increasing complexity and rapid pace of advances in industrial machinery hinder the development of precise mathematical models that rely solely on existing domain knowledge [10–12].

With the rapid advancement of vibration sensing technologies and the remarkable progress in machine learning (ML) and deep learning (DL), data-driven approaches have become the predominant choice for FD [13]. Methods that rely on vibration monitoring have received considerable attention [14–16]. A typical ML-based approach for bearing FD typically includes signal processing, feature extraction, selection of optimal features, and classification. Traditional ML algorithms, including support vector machine [17,18], random forest [19,20], and K-nearest neighbor [21,22], have been used extensively in FD. However, these techniques rely heavily on expertise for feature selection, reducing their adaptability and efficiency in real-world scenarios [23].

In contrast, DL-based FD can rely entirely on algorithms. These methods can also create end-to-end systems that eliminate the need to extract features manually [24–26] or to directly perform frequency analysis on time-series data [27]. Convolutional neural networks (CNNs) are well-suited for extracting representative features as they reduce dependence on expert knowledge. By combining convolutional and pooling layers, CNNs can efficiently identify spatial patterns in data. For example, an adaptive denoising CNN model is proposed that removes the requirement to manually adjust denoising parameters [28]. Similarly, bi-LSTM networks excel at processing time-dependent data, automatically extracting meaningful features without manual intervention [23]. In another study, an intelligent hybrid FD model that integrates a wavelet kernel network with a bi-LSTM enhanced by an attention mechanism was used for fault detection [29]. This approach successfully handled the temporal noise and overlapping signal challenges often seen in industrial bearing fault data.

In rolling bearing fault diagnosis, DL has recently attracted significant attention for its ability to automatically extract meaningful features from raw data, unlike traditional ML (like SVM and random forest), and to improve performance by increasing neural network depth [30]. However, DL networks must accommodate gradient flow issues, which can hinder parameter optimization and potentially reduce FD accuracy. In response to these challenges, He et al. introduced ResNet, which uses residual connections to prevent vanishing and exploding gradient problems [31]. Zhang et al. [32] proposed an attention-enhanced ResNet for diagnosing faults in gearboxes. Their approach successfully extracted time–frequency features, enhanced frequency-band information, and improved overall accuracy [32]. Similarly, Liang et al. developed a rolling bearing FD method using a wavelet transform combined with an improved ResNet [33], and Zhao et al. introduced a deep residual shrinkage network for FD [34]. These methods collectively demonstrate the benefits of ResNet over traditional CNN (without skip connections and batch normalization).

Recently, transformer architectures have been increasingly adopted in machinery fault diagnosis due to their strong capability for global feature representation and contextual dependency modeling. For instance, the Interpretable Domain Adaptation Transformer (IDAT) proposed by Liu et al. [35] employs a multi-layer domain adaptation transformer to align feature distributions between domains and introduces an ensemble attention weighting mechanism to enhance interpretability. While such models effectively address domain adaptation challenges, they mainly focus on transferring knowledge across do-

mains rather than integrating complementary feature representations. In contrast, the proposed attention-guided dual-path transformer framework aims to enhance intra-domain fault diagnosis by jointly learning spatial and time–frequency features through adaptive attention-based fusion, leading to richer and more discriminative feature representations.

All the discussed bearing fault diagnosis methods are summarized in Table 1, highlighting their core methodologies, extracted features, strengths, and limitations.

**Table 1.** Summary of representative bearing fault diagnosis studies.

| Author            | Methodology/Model Used                         | Feature Type            | Strengths                                     | Limitations/Gap                              |
|-------------------|------------------------------------------------|-------------------------|-----------------------------------------------|----------------------------------------------|
| Wang et al. [28]  | CNN-based denoising for vibration signals      | Spatial features        | Removes noise and enhances signal clarity     | Limited temporal modeling capability         |
| Guo et al. [29]   | Hybrid wavelet + LSTM with attention           | Time–frequency features | Captures temporal dependencies                | High model complexity and computation        |
| He et al. [31]    | Deep residual CNN                              | Spatial features        | Solves vanishing gradient issue               | Lacks temporal feature extraction            |
| Zhang et al. [32] | Attention-based ResNet for gearbox FD          | Time–frequency features | Enhances discriminative fault features        | Focused on a single feature domain           |
| Liang et al. [33] | CWT + ResNet combination                       | Time–frequency          | Improves frequency-domain learning            | Limited cross-domain fusion                  |
| Zhao et al. [34]  | ResNet with shrinkage for denoising            | Spatial features        | Effective for noisy environments              | Limited interpretability                     |
| Liu et al. [35]   | Multi-layer Transformer with domain adaptation | Latent-domain features  | Enhances domain transfer and interpretability | Focused on inter-domain transfer, not fusion |

The previously mentioned studies have significantly advanced DL-based techniques for bearing FD. However, these methods still face several challenges. One major issue is that traditional DL models often struggle to extract comprehensive features from one-dimensional signals due to their limited ability to capture both local and global patterns. This can be addressed by transforming the one-dimensional signals into two-dimensional scalograms, such as the continuous wavelet transform (CWT), which provides time–frequency representations and enhances feature extraction. However, each type of scalogram can capture only certain aspects of the signal. For example, CWT scalograms can capture transient and non-stationary features due to their adaptive time–frequency resolution, but they may fail to represent global or long-term signal trends and can introduce redundant information because of overlapping scales.

To overcome these issues, this study proposes a dual-path framework that transforms one-dimensional vibration signals into two-dimensional matrix images and CWT scalograms, extracting rich spatial and time–frequency features using ResNet-50. These features are fused using a lightweight transformer encoder with a learnable CLS token, enabling attention-based fusion for accurate fault classification on industrial data. The contributions of this study are as follows:

1. This study transforms raw vibration signals into both two-dimensional matrix images and CWT scalograms. The proposed framework extracts spatial and time–frequency

features using a fine-tuned ResNet-50 to ensure richer and more diverse feature representations.

2. The study introduces a lightweight transformer encoder to learn attention-based interactions between the features obtained from the pipelines. This attention-based mechanism helps the model focus more on the most important features during training.
3. The classifier is designed using a transformer-based architecture, where a learnable CLS token aggregates feature-level information from both pipelines. The output from this token is passed through a fully connected layer for fault classification.
4. The proposed approach is validated on a real-world industrial dataset. Quantitative results prove the capability and effectiveness of the model.

The structure of the remaining paper is as follows: Section 2 discusses the technical foundations, Section 3 describes the experimental setup, Section 4 explains the proposed method, and Section 5 provides the results and discussion, followed by conclusions in Section 6.

## 2. Technical Foundation

### 2.1. Two-Dimensional Matrix Images

Signal transformation is an important step for analyzing and extracting meaningful features from one-dimensional signals. These signals are mathematically denoted as

$$x = [x_1, x_2, x_3, \dots, x_N]; t \in [0, T] \quad (1)$$

where  $x$  is the amplitude of vibrations sampled over time. While one-dimensional signals provide useful temporal information, they lack the spatial structure necessary for extracting higher-level features using a deep neural network architecture such as ResNet-50. To overcome this limitation, signal transformation techniques are employed to convert one-dimensional time-series data into two-dimensional matrices, thereby introducing a structured spatial organization that reflects amplitude variations across time windows.

In this transformation, the one-dimensional signal is divided into smaller windows, each of which is mapped into a row of the resulting two-dimensional matrix. This transformation preserves short-term temporal relationships between adjacent segments while organizing the signal into a two-dimensional format suitable for image-based deep learning models. Mathematically, it can be represented as

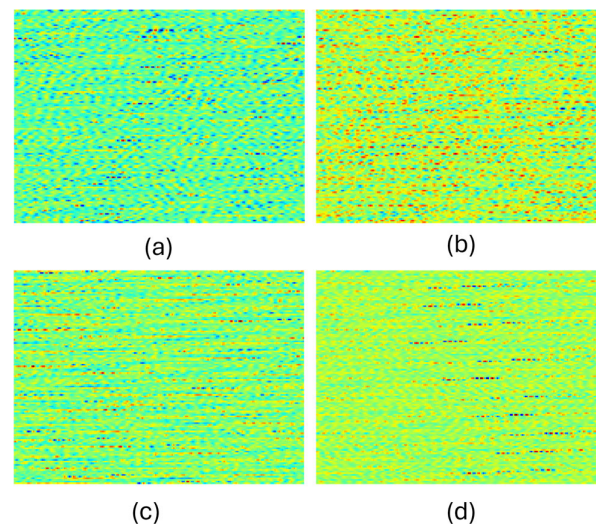
$$X_{2D} = \begin{bmatrix} x_1 & x_2 & \dots & x_m \\ x_{m+1} & x_{m+2} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(k-1)m+1} & x_{(k-1)m+2} & \dots & x_{km} \end{bmatrix} \quad (2)$$

where the matrix representation  $X_{2D}$  is the two-dimensional matrix representation of the signal,  $x_i$  denotes the amplitude values of the vibration signals, and  $m$  represents the number of columns (as determined by the window size). The term  $k$  denotes the integer-valued number of rows and is calculated as shown in Equation (3) [36].

$$k = \text{int}\left(\frac{N}{m}\right) \quad (3)$$

This transformation arranges the temporal information of the signal into a structured grid, where each row captures a specific window of the original signal. The resulting two-dimensional matrix image preserves the time-series integrity while introducing a spatial structure and is suitable for processing with deep neural network architectures

such as ResNet-50. Fault signatures, which may be localized in time, are more easily identifiable in a two-dimensional format, in which spatial arrangements can effectively highlight repeating patterns or anomalies. The transformation bridges the gap between one-dimensional time-series data and the two-dimensional spatial domain required by ResNet-50. By arranging the data into a matrix, local noise can be distributed across multiple cells, reducing its overall impact on the analysis [37]. The structured matrix representation scales efficiently for image-based processing pipelines, allowing deeper models to extract hierarchical features. Figure 1 shows two-dimensional matrix images of the faults.



**Figure 1.** Two-dimensional matrix images of bearing conditions: (a) inner race; (b) normal; (c) outer race; (d) roller.

## 2.2. Continuous Wavelet Transform

A CWT is a mathematical tool that maintains temporal resolution by localizing frequency components in time. This dual representation is essential for analyzing non-stationary signals, such as vibration signals from bearings, for which fault-related information is often localized in short time windows. A CWT decomposes a signal into scaled and translated versions of a mother wavelet, capturing both transient and periodic patterns. Mathematically, it can be represented as

$$\text{CWT}(a, b) = \int_{-\infty}^{+\infty} x(t) \overline{\psi_{(a,b)}(t)} dt \quad (4)$$

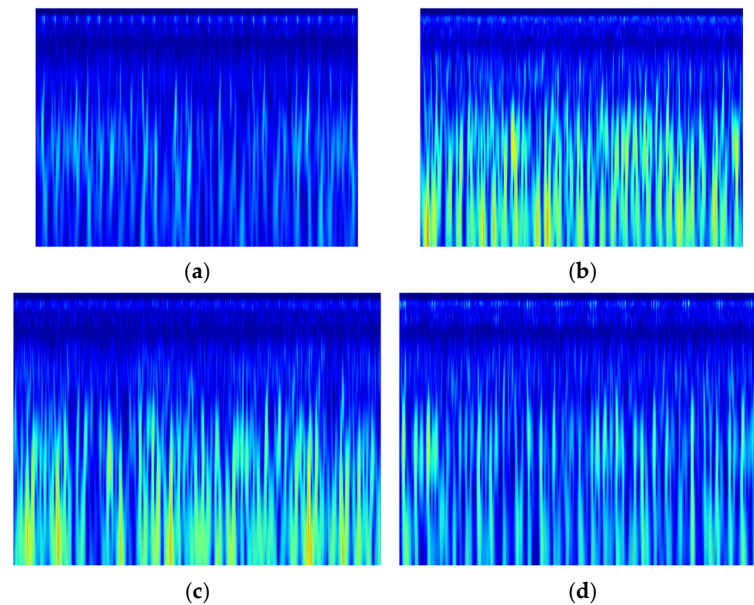
In Equation (4),  $a$  is the scale parameter responsible for controlling the frequency resolution,  $b$  denotes the translation parameter responsible for time localization, and  $\psi_{(a,b)}(t)$  represents the scaled and translated version of the mother wavelet. Mathematically,  $\psi_{(a,b)}(t)$  can be written as follows:

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (5)$$

In this study, the Morlet wavelet was used as the mother wavelet, owing to its superior time–frequency localization and suitability for non-stationary vibration signal analysis.

The resulting scalogram represents the energy distribution of the signal, providing rich information for fault analysis [38]. A CWT offers several advantages for signal analysis and is particularly effective for bearing FD. First, CWT provides time–frequency localization, enabling simultaneous representation of time and frequency information, which is

important for capturing dynamic changes in vibration signals. Second, it is well-suited for analyzing non-stationary vibration signals. Lastly, CWT generates scalograms, transforming raw signals into visually interpretable time–frequency images. These scalograms facilitate feature extraction and allow for more effective DL pattern analysis, enhancing the accuracy of FD. Figure 2 shows CWT scalograms of faults.



**Figure 2.** CWT images of bearing conditions: (a) inner race; (b) normal; (c) outer race; (d) roller.

Figure 2 presents the CWT scalograms generated using the Morlet (‘morl’) wavelet for four bearing conditions. The color intensity represents signal energy; yellow and green denote high energy, while blue indicates low energy. The normal bearing (b) shows uniform, low-energy patterns, reflecting stable operation. The inner race fault (a) displays intermittent high-energy streaks due to repeated impacts, while the outer race fault (c) exhibits periodic energy bursts at lower frequencies, indicating localized outer surface defects. The roller fault (d) shows irregular and dispersed high-energy zones caused by distributed surface wear. These distinct energy patterns demonstrate the CWT’s effectiveness in capturing transient, fault-related features for deep learning-based fault diagnosis.

### 2.3. ResNet-50

ResNet introduced residual learning through skip connections, enabling the training of networks with increased depth. It addresses the issue of gradient vanishing in the DL models. This architecture consists of 50 layers, primarily convolutional, with bottleneck residual blocks that allow gradients to flow smoothly during back-propagation. A key innovation of ResNet is the use of an explicitly defined identity mapping through skip connections, ensuring that useful features are retained across layers without degradation. This feature makes ResNet-50 particularly effective for extracting features from images generated from vibration signals. At its core, a residual block can be represented as follows:

$$y = F(x, W) + x \quad (6)$$

In Equation (6),  $x$  is the input,  $F$  represents the residual function parameterized by weights  $W$ , and  $y$  denotes the output feature map. The addition operation enforces identity mapping, allowing the network to learn residuals rather than try to directly map inputs to outputs. This identity mapping prevents the gradients from becoming excessively small

during back-propagation, solving the issue of vanishing gradients and allowing deep networks to converge effectively [31].

In ResNet-50, convolution layers are responsible for extracting features. Each convolution layer can be mathematically expressed as

$$F_{ij}^k = \sigma \left( \sum_m \sum_p \sum_q W_{pq}^{km} X_{((i+p),(j+q))}^m + b^k \right) \quad (7)$$

In Equation (7),  $F_{ij}^k$  is the output feature map for the  $k^{th}$  filter at position  $(i, j)$ , and  $\sigma$  denotes the activation function.  $W_{pq}^{km}$  is the filter weight at position  $(p, q)$  for the  $k^{th}$  filter at the  $m^{th}$  channel.  $X_{((i+p),(j+q))}^m$  represents the input feature map at position  $((i + p), (j + q))$  in the  $m^{th}$  channel.  $b^k$  is the bias term for the  $k^{th}$  filter. Figure 3 depicts the basic architecture of ResNet-50 [31].

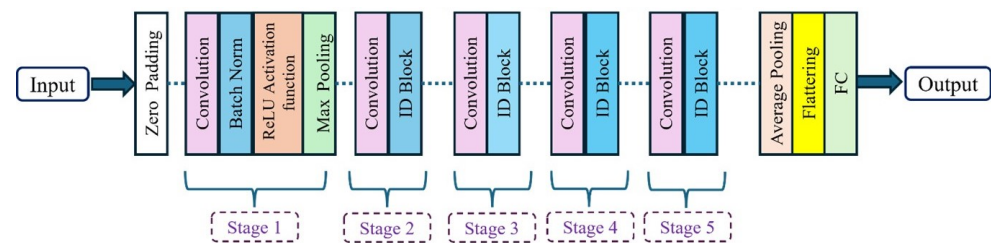


Figure 3. Basic ResNet-50 architecture.

### 3. Experimental Setup

To evaluate the effectiveness of the proposed model, vibration signals were collected from a bearing test rig developed at the Ulsan Industrial Artificial Intelligence (UIAI) Laboratory, University of Ulsan, Republic of Korea, as illustrated in Figure 4. The dataset comprises four bearing health states: normal condition, outer race fault (OR), inner race fault (IR), and roller fault. During data acquisition, the system was operated with a three-phase motor running steadily at 1800 rpm. The rotor shaft motion was transferred to the main shaft through a belt-drive mechanism connected to both sides of the test bearings. For precise and noise-free measurement, vibration sensors were mounted using a magnetic base on the left side of the target bearing, with two accelerometers oriented vertically and horizontally. Among them, the horizontal accelerometer data were used for analysis, as they provided more distinct fault-related signatures. Figure 5 depicts the schematic of the experimental setup, highlighting the arrangement and function of its components. This configuration enabled the collection of reliable vibration data suitable for assessing the proposed FD framework. The data acquisition system, outlined in Table 2, used a cylindrical roller-type bearing (FAG NJ206-3-TVP2). Signals were sampled at 25 kHz, ensuring acquisition of high-resolution data. The signals were segmented into 1 s intervals, and for each bearing condition, more than 300 data samples were obtained, with details summarized in Table 3.

The experimental process allows flexibility, as different fault types can be tested by simply replacing the test bearing in the existing setup without modifying other components. Figure 6 shows the bearings used in the experiment, highlighting the defect regions associated with each fault type.

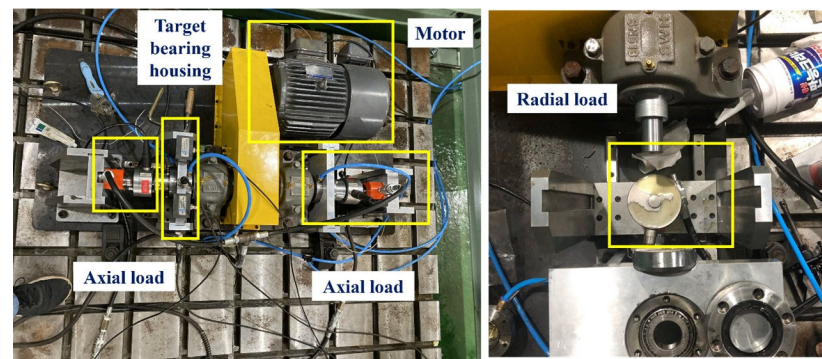


Figure 4. Experimental setup for bearing dataset collection.

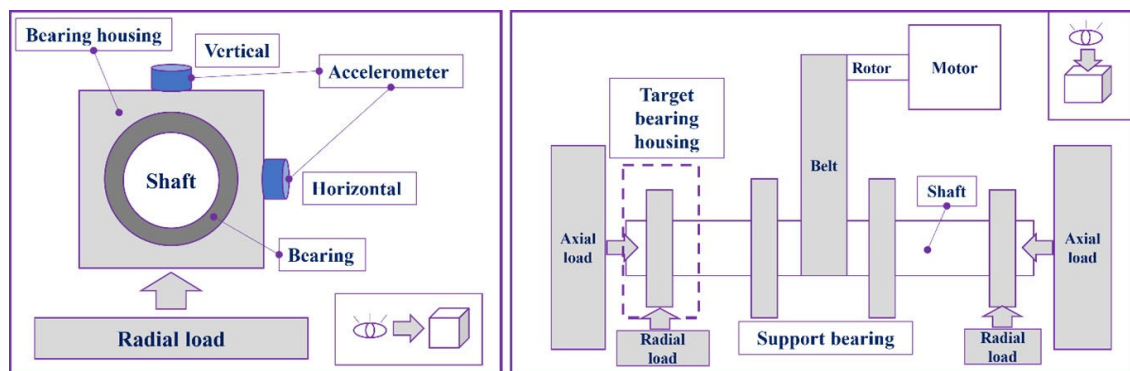


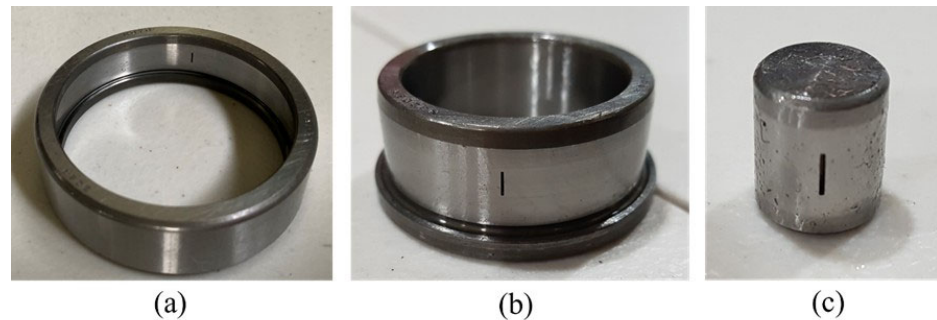
Figure 5. Schematics of experimental setup.

Table 2. Data acquisition system.

| Equipment                     | Parameter         | Value                                     |
|-------------------------------|-------------------|-------------------------------------------|
| Vibration sensor (PCB-622B01) | Sensor type       | IEPE-type piezoelectric accelerometer     |
|                               | frequency         | 0.2 to 15,000 Hz                          |
|                               | measurement range | $\pm 490 \text{ m/s}^2$                   |
|                               | sensitivity       | 100 mV/g                                  |
|                               | temperature       | $-40^\circ\text{C}$ to $70^\circ\text{C}$ |
| DAQ (NI9234)                  | dynamic range     | 102 dB                                    |
|                               | resolution        | 24 bits                                   |
|                               | Coupling          | AC (2 mA IEPE excitation)                 |

Table 3. Dataset details.

| Bearing State | No. of Samples | Individual Sample Time |
|---------------|----------------|------------------------|
| Normal        | 344            | 1 s                    |
| Inner race    | 370            | 1 s                    |
| Outer race    | 347            | 1 s                    |
| Roller        | 309            | 1 s                    |



**Figure 6.** Bearing fault conditions: (a) inner race; (b) outer race; (c) roller.

For model training and evaluation, the collected dataset was randomly divided into 80% for training and 20% for testing using a stratified split strategy to ensure balanced representation of all fault classes. The training subset was further divided internally into 90% for training and 10% for validation, which was used for hyperparameter tuning to improve generalization performance. No explicit data augmentation was applied, as the vibration dataset already included a wide range of fault conditions under controlled experimental settings, providing sufficient variability for effective model generalization.

#### 4. Proposed Method

This section presents a detailed explanation of the proposed FD approach, with its overall workflow illustrated in the flow diagram shown in Figure 7.

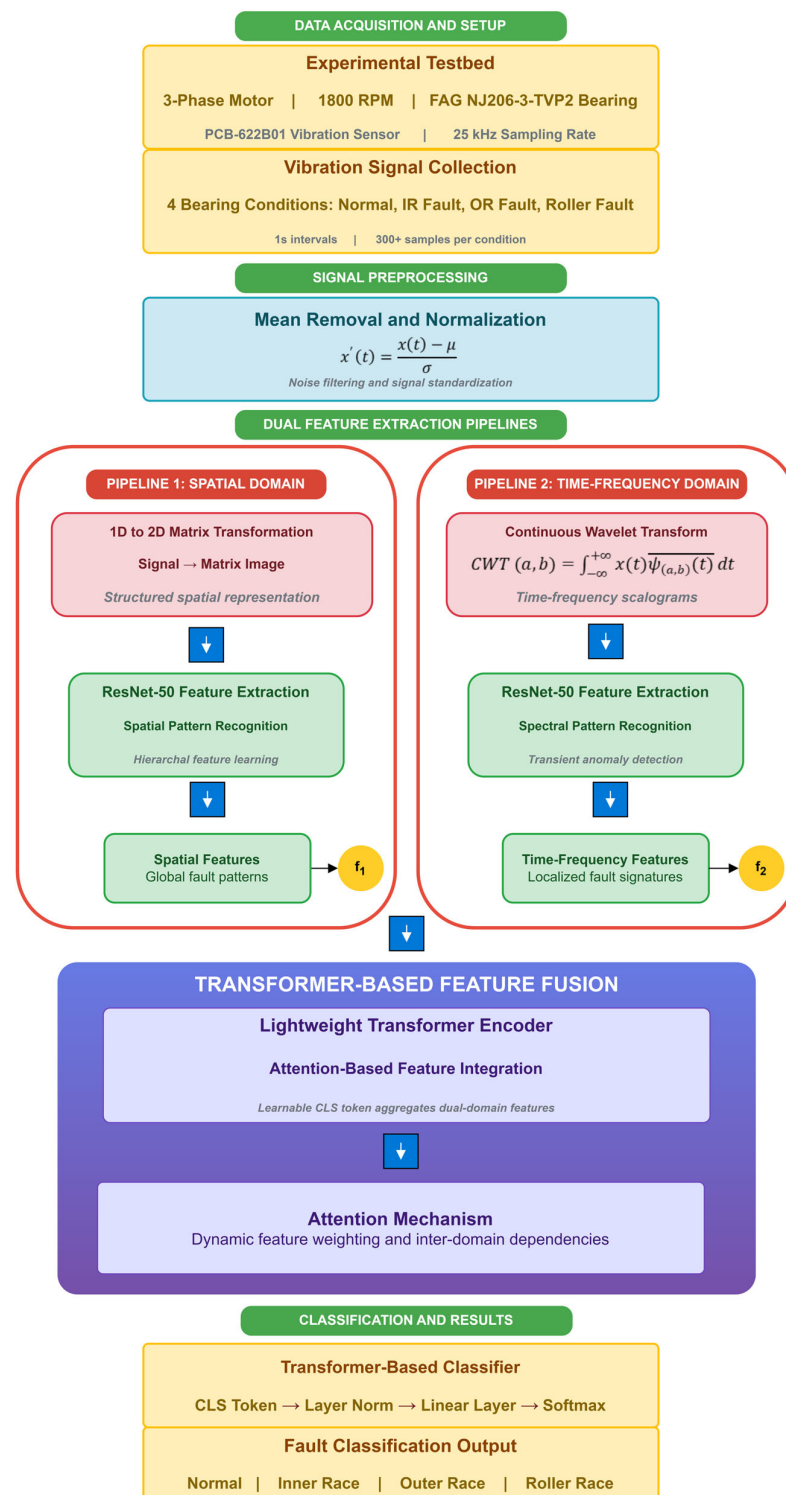
**STEP I:** The bearing FD framework begins with the collection of one-dimensional vibration signals as time-series data reflecting the dynamic behavior of bearings under varying conditions of normal operations and IR, OR, or roller defects. However, because raw vibration signals often contain noise, irregularities, and outliers that can hide meaningful fault-related patterns, preprocessing techniques are applied. In this study, means are removed to filter noise from the signals. This ensures that all signals are scaled to a standard range, represented mathematically as follows:

$$x'(t) = \frac{x(t) - \mu}{\sigma} \quad (8)$$

This operation standardizes the signal by centering it around zero mean and scaling it by its standard deviation. Such normalization improves numerical stability during training, ensures consistent feature representation across samples, and prevents features with larger magnitudes from disproportionately influencing the model.

**STEP II:** In pipeline 1, the preprocessed one-dimensional vibration signals (1-s segments sampled at 25 kHz) are transformed into two-dimensional matrices through structured mapping. The transformation involved segmenting the vibration signals into smaller windows, with each segment mapped into a row of the resulting two-dimensional matrix. The window size and overlap ratio were empirically selected to preserve adequate temporal context while maintaining computational efficiency. The resulting matrices were subsequently resized to  $224 \times 224$  pixels to match the ResNet-50 input dimension.

The advantage of this part of the proposed model is its ability to extract spatially hierarchical features. However, converting the signals into two-dimensional matrices may reduce access to subtle temporal variations, which are better retained in CWT-based scalograms. Despite this limitation, this part of the model excelled at capturing large-scale spatial dependencies and global fault patterns across the vibration signals. For consistency, each two-dimensional matrix was converted into a PNG image before being input to the ResNet-50 model for feature extraction.



**Figure 7.** Schematic diagram of the proposed method.

**STEP III:** In pipeline 2, the preprocessed signals were transformed into CWT scalograms. The CWT was applied directly to the signals, mapping them into the time–frequency domain. Unlike pipeline 1, this approach avoided additional preprocessing, thereby preserving the raw time–frequency characteristics and transient details present in the vibration signals. The resulting CWT scalograms were provided to the ResNet-50 network, where convolutional layers extracted discriminative features such as transient spikes, fault-related anomalies, and frequency-dependent patterns. This step is computationally efficient because it bypasses the intermediate matrix transformation stage. However, this process lacks

the hierarchical abstraction capabilities of pipeline 1. Despite these limitations, pipeline 2 effectively retained the intrinsic signal characteristics, making it suitable for identifying localized fault-related variations and transient events.

In the proposed dual-path design, each pipeline is tailored to capture complementary aspects of vibration signal characteristics. The first pipeline, which converts vibration signals into two-dimensional matrix images, emphasizes the global spatial organization of amplitude variations across consecutive time windows. This structured spatial representation helps identify periodic patterns, repetitive structures, and long-range dependencies within the signal. In contrast, the second pipeline based on CWT scalograms focuses on capturing localized and fine-grained time–frequency patterns that correspond to transient fault events, modulations, and frequency-dependent dynamics in the vibration response. Together, these representations provide both global spatial context and detailed temporal–spectral insights, enabling a more comprehensive fault characterization.

In the proposed dual-path architecture, two independent ResNet-50 networks are employed for feature extraction. Each network is initialized with ImageNet pre-trained weights, and weight sharing is not applied because the two input representations (two-dimensional matrix images and CWT scalograms) have distinct spatial and spectral characteristics. To adapt the networks effectively, all convolutional layers are frozen except the final residual block and the fully connected layer, which are fine-tuned independently to capture modality-specific discriminative features.

**STEP IV:** Once the features are extracted from the two pipelines, they are jointly processed using a lightweight transformer encoder. Instead of conventional concatenation, the model introduces a learnable classification token (CLS), which guides the transformer to perform attention-based feature fusion. This mechanism dynamically learns to weigh and relate spatial and time–frequency features while simultaneously producing a compact, lower-dimensional representation. The attention mechanism further emphasizes the most informative components and suppresses irrelevant or redundant information. As a result, the model not only enriches the feature representation but also improves robustness to varying signal characteristics and noise.

The lightweight transformer encoder consisted of a single encoder layer with four multi-head attention modules. Each head received the fused feature embeddings obtained by concatenating the outputs of both ResNet-50 branches. A learnable [CLS] token was prepended to the embedding sequence and served as a global representation during training. Each feature vector from both pipelines was treated as an individual token within the sequence. The output corresponding to the [CLS] token was passed through a normalization layer and a fully connected classifier to produce the final fault-type probabilities.

**STEP V:** The final classification is performed using the output of the CLS token from the transformer encoder. This token serves as a summary of the learned interaction between the input features from both pipelines. It is passed through a lightweight classification head comprising a normalization layer and a fully connected linear layer to generate class probabilities. Unlike conventional ANN-based classifiers, the use of a transformer encoder provides adaptive attention modeling, enabling the framework to generalize better across varying fault types and operating conditions. This approach also simplifies the architecture while maintaining high accuracy and interpretability. The full configuration details of the proposed model, including image size, ResNet-50 settings, transformer parameters, and training hyperparameters, are summarized in Table 4.

**Table 4.** Proposed model parameters.

| Parameter                         | Value/Description                                       |
|-----------------------------------|---------------------------------------------------------|
| Image size                        | $224 \times 224$                                        |
| Batch size                        | 32                                                      |
| Train/Test Split                  | 80%/20%                                                 |
| <b>ResNet-50 (Both Pipelines)</b> |                                                         |
| Pre-trained weights               | ImageNet                                                |
| Frozen layers                     | All except final block                                  |
| Output features                   | 2048 (per pipeline)                                     |
| <b>Transformer Encoder</b>        |                                                         |
| Transformer layers                | 1                                                       |
| Attention Heads                   | 4                                                       |
| Embedding dimension               | 2048                                                    |
| Feed-forward dimension            | 4096                                                    |
| Dropout rate                      | 0.1                                                     |
| Token sequence                    | Concatenated features from both pipelines + [CLS] token |
| Position encoding                 | Learnable, added to each token embedding                |
| CLS Token                         | $1 \times 1 \times 2048$                                |
| <b>Classifier</b>                 |                                                         |
| Classifier Head                   | Layer Norm (2048)                                       |
|                                   | Linear (2048 $\rightarrow$ 4)                           |
| Activation                        | None                                                    |
| <b>Training</b>                   |                                                         |
| Loss Function                     | Cross-Entropy Loss                                      |
| Optimizer                         | Adam                                                    |
| Learning rate                     | 0.0001                                                  |
| Epochs                            | 20                                                      |

## 5. Results and Discussion

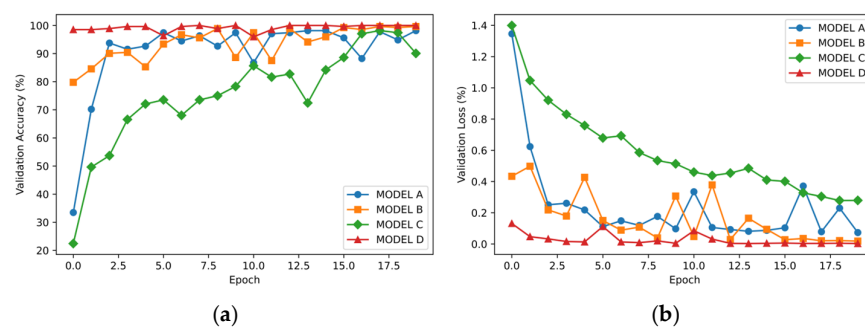
The proposed model is compared with three other models. In Model A, two-dimensional matrix images are passed through a pre-trained ResNet-50 model to extract deep spatial features. The extracted features are then fed into a lightweight transformer encoder. A learnable CLS token is provided to the input and interacts with the image features through self-attention. The output from the CLS token is passed to a linear classification layer, which produces the final class predictions.

Model B focuses on classifying bearing faults using time–frequency representations derived from raw vibration signals. These signals are first transformed into CWT scalograms, which capture localized frequency components over time. The resulting scalogram images are fed as input to a pre-trained ResNet-50 model, with the output layer removed to align with the number of output classes. This allows ResNet-50 to act solely as a deep feature extractor. The extracted features are then passed into a lightweight transformer encoder, where a learnable classification CLS token is provided to the input feature sequence. The transformer uses self-attention to understand how parts of the input relate to each other and

focuses on the most important features. The output from the CLS token is then processed by a linear layer that performs the final fault classification.

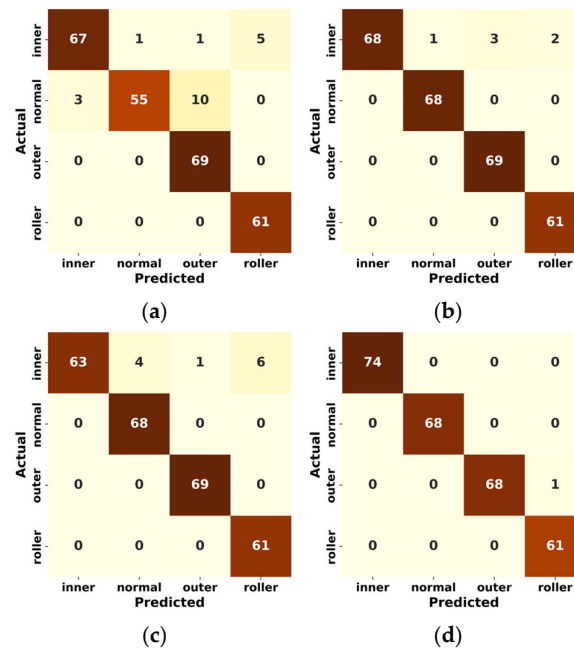
The third model, namely Model C, used for comparison, is a state-of-the-art model that uses a hybrid approach to generate CWT scalograms of one-dimensional signals. The generated scalograms are converted to greyscale, reducing computational complexity while retaining essential information. A custom CNN was then constructed, featuring three convolutional layers with rectified linear unit activation, followed by max-pooling layers for spatial reduction and dense layers for feature abstraction [39]. Separately, features extracted from the CNN were used to train a random forest classifier using extracted features and labels, and its accuracy was evaluated as validation features. This dual-step approach combines the CNN feature-extraction power with the decision-making capability of ensemble-based random forests, aiming to enhance classification performance. The CNN in Model C consists of three convolutional layers (32, 64, and 128 filters with  $3 \times 3$  kernels), each followed by ReLU activation and max-pooling layers, and two fully connected layers for feature abstraction. These features are then classified using a gcForest (deep forest) ensemble.

All three comparison models and the proposed model were tested on the collected dataset. In this study, 20% of the data were reserved for testing, while the remaining 80% were used for training. Each model was trained for 20 epochs, and the validation accuracy was monitored at every epoch. If a model achieved a higher validation accuracy than previously recorded, its parameters were saved, overwriting the earlier checkpoint. In this way, the best-performing model checkpoint based on validation accuracy was retained for each method. Figure 8 presents the validation accuracy and loss comparison plots. The proposed approach achieved higher validation accuracy compared to the reference models. Similarly, it also exhibited lower validation loss, further confirming its superior generalization performance.



**Figure 8.** Validation comparison: (a) accuracy; (b) loss.

A confusion matrix serves as a visual summary that compares the predicted labels of a classification model against the actual labels, helping assess how accurately the model distinguishes between different classes. It plays a crucial role in evaluating model performance by providing class-wise insight into correct and incorrect predictions. The confusion matrices of all the models are presented in Figure 9. In these matrices, the numbers represent the count of test samples for each fault class; diagonal values correspond to correctly classified samples, while off-diagonal values indicate misclassifications. This allows both overall and class-specific performance to be visually assessed, helping identify which fault types the model predicts most accurately and where confusion occurs.



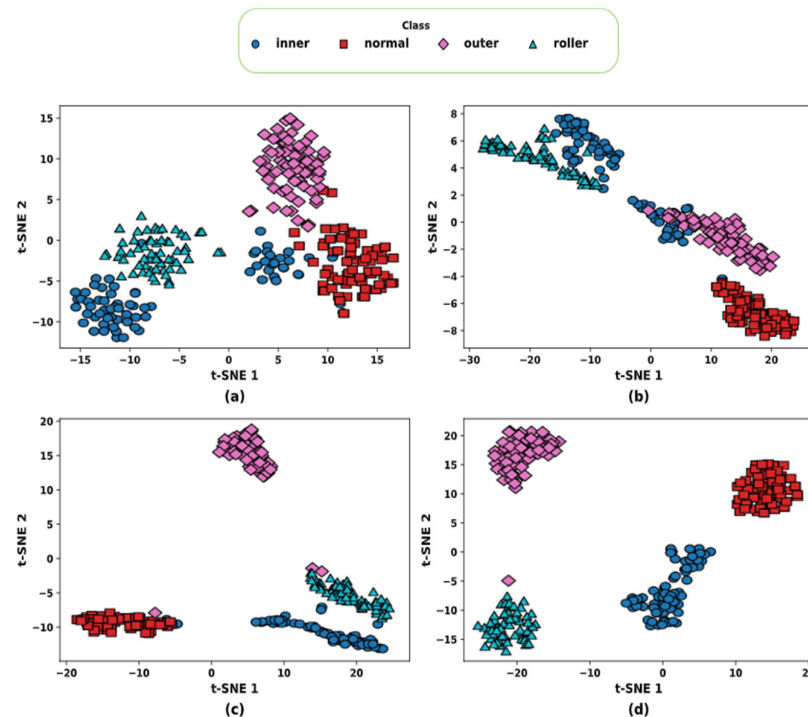
**Figure 9.** Confusion matrices for (a) Model A, (b) Model B, (c) Model C, and (d) Model D (proposed method).

The t-distributed stochastic neighbor embedding (t-SNE) plots of all the models are presented in Figure 10. The proposed model clearly separates the features of all the faults in a precise and clear manner compared to the other models. The proposed model achieved greater accuracy than the comparison models. Similarly, the proposed model had better precision, recall, and F1 score. Table 5 highlights the metric scores of all the models. Model A is robust; however, its reliance on a significant number of preprocessing steps can filter out raw, localized time–frequency details, missing potential key signal anomalies. The proposed model addressed this limitation through pipeline 2, which directly captures raw time–frequency details using CWT scalograms. This combination allowed the proposed model to capture both local and global fault patterns, resulting in a more comprehensive fault-detection framework.

**Table 5.** Fivefold average performance of the proposed and comparison models.

| Model            | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) | Trainable Parameters |
|------------------|--------------|---------------|------------|--------------|----------------------|
| Model A          | 96.86        | 96.88         | 96.91      | 96.84        | 25,419,244           |
| Model B          | 98.10        | 98.20         | 98.11      | 98.13        | 25,419,244           |
| Model C          | 93.43        | 94.47         | 93.43      | 93.29        | 18,907,652           |
| Model D-Proposed | 99.27        | 99.33         | 99.25      | 99.28        | 101,636              |

Model B retained raw time–frequency characteristics but struggled to balance the extraction of global hierarchical patterns and local details. The proposed model overcame this limitation through the dual pathways of pipeline 1 and pipeline 2. Pipeline 1 introduces a structured-matrix transformation, enabling hierarchical spatial feature extraction that complements the fine-grained time–frequency information captured by pipeline 2. By integrating the features extracted from both pipelines through attention feature fusion, the proposed model achieves a more comprehensive representation, enhancing its ability to detect both global and local fault patterns.



**Figure 10.** t-SNE plots: (a) Model A; (b) Model B; (c) Model C; (d) Model D (proposed).

The state-of-the-art Model C [39] uses a hybrid approach in which greyscale CWT scalograms are processed using a custom CNN for feature extraction, followed by classification through a random forest classifier. While this approach effectively combines CNN-based feature extraction with the ensemble decision-making capability of a random forest, relying on greyscale scalograms can lead to a loss of critical spectral details. Furthermore, the separate optimization of a CNN and a random forest may result in suboptimal utilization of the extracted features. In contrast, the proposed model retained richer information by utilizing both two-dimensional matrix images and CWT scalograms, providing diverse and informative feature representations.

The proposed dual-pipeline framework performed better than the other methods, as it utilizes both structured spatial and raw time–frequency features, integrating them through an attention-guided transformer encoder. This enables dynamic weighting of features and robust fault classification. The final classification, driven by the CLS token, benefits from enriched feature interactions, allowing the model to outperform all comparison models.

For complex fault classification tasks, deeper architectures such as ResNet-50 are essential to achieve effective generalization across varying operating conditions. However, training such deep networks from scratch requires large-scale datasets, the use of which was impractical in this study. Therefore, fine-tuned pre-trained ResNet-50 models were adopted to leverage previously learned representations while reducing overfitting risk. Although the proposed dual-pipeline framework employs two ResNet-50 branches, both operate in parallel to extract spatial and time–frequency features simultaneously. This design naturally increases the total number of parameters but does not significantly affect the computational time complexity, as both branches process inputs concurrently. Consequently, while the model contains more parameters than simpler networks, its efficiency and inference speed remain comparable, and the additional parameters contribute directly to improved feature representation and diagnostic accuracy.

To ensure a fair and robust evaluation of all models, we performed 5-fold cross-validation for each classification method, including the comparison models and the proposed model (Model D). In this setup, the dataset was partitioned into five equal subsets.

During each fold, four subsets were used for training while the remaining one was used for testing. This process was repeated five times, and the final performance metrics were reported as the average across all folds. The results, summarized in Table 5, clearly demonstrate the consistent superiority of Model D while validating the generalization capability of all evaluated models.

In addition to performance metrics, the number of trainable parameters for each model was also reported to ensure a fair comparison of model complexity. These values, listed in Table 5. Although the proposed model contains a higher number of trainable parameters, this increase is deliberate and necessary for handling the complexity of the task. Learning fine-grained, modality-specific discriminative features requires adapting deeper convolutional layers, which cannot be effectively trained from scratch on a limited dataset. To overcome this limitation, the two ResNet branches were briefly fine-tuned independently to adjust their high-level filters, and then frozen again to ensure stable and meaningful feature extraction. This controlled fine-tuning strategy increases the number of trainable parameters but ultimately enables the model to capture richer representations while avoiding overfitting, leading to superior performance.

To further verify the generalization capability of the proposed framework, an additional experiment was performed using the Paderborn University open source bearing fault dataset, which contains real bearing fault conditions collected under different operating speeds and loads [40]. The proposed model, trained on the original dataset, was evaluated on this external dataset without retraining. The results, summarized in Table 6, demonstrate that the model maintains consistently high performance across datasets, confirming its ability to generalize to unseen bearings and real fault conditions.

**Table 6.** Performance metrics of the proposed and comparison models on the Lab and Paderborn datasets.

| Model             | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) |
|-------------------|--------------|---------------|------------|--------------|
| Lab Dataset       |              |               |            |              |
| Model A           | 92.65        | 93.15         | 92.86      | 92.61        |
| Model B           | 97.79        | 97.80         | 97.97      | 97.83        |
| Model C           | 95.96        | 96.02         | 96.28      | 95.93        |
| Model D-Proposed  | 99.63        | 99.60         | 99.64      | 99.61        |
| Paderborn Dataset |              |               |            |              |
| Model A           | 70.39        | 77.68         | 69.02      | 67.67        |
| Model B           | 90.96        | 91.00         | 90.57      | 90.69        |
| Model C           | 84.87        | 85.41         | 84.00      | 84.39        |
| Model D-Proposed  | 98.43        | 98.47         | 98.32      | 98.38        |

## 6. Conclusions

In this study, an attention-guided hybrid framework for bearing fault diagnosis was proposed, combining dual-pipeline feature extraction and transformer-based classification. The first pipeline converted preprocessed vibration signals into structured two-dimensional matrix images to capture global spatial dependencies using a fine-tuned ResNet-50 network. The second pipeline transformed raw vibration signals into continuous wavelet transform (CWT) scalograms, effectively preserving localized and fine-grained time–frequency details associated with transient fault events. Features extracted from both pipelines were inte-

grated through a lightweight transformer encoder equipped with an attention mechanism and a learnable CLS token, enabling adaptive fusion and contextual understanding between spatial and time–frequency features.

Experimental validation on the laboratory-collected bearing dataset demonstrated the superior performance of the proposed model, achieving 99.87% classification accuracy, outperforming several benchmark models. Furthermore, evaluation on the Paderborn University open source bearing dataset confirmed the model’s strong generalization ability (98.43% accuracy) across different machines, operating speeds, and load conditions, establishing its robustness and adaptability. The proposed framework not only improves fault classification accuracy but also enhances interpretability through its attention-driven fusion, allowing insights into which features contribute most to decision-making.

These results collectively demonstrate that integrating multi-domain representations via an attention-based fusion strategy significantly improves diagnostic precision and reliability in vibration-based fault diagnosis. The proposed approach offers a promising direction toward developing generalizable, data-efficient, and interpretable deep learning systems for rotating machinery. Future work will focus on extending this framework to multi-sensor fusion, cross-machine transfer learning, and real-time industrial monitoring, as well as exploring lightweight deployment on embedded systems for intelligent predictive maintenance.

**Author Contributions:** Conceptualization, S.U., W.Z. and J.-M.K.; methodology, S.U., W.Z. and J.-M.K.; validation, S.U., W.Z. and J.-M.K.; formal analysis, S.U., W.Z. and J.-M.K.; resources, S.U., W.Z. and J.-M.K.; writing—original draft preparation, S.U. and W.Z.; writing—review and editing, J.-M.K.; visualization, S.U. and W.Z.; project administration, J.-M.K.; funding acquisition, J.-M.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Technology Innovation Program (‘20023566,’ ‘Development and Demonstration of Industrial IoT and AI-Based Process Facility Intelligence Support System in Small and Medium Manufacturing Sites’) funded by the Ministry of Trade, Industry, & Energy (MOTIE, Korea). This result was also supported by the “Regional Innovation System & Education (RISE)” through the Ulsan RISE Center, funded by the Ministry of Education (MOE) and the Ulsan Metropolitan City, Republic of Korea (2025-RISE-07-001).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors upon request.

**Conflicts of Interest:** Author Jong-Myon Kim was employed by the company Prognosis and Diagnostic Technologies Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Nomenclature

| Symbol                           | Description                                                                          |
|----------------------------------|--------------------------------------------------------------------------------------|
| $x(t)$                           | One-dimensional vibration signal as a function of time                               |
| $X_{2D}$                         | Two-dimensional matrix representation of the vibration signal                        |
| $\psi\left(\frac{t-b}{a}\right)$ | Mother wavelet function                                                              |
| $CWT(a, b)$                      | Continuous wavelet transform of signal $x(t)$                                        |
| $F(x, W)$                        | Residual mapping function parameterized by weights $W$                               |
| $W_{pq}^{km}$                    | Weight of the convolution kernel at position $(p, q)$ for filter $k$ and channel $m$ |

|            |                                                                        |
|------------|------------------------------------------------------------------------|
| $F_{ij}^k$ | Output feature map of the $k^{th}$ filter at spatial position $(i, j)$ |
| $b^k$      | Bias term associated with the $k^{th}$ filter                          |
| CLS        | Learnable classification token used in the transformer encoder         |

## References

- Mauricio, A.; Gryllias, K. Cyclostationary-based Multiband Envelope Spectra Extraction for bearing diagnostics: The Combined Improved Envelope Spectrum. *Mech. Syst. Signal Process.* **2021**, *149*, 107150. [\[CrossRef\]](#)
- Matania, O.; Dattner, I.; Bortman, J.; Kenett, R.S.; Parmet, Y. A systematic literature review of deep learning for vibration-based fault diagnosis of critical rotating machinery: Limitations and challenges. *J. Sound Vib.* **2024**, *590*, 118562. [\[CrossRef\]](#)
- Bazurto, A.J.; Quispe, E.C.; Mendoza, R.C. Causes and failures classification of industrial electric motor. In Proceedings of the 2016 IEEE ANDESCON, Arequipa, Peru, 19–21 October 2016; pp. 1–4. [\[CrossRef\]](#)
- Zhang, L.; He, X.; Chen, J.; Liu, J. Fault diagnoses of a nonlinear cracked rotor-bearing system based on vibration energy space and incremental learning approach. *J. Sound Vib.* **2025**, *600*, 118785. [\[CrossRef\]](#)
- Yu, Z.; Zhang, C.; Deng, C. An improved GNN using dynamic graph embedding mechanism: A novel end-to-end framework for rolling bearing fault diagnosis under variable working conditions. *Mech. Syst. Signal Process.* **2023**, *200*, 110534. [\[CrossRef\]](#)
- Feng, K.; Ji, J.C.; Ni, Q.; Li, Y.; Mao, W.; Liu, L. A novel vibration-based prognostic scheme for gear health management in surface wear progression of the intelligent manufacturing system. *Wear* **2023**, *522*, 204697. [\[CrossRef\]](#)
- Gao, S.; Yu, Y.; Zhang, Y. Reliability assessment and prediction of rolling bearings based on hybrid noise reduction and BOA-MKRV. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105391. [\[CrossRef\]](#)
- López, C.; Naranjo, Á.; Lu, S.; Moore, K.J. Hidden Markov Model based Stochastic Resonance and its Application to Bearing Fault Diagnosis. *J. Sound Vib.* **2022**, *528*, 116890. [\[CrossRef\]](#)
- Shi, J.; Hu, J.; Luo, Y.; Yu, Y.; Baddour, N.; Huang, W.; Shen, C.; Zhu, Z. Three-dimensional dynamic modeling and vibration analysis of roller bearings under compound fault excitation. *J. Sound Vib.* **2025**, *613*, 119188. [\[CrossRef\]](#)
- Wang, Y.; Wu, J.; Yu, Z.; Hu, J.; Zhou, Q. A structurally re-parameterized convolution neural network-based method for gearbox fault diagnosis in edge computing scenarios. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107091. [\[CrossRef\]](#)
- Chen, B.; Zhang, W.; Gu, J.X.; Song, D.; Cheng, Y.; Zhou, Z.; Gu, F.; Ball, A.D. Product envelope spectrum optimization-gram: An enhanced envelope analysis for rolling bearing fault diagnosis. *Mech. Syst. Signal Process.* **2023**, *193*, 110270. [\[CrossRef\]](#)
- Liu, Y.; Xiang, H.; Jiang, Z.; Xiang, J. Second-order transient-extracting S transform for fault feature extraction in rolling bearings. *Reliab. Eng. Syst. Saf.* **2023**, *230*, 108955. [\[CrossRef\]](#)
- Siddique, M.F.; Zaman, W.; Ullah, S.; Umar, M.; Saleem, F.; Shon, D.; Yoon, T.H.; Yoo, D.-S.; Kim, J.-M. Advanced Bearing-Fault Diagnosis and Classification Using Mel-Scalograms and FOX-Optimized ANN. *Sensors* **2024**, *24*, 7303. [\[CrossRef\]](#) [\[PubMed\]](#)
- Xu, G.; Liu, M.; Jiang, Z.; Shen, W.; Huang, C. Online Fault Diagnosis Method Based on Transfer Convolutional Neural Networks. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 509–520. [\[CrossRef\]](#)
- Shi, H.; Li, Y.; Bai, X.; Zhang, K.; Sun, X. A two-stage sound-vibration signal fusion method for weak fault detection in rolling bearing systems. *Mech. Syst. Signal Process.* **2022**, *172*, 109012. [\[CrossRef\]](#)
- Altaf, M.; Akram, T.; Khan, M.A.; Iqbal, M.; Ch, M.M.I.; Hsu, C.-H. A New Statistical Features Based Approach for Bearing Fault Diagnosis Using Vibration Signals. *Sensors* **2022**, *22*, 2012. [\[CrossRef\]](#)
- Wang, Z.; Yao, L.; Chen, G.; Ding, J. Modified multiscale weighted permutation entropy and optimized support vector machine method for rolling bearing fault diagnosis with complex signals. *ISA Trans.* **2021**, *114*, 470–484. [\[CrossRef\]](#)
- Wei, J.; Huang, H.; Yao, L.; Hu, Y.; Fan, Q.; Huang, D. New imbalanced bearing fault diagnosis method based on Sample-characteristic Oversampling Technique (SCOTE) and multi-class LS-SVM. *Appl. Soft Comput.* **2021**, *101*, 107043. [\[CrossRef\]](#)
- Prasojo, R.A.; Putra, M.A.A.; Ekojono; Apriyani, M.E.; Rahmanto, A.N.; Ghoneim, S.S.; Mahmoud, K.; Lehtonen, M.; Darwish, M.M. Precise transformer fault diagnosis via random forest model enhanced by synthetic minority over-sampling technique. *Electr. Power Syst. Res.* **2023**, *220*, 109361. [\[CrossRef\]](#)
- Chen, S.; Yang, R.; Zhong, M. Graph-based semi-supervised random forest for rotating machinery gearbox fault diagnosis. *Control Eng. Pract.* **2021**, *117*, 104952. [\[CrossRef\]](#)
- Elshenawy, L.M.; Chakour, C.; Mahmoud, T.A. Fault detection and diagnosis strategy based on k-nearest neighbors and fuzzy C-means clustering algorithm for industrial processes. *J. Frankl. Inst.* **2022**, *359*, 7115–7139. [\[CrossRef\]](#)
- Kumar, H.S.; Manjunath, S.H. Use of empirical mode decomposition and K- nearest neighbour classifier for rolling element bearing fault diagnosis. *Mater. Today Proc.* **2022**, *52*, 796–801. [\[CrossRef\]](#)
- Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [\[CrossRef\]](#)
- Khorram, A.; Khalooei, M.; Rezghi, M. End-to-end CNN + LSTM deep learning approach for bearing fault diagnosis. *Appl. Intell.* **2021**, *51*, 736–751. [\[CrossRef\]](#)

25. Chen, J.; Hu, W.; Cao, D.; Zhang, Z.; Chen, Z.; Blaabjerg, F. A Meta-Learning Method for Electric Machine Bearing Fault Diagnosis Under Varying Working Conditions with Limited Data. *IEEE Trans. Industr. Inform.* **2023**, *19*, 2552–2564. [\[CrossRef\]](#)
26. Hoang, D.T.; Kang, H.J. A Motor Current Signal-Based Bearing Fault Diagnosis Using Deep Learning and Information Fusion. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 3325–3333. [\[CrossRef\]](#)
27. Yan, X.; She, D.; Xu, Y. Deep order-wavelet convolutional variational autoencoder for fault identification of rolling bearing under fluctuating speed conditions. *Expert Syst. Appl.* **2023**, *216*, 119479. [\[CrossRef\]](#)
28. Wang, Q.; Xu, F. A novel rolling bearing fault diagnosis method based on Adaptive Denoising Convolutional Neural Network under noise background. *Measurement* **2023**, *218*, 113209. [\[CrossRef\]](#)
29. Wang, J.; Guo, J.; Wang, L.; Yang, Y.; Wang, Z.; Wang, R. A hybrid intelligent rolling bearing fault diagnosis method combining WKN-BiLSTM and attention mechanism. *Meas. Sci. Technol.* **2023**, *34*, 085106. [\[CrossRef\]](#)
30. Hakim, M.; Omran, A.A.B.; Ahmed, A.N.; Al-Waily, M.; Abdellatif, A. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Eng. J.* **2023**, *14*, 101945. [\[CrossRef\]](#)
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
32. Zhang, K.; Tang, B.; Deng, L.; Liu, X. A hybrid attention improved ResNet based fault diagnosis method of wind turbines gearbox. *Measurement* **2021**, *179*, 109491. [\[CrossRef\]](#)
33. Liang, P.; Wang, W.; Yuan, X.; Liu, S.; Zhang, L.; Cheng, Y. Intelligent fault diagnosis of rolling bearing based on wavelet transform and improved ResNet under noisy labels and environment. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105269. [\[CrossRef\]](#)
34. Zhao, M.; Zhong, S.; Fu, X.; Tang, B.; Pecht, M. Deep Residual Shrinkage Networks for Fault Diagnosis. *IEEE Trans. Ind. Inform.* **2020**, *16*, 4681–4690. [\[CrossRef\]](#)
35. Liu, D.; Cui, L.; Wang, G.; Cheng, W. Interpretable domain adaptation transformer: A transfer learning method for fault diagnosis of rotating machinery. *Struct. Health Monit.* **2025**, *24*, 1187–1200. [\[CrossRef\]](#)
36. Wang, Z.; Oates, T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In Proceedings of the Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–26 January 2015; pp. 91–96.
37. Ahmed, H.O.A.; Nandi, A.K. Vibration Image Representations for Fault Diagnosis of Rotating Machines: A Review. *Machines* **2022**, *10*, 1113. [\[CrossRef\]](#)
38. Yan, R.; Gao, R.X.; Chen, X. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Process.* **2014**, *96*, 1–15. [\[CrossRef\]](#)
39. Xu, Y.; Li, Z.; Wang, S.; Li, W.; Sarkodie-Gyan, T.; Feng, S. A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurement* **2021**, *169*, 108502. [\[CrossRef\]](#)
40. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition Monitoring of Bearing Damage in Electromechanical Drive Systems by Using Motor Current Signals of Electric Motors: A Benchmark Data Set for Data-Driven Classification. *PHM Soc. Eur. Conf.* **2016**, *3*. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.