ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 29 December 2025; Accepted: 6 March 2026

https://doi.org/10.17559/TV-20251229003248
Original scientific paper

Multi-Modal Attention-Based Automatic Diagnosis of Bearing Vibration Faults for Robust Mechanical Condition Monitoring

Canan TASTIMUR, Volkan KAYA*

Abstract: Bearing components are critical elements of rotating machinery, and their failures can lead to sudden shutdowns, reduced efficiency, and increased maintenance costs. To address the limitations of conventional fault diagnosis methods, this study proposes a multimodal deep learning framework that integrates raw time-domain vibration signals with stationary wavelet transform (SWT)-based time-frequency representations using a novel bidirectional cross-attention mechanism. This architecture enables effective and mutual interaction between heterogeneous features, significantly enhancing discriminative representation learning and fault classification performance in complex operating environments. The proposed model was evaluated on the public SUBVF1.0 dataset under three fault conditions, achieving training and validation accuracies of 99.17% and 95.63%, respectively. The model also demonstrated strong performance in terms of precision, recall, F1-score, and ROC analysis. An ablation study further confirms the contribution of each model component. These results indicate that the proposed approach is accurate, robust, and suitable for practical industrial fault diagnosis applications.

Keywords: bearing fault diagnosis; bidirectional cross-attention; multimodal deep learning; rotating machinery

1 INTRODUCTION

Bearings ensure the proper operation of gear mechanisms and rotating motors in industrial systems. While they are the most critical components of rotating machinery, they directly affect the system's performance, safety, and efficiency. Over time, bearing wear and load imbalances lead to failures. Consequently, bearing failures can result in increased maintenance costs and unexpected shutdowns of the rotating mechanism. Early detection of bearing failures will reduce maintenance costs and ensure continuity in industrial and manufacturing processes. Reliable diagnosis of bearing failures is challenging due to the complexity and uncertainty inherent in vibration signals. Although numerous studies in this area exist in the literature, traditional approaches operate in a single domain, either time or frequency, failing to capture all structural information in the signal. Methods based on manual feature extraction rely on data quality for the overall performance of the model. The performance of these methods remains limited under diverse and complex operating conditions. Recently, deep learning-based approaches have enabled automatic feature extraction from vibration signals.

Using only a single data type in bearing fault diagnosis limits the features that can be learned. The success of deep learning models in fault diagnosis stems from the automatic data representation. However, it is still beneficial to have a diverse data type. A monotonous input attribute limits performance. Fault diagnosis becomes complex for bearings, which exhibit variable operating conditions such as load, speed, and noise parameters in industrial processes. While frequency-based problem solving can capture periodic conditions, it may be insufficient for detecting transient faults. Solving problems using only the time domain can be insufficient for capturing spectral information. Therefore, an integrated approach combining both domains can provide higher performance.

There are many studies in this area in the literature. They conducted an AI-assisted study on fault diagnosis in rotating machinery [1]. The statistical analysis and adaptive classification approach relies on feature selection and a distance-based decision mechanism instead of time and frequency features. Its performance was particularly stable in noisy measurements. This study, which relied on distance measurements between classes, showed good classification performance on short-term signal segments. A drawback of my study is the lack of automatic feature extraction. Another study presented an approach based on feature extraction and feature fusion [2]. They obtained a rich feature representation space by combining time-frequency features. Furthermore, by removing unnecessary features through feature selection, the model's performance increased, and accuracy reached 87.87%. Experimental studies were conducted on the CWRU open database. A disadvantage of the study is that the feature extraction process is manual. Vibration signals were fed directly to a convolutional neural network (CNN) and learned using this architecture to obtain time-frequency representations of the samples [3]. The results were examined by testing the network depth, pooling strategies, and filter sizes for performance. Experimental studies were conducted on the CWRU dataset. This review [4] examines studies developed to address vibration faults, and studies based on time, frequency, and time-frequency feature spaces were examined.

Image-based vibration fault diagnosis in the bearing structure of a wind turbine was investigated and two parallel CNN architectures were proposed [5]. One of these parallel paths is a CNN and the other is a transformer architecture. Their robustness to synthetic noise is also investigated by adding different noise levels. In this study, which conducted vibration fault diagnosis using the CWRU dataset, they designed a classic cascading CNN architecture based on deep learning [6]. The average accuracy achieved was 97.50%. They applied LSTM to time series dependencies using one-dimensional time series methods and two-dimensional transformed signals. In this study, time-frequency features were extracted from vibration signals using Discrete Wavelet Transform (DWT), Fast Fourier Transform (FFT) and Adaptive Neural Fuzzy Inference System (ANFIS) and the obtained features were automatically classified, and 97% accuracy

was achieved using the CWRU dataset [7]. They applied a fault diagnosis model and an online diagnosis model for rolling bearing fault diagnosis [8]. The C4.5 algorithm, support vector machine (SVM) algorithm, and CART and BP algorithms were tested on the CWRU dataset, and the highest accuracy was 93.75%. They proposed a deep learning-based approach for rolling bearing fault diagnosis on the CWRU dataset [9]. The extracted high-level features were combined with expert knowledge and fed to the classifier block. The accuracy rate obtained was 94.20%.

It was developed to capture patterns in acoustic signals using a Swin Transformer-based model; the comparative results of sound-based data and vibration data were also examined [10]. 98.6% accuracy was achieved in acoustic data. It was emphasized that this method is suitable for complex and multi-scale signals. Another study converted vibration data to images using the continuous wavelet transform (CWT) and aimed for feature extraction with ResNet50 [11]. The extracted features were then supported with SVM, and 95.51% accuracy was achieved with ResNet50. The [12] proposed hybrid architecture based on Locality-Constrained Mapping and Stacked Sparse AutoEncoder to support feature extraction in noisy environments and detect vibration faults. The proposed method was also tested and compared with SVM and K-nearest neighbor (KNN), and a performance increase of 0.6% was achieved. In a study that developed an attenuation-based model to reduce single-modality dependency, an accuracy of approximately 95% was achieved by testing it on datasets such as Paderborn and Ottawa [13]. Tab. 1 compares recent bearing fault diagnosis studies in the literature in terms of adopted methods, input data modalities, and diagnostic performance.

Table 1 Comparison of bearing fault diagnosis studies in the literature

Refs.	Method	Dataset	Acc / %
[2]	Time-frequency feature extraction, feature fusion	CWRU	87.87
[6]	Cascading CNN and LSTM for temporal features	CWRU	97.50
[7]	FFT, DWT and ANFIS	CWRU	97.00
[8]	C4.5, SVM, CART, BP comparison	CWRU	93.75
[9]	Deep CNN , time-frequency fusion, expert knowledge	CWRU	94.20
[11]	CWT, ResNet50 and SVM	CWRU	95.51
[13]	Multimodal attention model	Paderborn	95.00

In this study, an innovative multimodal and bidirectional cross-attention-based framework is proposed for automatic and reliable bearing fault diagnosis. By jointly utilizing raw vibration signals in the time domain and SWT-based time-frequency representations, the proposed approach captures both transient temporal characteristics and long-term spectral patterns. Unlike conventional unidirectional fusion strategies, the bidirectional cross-attention mechanism enables mutual interaction between modalities, preserving complementary information and improving feature discrimination. Experimental results on the SUBVF1.0 dataset demonstrate the robustness and stability of the proposed model under noisy and complex operating conditions. The main contributions of this study are as follows:

- A simple and effective multimodal fault diagnosis approach is proposed by jointly using raw time-domain vibration signals and SWT-based time–frequency representations, enabling more reliable fault characterization.
- A bidirectional cross-attention mechanism is introduced to fuse multimodal features bidirectionally, preserving complementary information and improving feature discrimination compared to conventional fusion methods.
- Experimental results on a public bearing dataset demonstrate that the proposed model achieves stable and high diagnostic performance under noisy and complex operating conditions.

2 THE PROPOSED APRROACH

This study presents a holistic multimodal fault diagnosis framework that integrates time-domain vibration signals and time-frequency representations through a bidirectional cross-attention mechanism. The overall architecture consists of two parallel feature extraction branches: a convolutional neural network for learning local temporal patterns from raw vibration signals, and an SWT-based branch for capturing frequency distribution information. The extracted features are fused using bidirectional cross-attention to enhance feature interaction and discrimination.

2.1 Dataset and Preprocessing

The dataset SUBF V1.0, created for bearing fault diagnosis, was used [14]. It consists of a 0.25 HP, 50 Hz frequency, 1440 RPM speed, and 440 V voltage, a shaft bearing, and a base frame. This experimental setup is visualized in Fig. 1. Vibration signals were collected with a BeanDevice AX-3D wireless sensor at a sampling rate of 1000 Hz. The dataset consists of three types of data: intact, inner ring, and outer ring. A total of 18 hours of measurements were segmented into 10-second intervals, resulting in 2160 samples per condition. The dataset was randomly divided into training (70%), validation (15%), and test (15%) sets to ensure fair evaluation and generalization assessment. The generalizability of the proposed model is also enhanced when the dataset contains measurements under different operating conditions. Pre-processing steps were applied to the dataset. First, the mean of each signal record was extracted and the mean was reset. Z-Score normalization was applied by dividing by the standard deviation. The signals are segmented into segments of a specific length. Each signal was segmented into fixed-length windows and converted into two representations: (i) raw time-domain signals and (ii) SWT-based timefrequency representations. The Daubechies (db4) wavelet was used, as it provides a good balance between time and frequency localization for vibration signals. Decomposition levels 2, 3, and 4 were selected because they effectively capture fault-related frequency bands while preserving sufficient temporal resolution. These selections were further supported by preliminary experiments, which showed that db4-based SWT features at levels 2-4 yielded higher classification accuracy compared to other wavelet families and deeper decomposition levels. This configuration also provides a favorable trade-off between diagnostic performance and

computational efficiency, making it suitable for practical industrial applications.

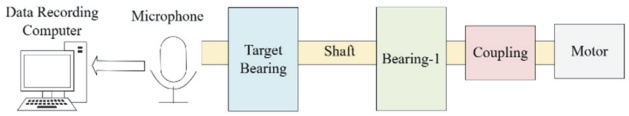


Figure 1 Block diagram of the in-house SUBF v1.0 data acquisition setup [14]

2.2 Proposed Multi-modal BCA-Based Model

In this study, a deep learning-based, multi-layered, multi-modal, and attention-based multi-feature fusion architecture capable of capturing information in both the time and time-frequency domains is presented for bearing fault diagnosis in as Fig. 2. The model processes data in the time and time-frequency domains, providing high discrimination in bearing fault diagnosis. This proposed multimodal architecture reduces information loss optimizing the attention mechanism by processing two information sources in parallel. Furthermore, this structure integrates information flows and reinforces each other through cross-attention blocks. The vibration signals in the dataset used are segmented and normalized to eliminate statistical imbalance, and training stability is increased by eliminating scale differences. The proposed model has two distinct and parallel processing pipelines. One is for processing raw vibration signals, and directly processes time-domain signals, and the other is the SWT pipeline, which expresses the time-frequency structure of the signals. Vibration signals in the raw data stream format are processed with a 1D convolution layer, and the resulting features are fed to the temporal attention block. The temporal attention block provides a short-term fault representation in the vibration signal. In the SWT pipeline, vibration signals are converted to 32 × 32 spectral tensors to prevent distortion of frequency features. These tensors are processed through CNN layers and directed to the channel-based attention mechanism. Information from the two streams was transferred to self-attention, and the relationships between time and frequency points were learned using a multi-attention mechanism. Information-rich regions were strengthened by modeling the internal relationships between the features. The modal alignment block then aligns the vector space of the two representations and aligns them with the joint representation of the two vectors. The bidirectional cross-attention, using raw and SWT bidirectional attention calculations, strengthens the common information components in both feature spaces. The cross-attention structure offers a more interactive context than classical fusion. Following this block, the representation space is passed to the classifier block for classification.

In the proposed study, a multimodal structure and cross-attention mechanism are used to query the feature representation in one modality, while the features in the other modality are associated with the key and value [15-17]. This mechanism is crucial because it allows us to learn the relationship between the raw vibration signals and the SWT spectral representations and minimize information loss. The raw modality query Q^r and the key K^s and value V^s of the SWT representation are expressed in Eq. (1), Eq. (2), and Eq. (3).

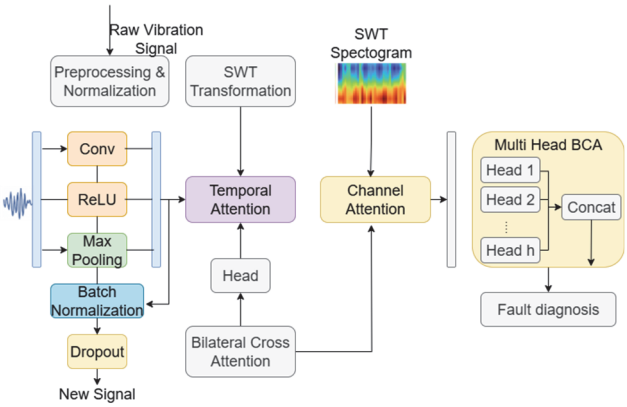


Figure 2 The overview of the proposed method

$$Q^r = X^r W_Q \tag{1}$$

$$K^s = X^s W_K \tag{2}$$

$$V^s = X^s W_V \tag{3}$$

$$\text{Attention}\left(Q^r, K^s, V^s\right) = \text{softmax}\left(\frac{Q^r\left(K^s\right)^T}{\sqrt{d_k}}\right)V^s \tag{4}$$

In Eq. (4), X^r is the feature matrix obtained from the raw vibration signals, and X^s is the feature matrix of the representation obtained by the SWT transformation. The W_Q represents the learned weight matrices, while d_k represents the key vector dimension. With two-way cross-attention, the mutual attention values of both modalities are calculated as in Eq. (5). This synchronizes the representation spaces of both modalities [18, 19]. Thus, the interaction of information in both time and frequency domains is optimized.

$$\begin{aligned}\text{BCA}\left(X^r, X^s\right) &= \hat{V}^r\left(\text{Attention}\left(Q^r, K^s, V^s\right)\right) + \\ &+\hat{V}^s\left(\text{Attention}\left(Q^r, K^s, V^s\right)\right)\end{aligned} \tag{5}$$

The features obtained with bidirectional cross-attention are fused. Features from different modalities, which have been calculated based on attention, are combined. The $\hat{V}^r$ and $\hat{V}^s$, the features combined based on attention, are the raw and SWT representations updated with BCA. The relative importance of each model is learned and weighted using α^r and α^s as in Eq. (6) and Eq. (7). The more informative modality is determined, and the features are combined effectively. The learned weights are also taken into account, and the feature fusion is performed as in Eq. (8).

$$\alpha^r = \frac{\exp\left(f\left(\hat{V}^r\right)\right)}{\exp\left(f\left(\hat{V}^r\right)\right) + \exp\left(f\left(\hat{V}^s\right)\right)} \tag{6}$$

$$\alpha^s = \frac{\exp\left(f\left(\hat{V}^s\right)\right)}{\exp\left(f\left(\hat{V}^r\right)\right)+\exp\left(f\left(\hat{V}^s\right)\right)} \tag{7}$$

$$X_{\text{fusion}} = \alpha^r \hat{V}^r + \alpha^s \hat{V}^s \tag{8}$$

As a result, this bidirectional cross attention structure integrates both time and frequency information. In addition to preventing information loss, the model can achieve higher performance by feeding it with different domains.

2.3 Raw Signal Extraction

A feature extraction process was performed over the time domain to identify anomalies related to fault conditions by capturing sudden changes, amplitude fluctuations, and dynamic behaviors from the raw vibration signal. High-resolution characteristics in the raw vibration signal were examined. A 1D convolutional neural network was used to extract features from the signal. Let $x(t)=\left[x_1, x_2, x_3, ..., x_N\right], x_i \in \mathbb{R}$ be the raw vibration signal in the time domain, f_s be the sampling frequency of the signal, and its length be N. As a preprocessing step, the raw vibration signal is first divided into fixed-length segments and normalized as in Eq. (9). In Eq. (9), μ_x represents the mean of the signal, and σ_x represents the standard deviation. The normalization process eliminates scale differences in the signal, ensuring training stability.

$$x_{\text{norm}}(t) = \frac{x(t)-\mu_x}{\sigma_x} \tag{9}$$

The $x_{\text{norm}}(t)$ obtained in Eq. (9) is applied with a 1D convolution layer Eq. (10). This captures local patterns. In Eq. (10), $w_m^{(k)}$ is the k-th filter weight, $b^{(k)}$ is the bias, and f(.) is the ReLU activation function. The resulting y_i^k represents short-term energy changes in the signal. Bearing faults generally appear as periodic pulses, so 1D convolution detects these pulses in the time domain. The extracted feature maps are then passed to the temporal attention block, which determines which time steps are more important for fault detection. The qttt weight value is learned for each time step using Eq. (11) and Eq. (12). In Eq. (12), httt is the feature vector at step t. The reweighted feature representation with attention is calculated in Eq. (13). This ensures that the important time regions in the signal are emphasized while the weaker regions are suppressed.

$$y_i^k = f\left(\sum_{m=1}^{M} w_m^{(k)} . x_{i+m-1} + b^{(k)}\right) \tag{10}$$

$$e_t = v^T \tanh\left(W_h h_t + b_h\right) \tag{11}$$

$$\alpha_t = \frac{\exp\left(e_t\right)}{\sum_{i=1}^{T}\exp\left(e_i\right)} \tag{12}$$

$$H' = \sum_{t=1}^{T} \alpha_t h_t \tag{13}$$

2.4 SWT Feature Extraction

By preventing the distortion of frequency components in vibration signals, SWT performs time-frequency based information processing. Compared to the classical Discrete Wavelet Transform (DWT), the sensitivity of the signal to translational changes is increased, and thus SWT is effective in vibration-based fault diagnosis. In the DWT process, the signal is sampled for each level, which is highly susceptible to data loss. In SWT, instead of resampling for each level, the filter coefficients are expanded in the time axis [20]. Thus, details of both high- and low-level components are obtained at each level [21, 22]. While $A_j[n]$ in Eq. (14) are the approximate coefficients at level j, $D_j[n]$ in Eq. (15) correspond to the detail coefficients. The $h^j[k]$ in Eq. (16) is the scaled version of the low-pass filter at level j, $g^j[k]$ and in Eq. (17) is the scaled version of the high-pass filter at level j. It is the scaled version of the level. Scaling of these filters is done by expanding the coefficients in SWT. In Eq. (16) and Eq. (17), the scale is enlarged by adding $2^{j-1}-1$ zeros between the filter coefficients.

$$A_j[n] = \sum_k h^j[k] . A_{j-1}[n-k] \tag{14}$$

$$D_j[n] = \sum_k g^j[k] . A_{j-1}[n-k] \tag{15}$$

$$h^j[k] = h[k] \uparrow 2^{j-1} \tag{16}$$

$$g^j[k] = g[k] \uparrow 2^{j-1} \tag{17}$$

2.5 Bidirectional Cross Attention Fusion

Bidirectional cross-attention is applied to meaningfully combine features obtained from raw vibration signals with features obtained through the SWT transformation. Unlike traditional fusion structures, a representation is learned that learns the interrelationships between raw and SWT modes. Information is exchanged between instantaneous changes in the time domain and frequency changes in vibrations. By enabling bidirectional information interaction between the two modalities, it optimizes the attention allocation of each modality relative to the other. It is an attention mechanism that determines which frequency regions are associated with a specified feature vector in the time domain and which time samples a feature map in the frequency domain uses for learning. Let the feature vectors for the two modalities be $X_r \in \mathbb{R}^{T_r \cdot d_r}$ and $X_s \in \mathbb{R}^{T_s \cdot d_s}$. The X_r are features from the raw vibration signals, and X_s are SWT features. Tr and T_s are the number of time samples, while d_r and d_s represent the feature dimensions. BCA consists of two distinct cross-attention blocks. This bidirectional structure allows both modalities to exchange information with each other. When applying cross-attention from the raw modality to the SWT modality, the raw features are defined as query

$Q_r = X_r W_r^Q$, swt attributes are defined as key $K_s = X_s W_s^K$, and value $V_s = X_s W_s^V$. Here, $W_r^Q \in \mathbb{R}^{d_r \cdot d_k}$, W_s^K , and W_s^V are the learnable transformation matrices. Attention scores for each time sample are calculated in Eq. (20). The model learns which frequency regions in the SWT domain are associated with the temporal changes in the raw signal. Information from the SWT space is weighted with attention scores and directed to the raw vibration space, resulting in an enriched version of the feature representation as in Eq. (21).

$$A_{r\to s} = \text{softmax}\left(\frac{Q_r K_s^T}{\sqrt{d_k}}\right) \tag{20}$$

$$Z_{r\to s} = A_{r\to s} V_s \tag{21}$$

The SWT modality is the query function, which is symmetrical to the block where features are extracted from raw vibration signals. The key and value values are obtained from the raw vibration signal modality. Let query, key, value be $Q_s = X_s W_s^Q$, $K_r = X_r W_r^K$, and $V_r = X_r W_r^V$, respectively. Attention scores are calculated in Eq. (22). The information extracted from the signal is transferred to the SWT space and calculated using Eq. (23). The SWT features are formatted to capture patterns in the time domain.

$$A_{s\to r} = \text{softmax}\left(\frac{Q_s K_r^T}{\sqrt{d_k}}\right) \tag{22}$$

$$Z_{s\to r} = A_{s\to r} V_r \tag{23}$$

The representations $Z_{r\to s}$ and $Z_{s\to r}$ obtained as a result of bidirectional attention are transmitted to the modal alignment block. In this block, the dimensions of $Z_r^{'}$ and $Z_s^{'}$ are first aligned with Eq. (24) and Eq. (25). Qrr and Qsss in these equations represent the 1D convolutional dimension matching. The representations obtained from Eq. (24) and Eq. (25) are combined with bidirectional fusion, as in Eq. (26). The γ_1, γ_2 and γ_3 in Eq. (26) are the learnable weight coefficients, and $\odot$ is the point-based multiplication operation. The most important element to note here is the bidirectional information reinforcement. While information interacts unidirectionally in classical attention, in the BCA block, information can flow in both directions. Common information components become more meaningful, and noisy components are neutralized. To optimize the representational similarity between the two modalities, the bidirectional consistency loss (BCL) in Eq. (27) is calculated.

$$Z_r^{'} = \phi_r\left(Z_{r\to s}\right) \tag{24}$$

$$Z_s^{'} = \phi_s\left(Z_{s\to r}\right) \tag{25}$$

$$Z_{\text{fusion}} = \gamma_1 Z_r^{'} + \gamma_2 Z_s^{'} + \gamma_3\left(Z_r^{'} \odot Z_s^{'}\right) \tag{26}$$

$$\mathcal{L}_{\text{BCL}} = Z_{r\to s} - Z_{s\to r}{}^2 \tag{27}$$

2.6 Classification and Training Details

The main objective of the proposed model is to combine time and frequency modes obtained from bearing vibration signals to perform high-performance fault classification. In this architecture, the representations obtained from the time mode were learned using 1D convolution blocks, while the representations obtained from the frequency mode were learned using 2D convolution blocks. This enabled the model to extract features from different information levels. The initial representations of the features in these two modes were fed to the feature fusion block. This block provides bidirectional attention representation for both channel and spatial contexts, thus increasing the model's representational power. The final layers of the proposed network include global average pooling, dense, and softmax layers. Categorical cross anthropy was used as the loss function, and the Adam optimizer was used for weight updates. A dropout layer was used during the training process to prevent overfitting. The batch size was set to 64 and the epoch was set to 20. The training process was performed on an NVIDIA RTX 4090 GPU using Python 3.12 and TensorFlow 2.16. The number of learnable parameters for the model was calculated as 310.403, indicating that the model has a balanced architecture in terms of depth and breadth.

3 EXPERIMENTAL RESULTS

The performance of this study, which utilizes deep learning and attention-based vibration signals for fault diagnosis, was examined in detail. During the experimental studies, training and testing were conducted on the SUBVF1.0 dataset, which contains three types of vibration defects. A steady decrease in the loss function was observed during training. This demonstrates that the model is robust to overfitting and that the learning process is optimized. To evaluate the robustness of the proposed model under noisy conditions, Gaussian white noise with different signal-to-noise ratio (SNR) levels was artificially added to the vibration signals during testing. The classification performance was then assessed at each SNR level, demonstrating that the proposed model maintains stable accuracy under increasing noise levels.

3.1 Evaluation Metrics

To measure the classification success and generalization ability of the proposed model, the accuracy, precision, recall, and f1-score metrics were evaluated during training and validation. Accuracy is the overall performance of the model and is calculated by Eq. (28). To indicate the degree of selectivity of the model and to minimize the false alarm rate, it is calculated by Eq. (29). The recall metric indicates the sensitivity of the model in detecting errors and the risk of classification as false

negatives, as in Eq. (30). The F1-Score metric is the measure of agreement between the precision and recall metrics and is calculated by Eq. (31). In the calculation of these metrics, TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative. As shown in Fig. 3, the proposed model exhibits stable training behavior and achieves high performance across the evaluated metrics.

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{28}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{29}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{29}$$

$$\text{F1-Score}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{31}$$

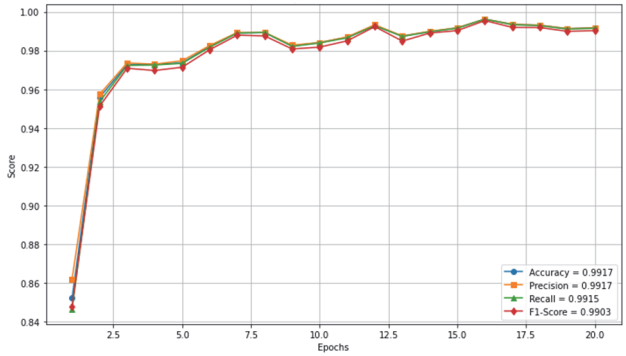


Figure 3 Performance metrics of the proposed model in training

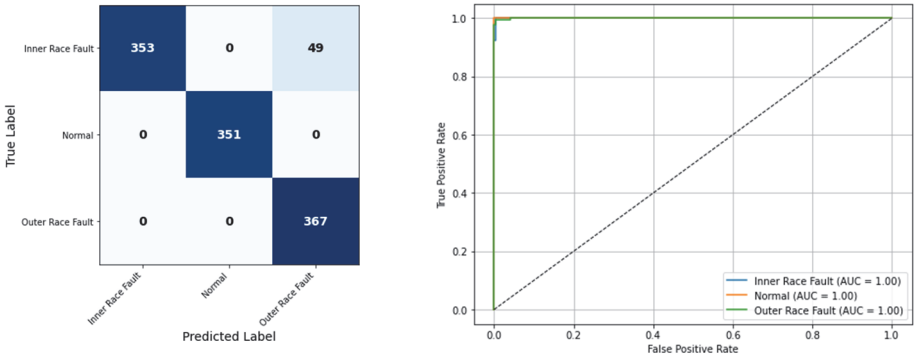
Tab. 2 shows the comparison of training and validation performance metrics. It is observed that the model achieved 99.17% training accuracy and 95.62% validation accuracy after 20 epochs.

Table 2 Comparison of training and validation performance metrics of the proposed model

Metrics	Training	Validation
Accuracy	0.9917	0.9562
Precision	0.9918	0.9563
Recall	0.9915	0.9563
F1-Score	0.9902	0.9554
Loss	0.0316	0.1435

3.2 Visualization of Results

Some images are presented to better visualize the decision-making process of the proposed model. The confusion matrix is shown in Fig. 4a to show the correct prediction rates for each class in the proposed model. Examining these results, 353 samples belonging to the inner race fault class were correctly classified, while 49 were incorrectly labeled as outer race faults. All 351 samples belonging to the normal class were correctly classified and labeled as normal. Similarly, all 367 outer race faults were correctly labeled and assigned to the outer race class. An examination of the ROC curve in Fig. 4b reveals an AUC value of 1.00 for all classes, demonstrating the model's high discriminatory power. The median distribution graphs in Fig. 5 are presented using box plots. These visualizations reflect the average feature behavior in each class. It can be seen that the distribution within each class type has statistical separation. Finally, the Uniform Manifold Approximation and Projection (UMAP) visualization in Fig. 6 demonstrates how the model distinguishes and classifies each data point from the original data points and the model's prediction result.



(a) **(b)**

Figure 4 (a) Confusion matrix and (b) ROC curve of the proposed model

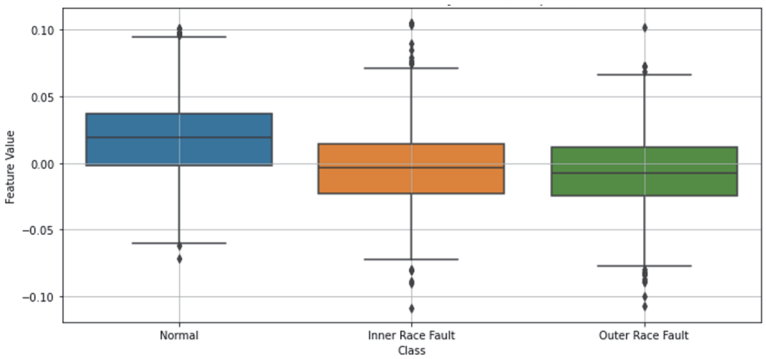
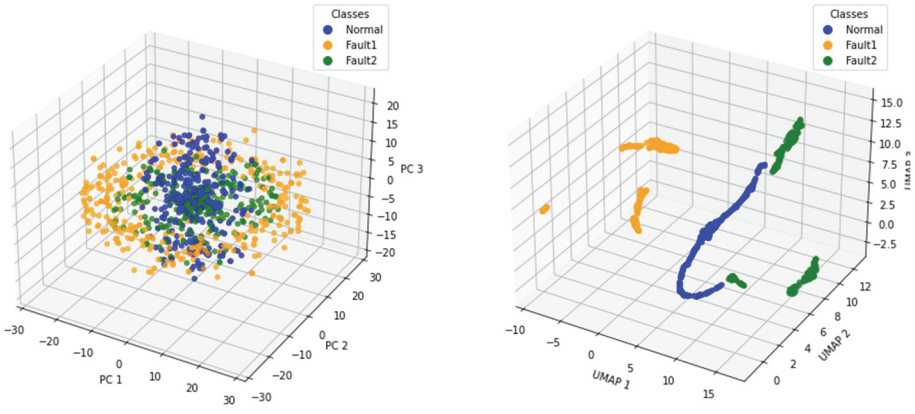


Figure 5 Distribution of median by class of the proposed model



(a) (b)

Figure 6 UMAP visualization (a) input space and (b) learned feature space of the proposed model

3.3 Ablation Studies

The proposed model consists of multiple components. An ablation study was conducted to quantify the contribution of these components to the baseline model. Therefore, certain layers were removed from the model, and the model was retrained. The resulting performance results are presented comparatively. In the proposed model, the representations obtained from the raw vibration mode and the SWT mode were combined and directed to the BCA block. First, the raw vibration mode was removed from the proposed model, and classification was performed using only the representation obtained from the SWT mode. This is labeled Ablation-1. In contrast, Ablation-2 used only the representations obtained from the raw vibration mode, and the representations obtained from the SWT mode were removed. In Ablation-3, classification was performed using the raw vibration mode representations and the SWT mode representations, and the BCA block was replaced with the classical attention block. Fig. 7, Fig. 8, Fig. 9 and Tab. 3 present the comparison of the ablation studies with the proposed model across all metrics in training and validation. Ablation-1, there was a performance loss of approximately 3% compared to the proposed model during the training phase, indicating that the contribution of the raw vibration signal was quite significant. In Ablation-2, there was a performance loss of 1% compared to the proposed model during the training phase. This decrease was smaller compared to Ablation-1. In Ablation-3, using classical attention methods in the training model again led to a performance loss of approximately 2%. This shows that BCA improves performance by preventing information loss in the model. In Ablation-1, there was a performance loss of approximately 2% compared to the proposed model during the validation phase. In Ablation-2, there was a performance loss of 1% compared to the proposed model during the validation phase. This decrease was smaller compared to Ablation-1, indicating that the raw vibration signal carries more important information about the model compared to SWT. In Ablation-3, using classical attention methods in the validation model again resulted in a performance loss of approximately 12%. This further decrease in performance compared to training, and this significant decrease in the model's generalization ability, demonstrates the main contribution of the BCA model. Fig. 10 shows comparison of classification accuracy between the proposed method and existing studies. The results clearly demonstrate that the proposed method achieves the highest accuracy, outperforming previous approaches reported in the literature.

Table 3 Comparison of performance metrics between the proposed model and its ablated variants

Metrics	Ablation-1		Ablation-2		Ablation-3		Proposed	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Accuracy	0.9647	0.9312	0.9812	0.9473	0.9734	0.8366	0.9917	0.9562
Precision	0.9649	0.9312	0.9823	0.9473	0.9745	0.8380	0.9918	0.9563
Recall	0.9645	0.9304	0.9801	0.9473	0.9714	0.8357	0.9915	0.9563
F1-Score	0.9625	0.9310	0.9790	0.9475	0.9713	0.8513	0.9902	0.9554

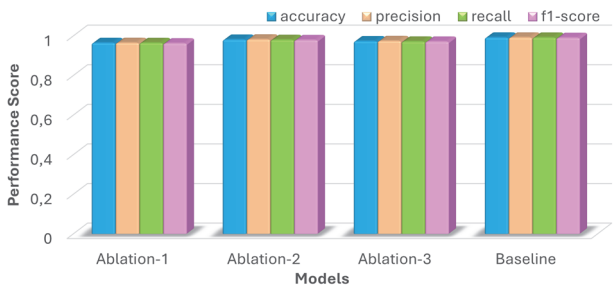


Figure 7 Ablation study results illustrating the effect of removing each component on the overall performance metrics of training

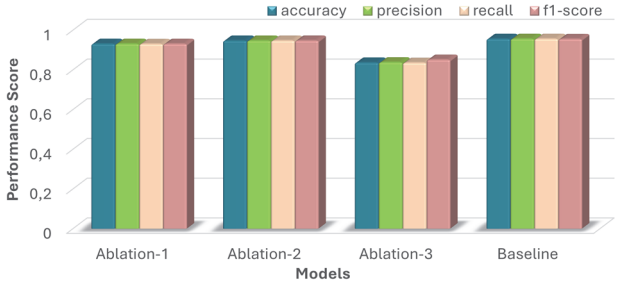


Figure 8 Ablation study results illustrating the effect of removing each component on the overall performance metrics of validaiton

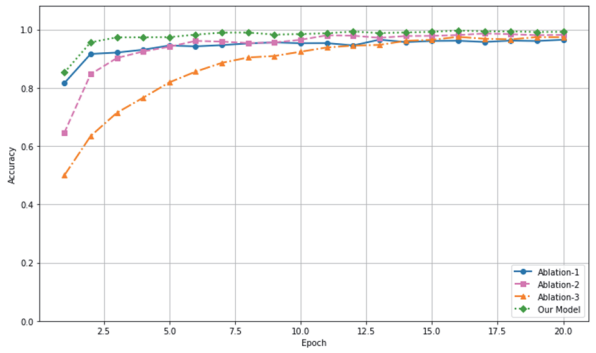


Figure 9 Training accuracy values obtained from the proposed model and different ablation configurations

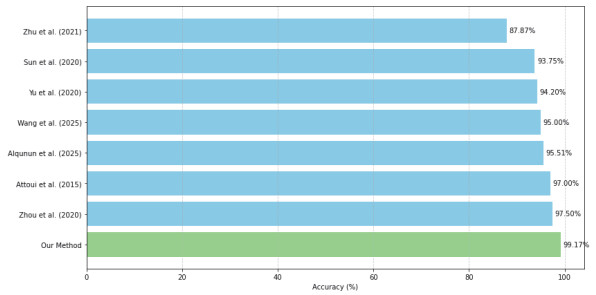


Figure 10 Comparison of classification accuracy between the proposed method and existing approaches reported in the literature

4 DISCUSSION

The proposed multimodal framework achieves high diagnostic performance by effectively fusing raw time-domain and SWT-based time-frequency features through bidirectional cross-attention. This fusion strategy enhances feature discrimination and class separability, as confirmed by both performance metrics and ablation results. In addition, the model demonstrates strong robustness under noisy conditions and maintains stable performance, making it suitable for real-world industrial applications. Future work will focus on extending the approach to more diverse operating conditions and embedded system deployments.

5 CONCLUSIONS AND RECOMMENDATIONS

In this study, a multimodal bearing fault diagnosis framework based on bidirectional cross-attention was proposed. By integrating SWT-based time-frequency features with raw vibration signal representations, the model effectively captures both spectral and temporal characteristics of bearing faults. The bidirectional cross-attention mechanism enhances feature interaction and class discrimination. Experimental results on the SUBVF1.0 dataset demonstrate strong classification performance, achieving accuracy, precision, sensitivity, and F1-score values of 99.17%, 99.18%, 99.15%, and 99.02%, respectively. The model also exhibits high generalization capability and robustness. Future work will focus on extending the framework to different operating speeds, load conditions, and sensor types, as well as deploying the model in real-time and embedded systems. Despite its strong performance, the proposed model has certain limitations. The current study focuses on a single dataset and a limited number of fault types. Future research should investigate the generalizability of the approach across a wider range of operating conditions, fault severities, and sensor configurations to further validate its robustness and scalability.

Acknowledgement

This work was supported by the Scientific Research Projects Coordination Unit of Erzincan Binali Yildirim University under Grant No: FHD-2025-1108.

6 REFERENCES

[1] Lei, Y., He, Z., & Zi, Y. (2008). A new approach to intelligent fault diagnosis of rotating machinery. *Expert Systems with Applications*, *35*(4), 1593-1600. https://doi.org/10.1016/j.eswa.2007.08.072

[2] Zhu, H., He, Z., Wei, J., Wang, J., & Zhou, H. (2021). Bearing fault feature extraction and fault diagnosis method based on feature fusion. *Sensors*, *21*(7), 2524. https://doi.org/10.3390/s21072524

[3] Hoang, D. T. & Kang, H. J. (2019). Rolling element bearing fault diagnosis using convolutional neural network and vibration image. *Cognitive Systems Research*, *53*, 42-50. https://doi.org/10.1016/j.cogsys.2018.03.002

[4] Althubaiti, A., Elasha, F., & Teixeira, J. A. (2021). Fault diagnosis and health management of bearings in rotating equipment based on vibration analysis - a review. *Journal of Vibroengineering*, *24*(1), 46-74. https://doi.org/10.21595/jve.2021.22100

[5] Chen, Q., Zhang, F., Wang, Y., Yu, Q., Lang, G., & Zeng, L. (2025). Bearing fault diagnosis based on efficient cross space multiscale CNN transformer parallelism. *Scientific Reports*, *15*(1), 12344. https://doi.org/10.1038/s41598-025-95895-x

[6] Zhou, W. & Hou, Y. (2020). A Deep Learning Method for Rolling Bearing Fault Diagnosis through Heterogeneous Data. *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*, 1214-1219). https://doi.org/10.1109/DDCLS49620.2020.9275189

[7] Attoui, I., Boutasseta, N., Fergani, N., Oudjani, B., & Deliou, A. (2015). Vibration-based bearing fault diagnosis by an integrated DWT-FFT approach and an adaptive neuro-fuzzy inference system. *2015 3rd International Conference on Control, Engineering & Information Technology (CEIT)*, 1-6. https://doi.org/10.1109/CEIT.2015.7233098

[8] Sun, J., Yu, Z., & Wang, H. (2020). On-line fault diagnosis of rolling bearing based on machine learning algorithm. *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, 402-407. https://doi.org/10.1109/ISCTT51595.2020.00075

[9] Yu, S., Li, X., Yang, B., & Lei, Y. (2022). A Hybrid Intelligent Method for Rolling Bearing Fault Diagnosis Integrated with Expert Knowledge and Deep Learning. *2022 4th International Conference on Industrial Artificial Intelligence (IAI)*, 1-5. https://doi.org/10.1109/IAI55780.2022.9976758

[10] Jeong, H., Kim, S., Seo, D., & Kwon, J. (2025). Source-free domain adaptation framework for rotary machine fault diagnosis. *Sensors*, *25*(14), 4383. https://doi.org/10.3390/s25144383

[11] Alqunun, K., Bechiri, M. B., Naoui, M., Khechekhouche, A., Marouani, I., Guesmi, T., Alshammari, B. M., AlGhadhban, A., & Allal, A. (2025). An efficient bearing fault detection strategy based on a hybrid machine learning technique. *Scientific Reports*, *15*(1), 18739. https://doi.org/10.1038/s41598-025-02439-4

[12] Deng, Z., Li, Y., Zhu, H., Huang, K., Tang, Z., & Wang, Z. (2020). Sparse stacked autoencoder network for complex system monitoring with industrial applications. *Chaos, Solitons & Fractals*, *137*, 109838.

https://doi.org/10.1016/j.chaos.2020.109838
[13] Wang, Y., Wang, H., Bai, R., Shi, Y., Chen, X., & Xu, Q. (2025). Enhanced rolling bearing fault diagnosis using multimodal deep learning and singular spectrum analysis. *Applied Sciences*, *15*(9), 4828. https://doi.org/10.3390/app15094828
[14] Aziz, S., Khan, M. U., Faraz, M., & Montes, G. A. (2023). Intelligent bearing faults diagnosis featuring automated relative energy-based empirical mode decomposition and novel cepstral autoregressive features. *Measurement*, *216*, 112871. https://doi.org/10.1016/j.measurement.2023.112871
[15] Song, J. & Liu, Y. (2025). Cross-Domain Multiscale Feature Fusion with Multi-Source Data for Enhanced Motor Fault Diagnosis. *2025 37th Chinese Control and Decision Conference (CCDC)*, 4326-4332. https://doi.org/10.1109/CCDC65474.2025.11090807
[16] Luo, J. & Chen, X. (2024). An Improved NGO Algorithm for Optimizing VMD-CNN-BILSTM-Cross Attention Bearing Fault Diagnosis Method. *2024 IEEE International Conference on Cognitive Computing and Complex Data (ICCD)*, 341-346. https://doi.org/10.1109/ICCD62811.2024.10843485
[17] Jang, G. B., Kim, J. Y., & Cho, S. B. (2021). Cross-domain fault diagnosis of rotating machinery using discriminative feature attention network. *IEEE Access*, *9*, 99781-99793. https://doi.org/10.1109/ACCESS.2021.3096145
[18] Zuo, D. & Wang, X. (2024). Fault Diagnosis Method Based on Multi-Attention Convolutional Neural Network. *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)*, 1-3. https://doi.org/10.1109/CSRSWTC64338.2024.10811656
[19] Tao, T., Jia, X. F., Su, D. F., Zhang, X. L., Gao, J. F., Xu, R. Z., Mo, L., & Wang, X. J. (2024). Fault Diagnosis of Rolling Bearings Based on Cross Attention Network with Multi-Scale Feature Fusion. *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)*, 1-3. https://doi.org/10.1109/CSRSWTC64338.2024.10811569
[20] Chen, J., Lin, C., Lu, Q., Yang, C., Li, P., Yu, P., & Ge, H. (2024). A decision fusion SWT-RF method for rolling bearing enhanced diagnosis under low-quality data. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1-11. https://doi.org/10.1109/TIM.2024.3350130
[21] Wu, W., Jing, S., Leng, J., Luo, C., & Hua, W. (2022). Gearbox Fault Feature Extraction Method Based on SWT and Fast Kurtogram. *2022 16th Symposium on Piezoelectricity, Acoustic Waves, and Device Applications (SPAWDA)*, 296-300. https://doi.org/10.1109/SPAWDA56268.2022.10045855
[22] Kompella, K. D., Rao, M. V. G., & Rao, R. S. (2017). SWT based bearing fault detection using frequency spectral subtraction of stator current with and without an adaptive filter. *TENCON 2017-2017 IEEE Region 10 Conference*, 2472-2477. https://doi.org/10.1109/TENCON.2017.8228277

Contact information:

Canan TASTIMUR, Asst. Prof. Dr.
Erzincan Binali Yildirim University,
Department of Computer Engineering,
Erzincan, 24100, Türkiye
E-mail: ctastimur@erzincan.edu.tr

Volkan KAYA, Assoc. Prof. Dr.
(Corresponding author)
Erzincan Binali Yildirim University,
Department of Computer Engineering,
Erzincan, 24100, Türkiye
E-mail: vkaya@erzincan.edu.tr