machines

MDPI

Article

Transformer-Based Bearing Fault Classification with VMD-Based Noise Suppression and rCCA-Enhanced Correlation Modeling

Tarkan Koca [1,*], Mehmet Bilal Er [2] and Aydın Çıtlak [3]

[1] Department of Mechanical Engineering, Inonu University, 44000 Malatya, Türkiye
[2] Department of Software Engineering, Harran University, 63050 Şanlıurfa, Türkiye; bilal.er@harran.edu.tr
[3] Department of Mechanical Engineering, Fırat University, 23119 Elazığ, Türkiye; acitlak@firat.edu.tr
* Correspondence: tarkan.koca@inonu.edu.tr

Abstract

Early detection of bearing faults in rotating machinery is essential for ensuring system reliability and effective maintenance planning. Vibration signals inherently contain characteristic fault-related frequency components, providing rich information for both physically interpretable and data-driven analyses. In this study, a multi-representation and correlation-aware feature extraction framework is proposed for automatic classification of bearing faults from vibration signals. Experimental evaluations are conducted using the Case Western Reserve University (CWRU) Bearing Dataset. The dataset includes vibration recordings corresponding to inner race, outer race, ball faults, and healthy conditions under different damage severities. The proposed approach first applies Variational Mode Decomposition (VMD) to suppress noise and enhance frequency-related characteristics. Three different feature representations are then constructed: analytical spectral descriptors, raw Transformer-based deep representations, and a hybrid feature vector obtained by combining these two representations. The hybrid structure is further enhanced through regularized Canonical Correlation Analysis (rCCA), which models the relationship between Transformer representations and spectral descriptors, enabling correlation-aware feature fusion. Spectral, raw Transformer, and rCCA-enhanced hybrid feature vectors are evaluated separately using SVM, Random Forest, and XGBoost classifiers. The results demonstrate that both spectral and Transformer-based representations provide strong performance individually; however, integrating these complementary information sources while modeling their correlations leads to superior and more balanced classification performance. In particular, the rCCA-enhanced hybrid feature vector achieves the best results across all performance metrics. The findings indicate that combining physically meaningful frequency-domain information with data-driven deep representations yields a more robust and generalizable solution for bearing fault diagnosis.

Keywords: bearing fault diagnosis; VMD; transformer; hybrid feature vector; regularized canonical correlation analysis

Check for updates

Academic Editor: Gómez María Jesús

Received: 11 March 2026
Revised: 16 April 2026
Accepted: 20 April 2026
Published: 1 May 2026

Copyright: © 2026 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.

1. Introduction

Rotary machinery is widely used in many critical industrial areas such as automotive, aerospace, energy production, agriculture, and manufacturing industries. These machines consist of multiple mechanical components such as shafts, gears, couplings, and bearings, and the reliability of these components directly affects the overall performance of the system. Bearings, in particular, are among the most critical elements of rotary machinery due to their

role in supporting rotational motion and carrying dynamic loads. Industrial analyses show that a significant portion of failures in rotary machinery are bearing-related [1]. Sudden bearing failures can lead to production losses, increased maintenance and repair costs, and in some cases, serious safety risks. Bearing failures develop over time due to the influence of many factors such as faulty design, incorrect assembly, overloading, insufficient lubrication, wear, fatigue, and harsh operating environments. These failures usually occur in the inner ring, outer ring, or rolling elements and, if not detected early, can cause the machine to stop unexpectedly. This situation is not limited to economic losses but also poses serious threats to human safety in critical industrial applications. Therefore, continuous monitoring of bearing operating conditions and early detection of failures are of great importance for predictive maintenance and reliability engineering [2]. Vibration signal analysis is one of the most effective and widely used non-destructive methods for monitoring the condition of rotating machinery. Vibration signals contain rich information about the operating conditions of the machine and the health status of the mechanical components. In studies, numerous conventional signal processing methods based on time domain, frequency domain, and time-frequency domain have been developed for the diagnosis of bearing failures [3]. While time domain methods generally rely on statistical properties, frequency domain analyses make it possible to examine spectral components. Time-frequency-based methods can simultaneously address both the temporal and frequency content of the signal. However, conventional signal processing methods are mostly based on linear and stationary system assumptions. Vibration signals obtained in real industrial environments generally exhibit noisy, non-stationary, and nonlinear characteristics [4]. Many types of failures in rotating machinery are directly related to the nonlinear dynamic behavior of the system. In nonlinear systems, responses to harmonic excitations are not limited only to fundamental frequency components; lower harmonics, upper harmonics, quasi-periodic motions, and chaotic behaviors can also occur. This situation causes classical spectral analysis methods to fail to adequately capture early fault indications. In this context, nonlinear dynamics and chaos theory-based analysis approaches allow for a more in-depth examination of vibration signals. Features obtained from chaotic field (CD) analysis can quantitatively express the complexity, stability, and dynamic evolution of the system and offer significant advantages in early fault detection [5]. Chaos-based analyses contribute to the determination of fault characteristics that are difficult to detect with traditional methods by revealing irregular behaviors in time series. Numerous features can be extracted from vibration signals in time, frequency, time-frequency, and chaotic domains. However, this brings with it the problem of high-dimensional data. Inclusion of unnecessary or low-discrimination features in the model increases computational costs, prolongs the training time of algorithms, and negatively affects classification performance [6]. Therefore, selecting features with high discriminability and applying dimensionality reduction strategies in fault diagnosis is of great importance. Effective feature selection increases the accuracy, stability, and generalization ability of the model while minimizing information loss. In recent years, with the increase in computing capacity and the accessibility of large-scale condition monitoring data, machine learning and deep learning-based approaches have become widely used in the field of bearing fault diagnosis and remaining service life estimation [7]. These methods can learn meaningful patterns from complex and nonlinear data structures and provide higher accuracy compared to traditional methods. However, the success of these models largely depends on the quality and representational power of the input features used. In research conducted in the field of bearing fault diagnosis, the PRONOSTIA, CWRU, IMS, and Paderborn datasets are widely referenced. These datasets are used under different load, speed, and operating conditions. The CWRU dataset provides vibration signals collected at different load levels and sampling frequencies for induction motors, while

the Paderborn dataset allows for multi-sensor analyses with synchronized vibration and current data of healthy and failed bearings. The IMS dataset covers the temporal evolution of the bearing failure process through accelerated life tests, while the PRONOSTIA dataset represents the failure dynamics under different operating scenarios [8]. However, the vast majority of these datasets focus on a specific bearing type and a limited number of failure cases. In real industrial environments, bearings operate under variable speeds, complex load profiles, and multiple failure combinations. This severely limits the generalization ability of models trained on existing datasets and leads to performance loss in real field applications. Developed to address this deficiency, the HUST dataset offers a more realistic evaluation environment by covering different bearing types, integrated failure scenarios, and variable operating conditions. In this respect, HUST is an important data source that allows the analysis of scenarios close to industrial conditions [9]. Bearing failure diagnosis methods are generally classified as signal-based, model-based, knowledge-based, and hybrid approaches [10]. Signal-based methods rely directly on vibration or current signals, while model-based approaches are based on physical deterioration models. Knowledge-based methods use inferences based on expert knowledge, while hybrid approaches aim to combine the strengths of these methods. In early studies, classical machine learning algorithms such as support vector machines, principal component analysis, and artificial neural networks were widely applied. However, these methods cannot always provide sufficient performance in complex vibration signals because they require manual feature extraction and can represent nonlinear dynamics to a limited extent [11]. In order to overcome these limitations, deep learning-based approaches have come to the fore in recent years. Convolutional Neural Networks (CNNs) have been able to successfully represent the local structural features of vibration signals thanks to their ability to automatically learn spatial patterns. Long-short-term memory (LSTM) networks have improved fault diagnosis performance by modeling long-term dependencies in time-series data. In particular, CNN-LSTM-based hybrid architectures have achieved high accuracy rates in bearing failure classification due to their ability to process spatial and temporal information simultaneously. Despite the extensive literature on bearing fault diagnosis using the CWRU dataset, most prior studies have primarily emphasized either new classifiers or isolated feature extraction pipelines. Recent review and benchmarking studies have also highlighted that strong performance on CWRU alone does not automatically establish methodological novelty or robust generalization [12–15]. In addition, many hybrid approaches still rely on direct feature concatenation, which may retain redundant components and may not explicitly model the dependency between handcrafted and learned feature spaces. Motivated by this gap, the present study focuses on correlation-aware representation learning rather than on merely combining multiple methods. In contrast to conventional pipelines, the proposed framework integrates three complementary design choices: (i) VMD-based denoising to stabilize fault-related spectral content, (ii) a frequency-token Transformer that learns contextual dependencies among physically meaningful band-energy tokens instead of raw samples, and (iii) regularized canonical correlation analysis (rCCA) to align analytical spectral descriptors with Transformer representations in a shared correlation-aware subspace. Therefore, the novelty of this study lies in the explicit modeling of cross-view relationships and the progressive analysis of spectral, Transformer, direct hybrid, and rCCA-enhanced hybrid representations under the same experimental protocol.

The main contributions of this study are summarized as follows:

- A VMD-based adaptive noise suppression framework is employed to enhance fault-related frequency components in bearing vibration signals, improving signal quality before feature extraction.

- A Transformer-based frequency-domain representation is developed by tokenizing band energy distributions, enabling the model to capture global dependencies and complex spectral patterns directly from vibration data.
- A correlation-aware hybrid feature construction strategy is introduced by integrating analytical spectral descriptors with Transformer-learned deep representations.
- A regularized Canonical Correlation Analysis (rCCA)-based feature fusion mechanism is proposed to model and strengthen the relationship between spectral and Transformer feature spaces, resulting in a more discriminative and compact hybrid feature vector.
- A comprehensive experimental evaluation using multiple classifiers (SVM, Random Forest, and XGBoost) demonstrates that the proposed rCCA-enhanced Transformer framework achieves superior classification performance compared to standalone spectral or deep representations. The proposed framework provides a robust and generalizable solution for vibration-based bearing fault diagnosis, offering practical potential for industrial condition monitoring and predictive maintenance applications.

The remainder of the paper is organized as follows. Section 2 presents a review of the related literature on vibration-based bearing fault diagnosis. Section 3 describes the proposed method in detail. Section 4 introduces the dataset and experimental setup, including the evaluation criteria and obtained results. Finally, Section 5 concludes the paper by summarizing the main findings and outlining directions for future work.

2. Related Works

Bearing failure classification has long been a key research area of intense interest to researchers in terms of ensuring the operational reliability of rotating machinery and effectively implementing predictive maintenance strategies. Initially, studies in this area focused primarily on traditional signal processing and feature extraction methods but have subsequently evolved significantly towards improved failure classification approaches based on machine learning and deep learning, possessing automated feature learning capabilities. Prabhakar et al. [16] proposed a Discrete Wave Transform (DWT)-based method for detecting inner ring, outer ring, and combination failures in ball bearings. The study demonstrated that both single and multiple failure states could be distinguished through time-frequency analysis. The results show that wavelet-based methods are effective in revealing local failure characteristics, but their generalization ability under variable operating conditions may be limited. Malhi and Gao [17] developed a feature selection approach based on principal component analysis for failure classification in mechanical systems. In the study, it was shown that converting the high-dimensional feature space derived from vibration signals into a lower-dimensional representation positively affects the classification accuracy. The findings reveal that dimensionality reduction operations both reduce the computational load and increase the stability of the classification process. However, it was emphasized that the overall performance of the method depends significantly on the quality of the initially defined feature set. Nouri Khajavi and Norouzi Keshtan [18] developed a method integrating Discrete Wave Transform with artificial neural networks for the identification of inner and outer ring failures in ball bearings. In the proposed approach, it was shown that failure-specific components can be separated at the local level through analyses performed in the time-frequency domain. It was experimentally shown that the obtained wavelet-based features improve the classification performance. However, it was stated in the study that the method could not exhibit sufficient generalization ability under different speed conditions. Gupta and Pradhan [19] discussed different machine learning-based classification approaches for bearing failure diagnosis in a comparative framework. In this study, the diagnostic performance of support vector machines, k-nearest

neighbor algorithms, and decision trees was analyzed using statistical features derived from the time and frequency domains. The results show that SVM-based models offer higher classification performance, especially in scenarios with limited data volumes. However, the dependence of the method on manual feature extraction was noted as a fundamental limitation in terms of practical applications. Soleimani and Khadem [20] developed a nonlinear diagnostic approach based on chaotic vibration behavior for detecting failures in ball bearings and gearboxes. Healthy and faulty states were modeled using phase-space reconstruction; chaotic metrics such as the largest Lyapunov exponent, approximate entropy, and correlation dimension were reported to provide high discrimination. The study revealed that chaos-based features can represent dynamic changes that cannot be captured by traditional statistical methods. Li et al. [21] proposed a CNNEPDNN model that can learn local and global features together. The model was compared with classical CNN architectures under different load conditions and was shown to provide higher classification accuracy. The model's discriminative representation power was validated by visualizing the feature space, but it was noted that variable speed and high noise conditions could negatively affect performance. Magar et al. [22] developed FaultNet, a data-driven and vibration-based deep learning architecture for rolling element bearings. It was shown that the multi-channel input structure provides richer feature representation compared to single-channel models. The high accuracy rates obtained in the CWRU and Paderborn datasets support the suitability of the method for online fault diagnosis. Zhao et al. [23] examined the use of deep learning approaches in the field of machine health monitoring in detail. In the study, it was emphasized that convolutional neural networks, in particular, can learn discriminative features from raw vibration data without the need for human intervention, thus providing higher diagnostic accuracy compared to traditional signal processing-based methods. However, the need for large-scale datasets and high computational costs for effective training of deep learning-based models are among the main limitations of the method. Shao et al. [24] developed an approach for classifying bearing failure types using multilayer convolutional neural network architectures. In the study, it was experimentally shown that the proposed model could distinguish different failure scenarios with high accuracy rates without the need for any manual feature extraction process. The findings show that deep CNN-based structures have the capacity to effectively learn complex and nonlinear patterns related to bearing failures. Luo et al. [25] proposed an improved failure classification approach by integrating spectral representations obtained at different scales with a convolutional neural network-based structure. In the study, it was shown that the diagnostic performance was significantly improved by using information derived from various frequency bands simultaneously. However, it was stated that the computational cost of the method is relatively high due to the combined use of multi-scale analysis and deep learning components. Alam et al. [26] developed a one-dimensional convolutional neural network model supported by transfer learning strategies for classifying bearing failures under variable operating conditions. In the proposed approach, it has been experimentally shown that high diagnostic performance can be achieved even in limited labeled data scenarios by performing information transfer between source and target datasets. The findings reveal that transfer learning offers an effective solution in terms of adapting to different operating conditions within the scope of improved fault classification methods. Hatipoğlu et al. [27] proposed a hybrid bearing fault diagnosis method combining features obtained from time, frequency, and chaotic domains. The feature set optimized with LASSO was evaluated using LSTM and attention mechanisms. High F1 scores obtained in CWRU and HUST datasets showed that the inclusion of chaotic features in the model improved the diagnostic performance. Sinitsin et al. [28] proposed a model for bearing fault diagnosis based on a hybrid CNN–MLP architecture integrating information from different data types.

In the study, it was experimentally shown that the multiple input structure can represent fault-related features more comprehensively and that this approach provides a significant improvement in classification performance. Chen et al. [29] systematically examined the studies conducted in the field of bearing failure diagnosis using bibliometric methods. The findings show that deep learning-based approaches have increased significantly in recent years, but these methods still have significant research gaps in terms of generalizability and model explainability in real industrial applications. Sahu et al. [30] compiled current studies in the field of bearing failure diagnosis and comprehensively examined the development of deep learning-based methods. It was emphasized that deep neural networks provide high accuracy by learning complex time-frequency patterns from raw sensor data, and it was predicted that multilayer hybrid architectures will offer more effective solutions in the future. Jamil and Khanam [31] investigated the effect of statistical feature ranking methods in the classification of bearing failures. Features selected with one-way ANOVA and Kruskal–Wallis tests were evaluated with classifiers such as SVM, KNN, and ANN. It was reported that features selected with the Kruskal–Wallis test significantly increased the classification accuracy. Ali et al. [32] proposed a Weighted Probability Ensemble Deep Learning (WPEDL)-based approach for the simultaneous diagnosis of bearing, rotor, and stator failures in induction motors. The ensemble structure, built on time-frequency representations extracted from vibration and current signals using STFT, produced more stable and higher accuracy results compared to single deep learning models. The study demonstrates that ensemble-based deep learning is a powerful alternative for industrial applications in multiple failure scenarios. Abbasi et al. [33] developed a CNN-LSTM-based multitasking deep learning model for the detection and classification of bearing failures using the HUST dataset. The model achieved high accuracy under different operating conditions, but dependence on a large amount of labeled data was noted as a significant limitation. The study highlights the need for novel methods for limited data conditions. A recent study proposed a multiscale residual antinoise network (MRANet) with a dynamic recalibration mechanism to improve fault diagnosis under limited data conditions. By combining STFT-based time–frequency representations with multibranch convolution and residual learning, the method enhances feature extraction and achieves high accuracy under varying load and speed conditions [34]. Another approach integrates a backtracking strategy, improved VMD, and infogram analysis for early fault detection in noisy signals. The method optimizes VMD parameters and selects informative components for signal reconstruction, providing more accurate estimation of incipient fault time compared to conventional techniques [35].

To clarify the research gap relative to representative recent studies, a concise comparison is provided in Table 1.

Table 1. Representative recent CWRU-based studies and the methodological position of the proposed framework.

Study	Main Strategy	Signal/View	Fusion Mechanism	Gap Addressed in This Study
Neupane and Seok [15]	Review of deep-learning-based bearing fault diagnosis studies using the CWRU dataset	CWRU vibration data in prior deep-learning studies.	No explicit feature-level fusion; survey of model-centric DL methods.	Motivates a framework that goes beyond stand-alone deep models by explicitly integrating complementary representations rather than only comparing architectures.
Hendriks et al. [14]	Benchmarking study for CWRU under a more realistic train/test split, showing flaws in the common setup	CWRU vibration data; original vs. proposed benchmark partitions with independent bearings.	No correlation-aware feature fusion; includes a time-frequency data fusion benchmark variant, but the paper's central contribution is benchmarking rigor.	Highlights that evaluation protocol and leakage-resistant benchmarking are crucial, not just high reported accuracy; this supports the need for a more principled methodology and fair validation.

Table 1. *Cont.*

Study	Main Strategy	Signal/View	Fusion Mechanism	Gap Addressed in This Study
Zhang et al. [13]	Review of machine-learning-based CWRU fault diagnosis methods, including dataset characteristics, feature selection, and classifiers.	CWRU vibration signals with emphasis on engineered features + ML pipelines.	Mostly direct/model-specific combinations, not an explicit cross-view alignment framework.	Motivates explicit study of complementary handcrafted and learned views, rather than relying only on conventional feature engineering or classifier selection.
Alonso-González et al. [12]	Envelope analysis with classical machine-learning classifiers for bearing diagnosis	Frequency-domain/envelope-spectrum features from CWRU vibration data; amplitudes at characteristic fault frequencies.	No multi-view fusion; conventional ML over envelope-derived predictors.	Shows that informative spectral features are useful, but cross-view dependency modeling is limited; this leaves room for integrating spectral descriptors with learned contextual features.
This study	VMD + frequency-token Transformer + rCCA	Spectral descriptors + deep contextual tokens	Correlation-aware rCCA alignment	Reduces redundancy and strengthens complementary information across handcrafted spectral and learned token-level views.

3. Method and Materials

The aim of this study is to automatically classify failure types from bearing vibration signals measured within the CWRU dataset. The proposed approach consists of VMD-based noise suppression, a Transformer-based hybrid feature vector that tokenizes band energies in the frequency domain, rCCA, which captures the relationship between Transformer representations and analytical spectral descriptors, and classification. The overall workflow of the proposed framework is summarized in Algorithm 1, which outlines the VMD-based denoising, frequency-token Transformer representation, rCCA-based correlation alignment, and the final classification stage. The overall framework of the proposed VMD–Transformer–rCCA-based fault diagnosis method is illustrated in Figure 1. For clarity, Equations (1) and (2) define the record-level notation and labels, Equations (3)–(8) describe the VMD-based denoising stage, Equations (9)–(18) define the analytical spectral descriptors, Equations (19)–(35) formalize token construction and Transformer-based representation learning, and Equations (36)–(42) describe the regularized CCA-based correlation alignment stage. These grouped references were added to make the mathematical flow easier to follow. Before feature extraction, each vibration record is segmented into partially overlapping windows so that local stationary behavior can be analyzed while preserving sufficient fault-related periodicity. To avoid scale dominance among heterogeneous descriptors, feature normalization is applied after feature extraction and before classifier training. In addition, the train/test split and 10-fold cross-validation procedures are performed at the record level to reduce optimistic leakage between highly similar windows originating from the same signal. To improve reproducibility, the revised manuscript now explicitly summarizes the practical settings used for VMD, Transformer training, and rCCA-based fusion. The vibration signals were segmented using a window length of 2048 samples with an overlap ratio of 50%. In the VMD stage, the number of modes was set to $K = 6$ and the penalty factor was selected as $\alpha = 2000$. An energy-threshold of $\tau_e = 0.05$ was applied to retain dominant modes during reconstruction. The frequency domain representation was constructed using $B = 16$ frequency bands. For the Transformer architecture, each frequency band was embedded into a latent space of dimension $d = 64$. The encoder consisted of 2 layers with 4 attention heads per layer. The model was trained using a batch size of 32 for 100 epochs with the Adam optimizer. The learning rate was set to 1×10^{-3}. In the rCCA stage, the regularization coefficients were selected as $\lambda_x = 0.1$ and $\lambda_y = 0.1$ to ensure numerical stability and prevent overfitting during covariance estimation.

Algorithm 1. VMD-Denoising + Frequency-Token Transformer + rCCA-Based Hybrid Feature Learning

Input: Vibration record $s^{(n)}[m](m = 0, \ldots, M-1)$, sampling frequency f_s; window length L, overlap; VMD params (K_v, η); bands B; Transformer params $(d_{\text{model}}, h, L_e)$; rCCA params $(\lambda_z, \lambda_g, q)$.

Output: Predicted label $\hat{y}^{(n)} \in \{1, \ldots, K\}$.

1. Segment $s^{(n)}[m]$ into windows $w_\tau(t)$, $\tau = 1, \ldots, T$.
2. **VMD:** Decompose each window: $w_\tau(t) = \sum_{k=1}^{K_v} u_{\tau,k}(t)$.
3. Compute mode energies $E_{\tau,k} = \sum_{n=0}^{L-1} u_{\tau,k}[n]^2$
4. Compute ratios $r_{\tau,k} = E_{\tau,k} / \sum_{j=1}^{K_v} E_{\tau,j}$
5. Select modes $K_\tau = \{k : r_{\tau,k} \geq \eta\}$.
6. Reconstruct denoised window $\widetilde{w}_\tau[n] = \sum_{k \in K_\tau} u_{\tau,k}[n]$
7. Compute DFT $X_\tau[l] = \sum_{n=0}^{L-1} \widetilde{w}_\tau[n] e^{-j2\pi ln/L}$, power $P_\tau[l] = \mid X_\tau[l] \mid^2$.
8. Map frequency bins $f_l = (l f_s)/L$
9. Extract spectral descriptors $g_\tau = \{SC_\tau, SE_\tau, PA_\tau, FV_\tau, \ldots\}$
10. Compute band energies $BP_\tau^{(b)} = \sum_{l: f_l \in [f_a^{(b)}, f_b^{(b)}]} P_\tau[l]$, form token vector $b_\tau = \left[BP_\tau^{(1)}, \ldots, BP_\tau^{(B)}\right]^T$
11. Embed each band token $e_{\tau,b} = W_b\, b_\tau[b] + c_b$, add band-position $e_{\tau,b} \leftarrow e_{\tau,b} + p_b$
12. Form Transformer input $E_\tau = \left[e_{\tau,1}, \ldots, e_{\tau,B}\right]^T$
13. Apply encoder self-attention $\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$ obtain contextual bands H_τ
14. Attention pooling : $a_{\tau,b} = u^T \tanh\left(W_p h_{\tau,b}\right)$,
15. Record-level stats : $\bar{z} = \frac{1}{T}\sum_\tau z_\tau$, $\text{Var}(z) = \frac{1}{T}\sum_\tau (z_\tau - \bar{z}).(z_\tau - \bar{z})$; similarly $(\bar{g}, \text{Var}(g))$
16. Form hybrid vector $h = [\bar{z}; \text{Var}(z); \bar{g}; \text{Var}(g)]$.
17. Build window matrices $Z = \left[z_1^T; \ldots; z_T^T\right]$, $G = \left[g_1^T; \ldots; g_T^T\right]$, covariances $\Sigma_{ZZ}, \Sigma_{GG}, \Sigma_{ZG}$
18. **rCCA:** Regularize $\Sigma_{ZZ}^\lambda = \Sigma_{ZZ} + \lambda_z I$
19. Keep top-q correlations $c_{\text{CCA}} = [\rho_1, \ldots, \rho_q]$
20. Classify with final feature $\phi = [h; c_{\text{CCA}}]$ to obtain $\hat{y}^{(n)}$ (e.g., SVM/RF/XGBoost).

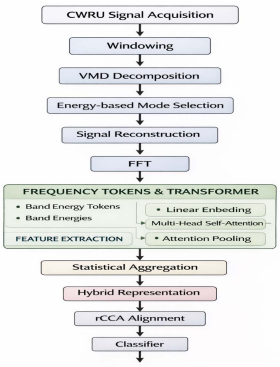


Figure 1. Flowchart of the proposed VMD-Transformer-rCCA-based bearing fault diagnosis framework.

The CWRU dataset contains vibration records obtained at different failure locations (inner ring, outer ring, ball) and different damage levels (including failure-free/healthy); these records are a widely used reference measurement source in machine health monitoring and predictive maintenance applications. Each measurement record is treated as a single-channel time series obtained from an accelerometer. Accordingly, let the n-th record be defined as follows:

$$s^{(n)}[m] \in \mathbb{R}, m = 0, 1, \ldots, M-1, n = 1, 2, \ldots, N, \quad (1)$$

Here, n denotes the index of the record in the dataset, and N represents the total number of records. The variable m is the discrete-time sample index, while M indicates the total number of samples in a single record. The term $s^{(n)}[m]$ represents the vibration

amplitude (acceleration) at the m-th sample of the n-th record. The sampling frequency of the recordings is denoted by f_s. Since bearing faults typically generate impulsive vibration patterns with periodic components, the signal carries meaningful information in both the time domain (impacts, transient structures) and the frequency domain (characteristic fault frequencies and sidebands). Therefore, the representation learning stages applied to each record aim to preserve as much information as possible from both domains. The true class label corresponding to each record is defined as follows:

$$y^{(n)} \in \{1, 2, \ldots, K\}, \tag{2}$$

Here, K denotes the total number of fault classes in the classification problem (e.g., Healthy, Inner Race Fault, Outer Race Fault, and Ball Fault). These labels specify which fault type is represented by the corresponding vibration record.

3.1. Denoising via Variational Mode Decomposition (VMD)

Each windowed signal $w_\tau(t)$ is decomposed into K_v intrinsic modes using Variational Mode Decomposition (VMD) [36,37]:

$$w_\tau(t) = \sum_{k=1}^{K_v} u_{\tau,k}(t) \tag{3}$$

Here, $u_{\tau,k}(t)$ denotes the k-th mode of the τ-th window. Each mode is assumed to be a narrow-band component centered around a specific frequency. VMD determines these modes by solving the following constrained optimization problem (in summarized form) [38]:

$$\min_{\{u_k\},\{\omega_k\}} \sum_{k=1}^{K_v} \| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \|_2^2 \text{ s.t.} \sum_{k=1}^{K_v} u_k(t) = w(t) \tag{4}$$

In this formulation, the objective function minimizes the bandwidth of each mode. The term inside the norm computes the analytic signal of each mode (via Hilbert transform), shifts it to baseband by $e^{-j\omega_k t}$, and evaluates its smoothness through the time derivative. Minimizing this quantity forces each mode to be compact around its center frequency ω_k. The reconstruction constraint ensures that the sum of all modes equals the original signal.

For the τ-th window, the energy contribution of each mode is computed as

$$E_{\tau,k} = \sum_{n=0}^{L-1} u_{\tau,k}[n]^2 \tag{5}$$

The relative energy ratio of each mode is then defined by

$$r_{\tau,k} = \frac{E_{\tau,k}}{\sum_{j=1}^{K_v} E_{\tau,j}} \tag{6}$$

This ratio quantifies the contribution of the k-th mode to the total signal energy. Modes with very low relative energy are typically associated with noise or insignificant fluctuations. To perform denoising, a subset of relevant modes is selected based on an energy threshold η:

$$K_\tau = \{k : r_{\tau,k} \geq \eta\} \tag{7}$$

Only the modes whose energy ratios exceed η are retained. The denoised window is then reconstructed as

$$\widetilde{w}_\tau[n] = \sum_{k \in K_\tau} u_{\tau,k}[n] \tag{8}$$

This reconstruction step suppresses low-energy, noise-dominated components while preserving dominant structural oscillations. As a result, irregular fluctuations in the spectrum are reduced, leading to more stable band energy estimates and more reliable spectral descriptors in subsequent processing stages. Example vibration signals from each fault class of the CWRU dataset and their corresponding VMD-denoised versions are shown in Figure 2. In each subplot, the blue curve represents the original raw vibration signal, while the orange curve shows the reconstructed signal obtained after applying VMD.

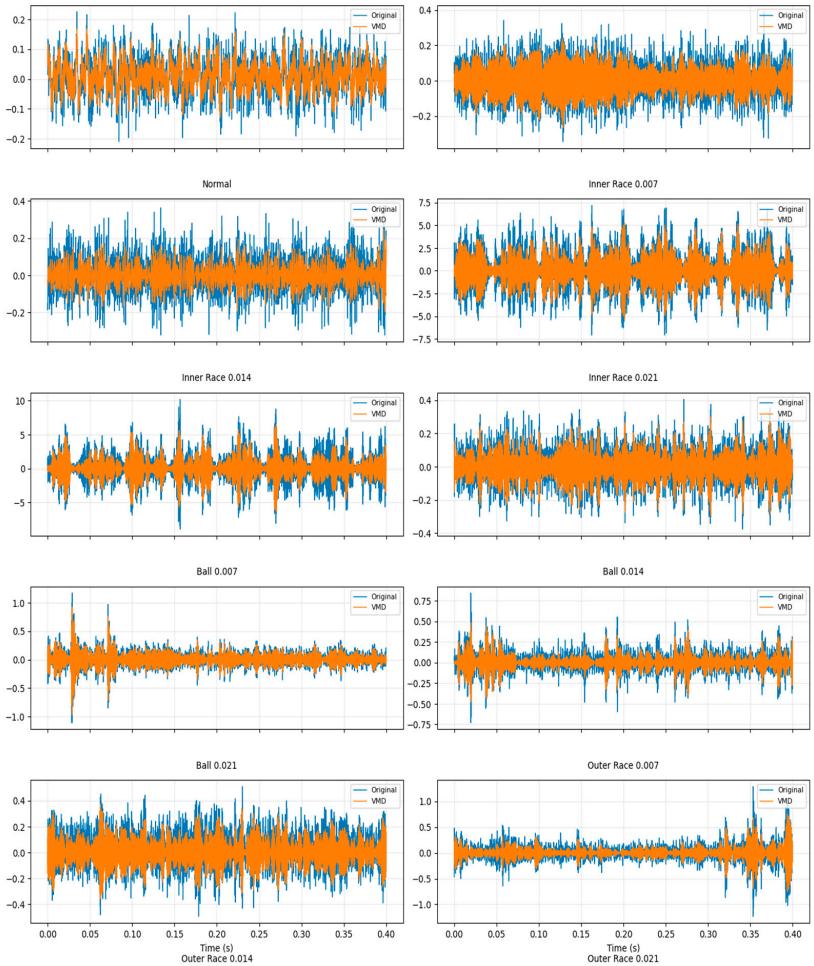


Figure 2. Original and VMD-denoised vibration signals for each fault class in the CWRU dataset.

3.2. Spectral Feature Extraction in the Frequency Domain

For each denoised window $\widetilde{w}_\tau[n]$, the discrete Fourier transform (DFT) is computed as

$$X_\tau[l] = \sum_{n=0}^{L-1} \widetilde{w}_\tau[n] e^{-j2\pi ln/L}, l = 0, \ldots, L-1 \tag{9}$$

This transformation maps the signal from the time domain into the frequency domain. Each coefficient $X_\tau[l]$ represents the complex amplitude of the frequency component indexed by l. The corresponding power spectrum is defined as

$$P_\tau[l] = \mid X_\tau[l] \mid^2 \tag{10}$$

which measures the energy content at each discrete frequency bin. The physical frequency axis associated with index l is given by

$$f_l = \frac{l f_s}{L} \tag{11}$$

where f_s is the sampling frequency. Using this frequency representation, several analytical spectral descriptors are computed for each window. Spectral centroid defined as [39–41];

$$SC_\tau = \frac{\sum_l f_l P_\tau[l]}{\sum_l P_\tau[l]} \tag{12}$$

The spectral centroid corresponds to the energy-weighted mean frequency. It can be interpreted as the "center of gravity" of the spectrum. If energy shifts toward higher frequencies—as often observed in faulty bearings due to impulsive components—the centroid increases accordingly. Spectral Entropy: First, the normalized spectral distribution is defined as

$$p_\tau[l] = \frac{P_\tau[l]}{\sum_l P_\tau[l]} \tag{13}$$

This normalization converts the power spectrum into a probability distribution over frequencies.

Spectral entropy is then computed as [41,42];

$$SE_\tau = -\sum_l p_\tau[l] \log(p_\tau[l] + \epsilon) \tag{14}$$

This quantity measures the disorder or irregularity of the spectral distribution. A highly concentrated spectrum (dominant frequency components) yields low entropy, whereas a more uniformly distributed spectrum—often associated with noise or complex fault behavior—produces higher entropy values. Band Energies: Let each frequency band b be defined over the interval

$$\left[f_a^{(b)}, f_b^{(b)}\right] \tag{15}$$

The energy contained within band b is calculated as [41–43];

$$BP_\tau^{(b)} = \sum_{l: f_l \in [f_a^{(b)}, f_b^{(b)}]} P_\tau[l], b = 1, \ldots, B \tag{16}$$

This feature aggregates spectral energy over predefined frequency ranges. The selection of band limits can be aligned with characteristic fault frequencies, making these features physically interpretable. The choice of the predefined frequency bands is guided by two considerations. First, band-wise aggregation reduces the sensitivity of the representation to isolated noisy bins and small frequency drifts. Second, it preserves interpretable regional energy patterns that are more consistent with fault-related harmonic redistribution than raw bin-level inputs. Accordingly, the Transformer receives compact tokens that summarize physically meaningful spectral neighborhoods while still allowing self-attention to capture long-range inter-band dependencies.

$$PA_\tau = \max_l \mid X_\tau[l] \mid \tag{17}$$

The peak amplitude represents the maximum magnitude of the Fourier coefficients. It reflects the strongest oscillatory component within the window and is particularly sensitive to periodic fault-induced vibrations. Frequency variance quantifies the spread of spectral energy around the centroid and is defined as

$$FV_\tau = \frac{\sum_l (f_l - SC_\tau)^2 P_\tau[l]}{\sum_l P_\tau[l]} \tag{18}$$

Frequency variance quantifies the spread of spectral energy around the centroid. While the centroid indicates where the spectrum is centered, the variance describes how widely

energy is distributed. Broader distributions typically correspond to more complex or broadband vibration behavior. Together, these spectral descriptors provide a complementary analytical representation of each window. While the Transformer-based representation captures learned inter-band relationships, these handcrafted spectral features encode physically interpretable information about energy distribution, concentration, and dispersion in the frequency domain.

3.3. Construction of Frequency Tokens and Token Embedding

In the proposed framework, the input tokens of the Transformer are constructed from band energy features extracted in the frequency domain. For each window τ, the band energy vector is defined as More specifically, each denoised window is converted into a sequence of scalar band-energy values, one token per frequency band. After linear projection and addition of learnable band-position embeddings, the resulting token matrix is processed by the Transformer encoder. This design differs from raw-signal and image-based Transformer inputs because it preserves domain interpretability while remaining computationally compact.

$$b_\tau = \left[BP_\tau^{(1)}, \ldots, BP_\tau^{(B)}\right]^T \in \mathbb{R}^B \tag{19}$$

where B denotes the total number of predefined frequency bands, and each component represents the total spectral energy within the corresponding band. By treating each frequency band as an individual token, the Transformer processes physically meaningful spectral summaries rather than raw time samples. This design enables the model to learn inter-band relationships such as energy redistribution, harmonic coupling, and spectral shifts that are characteristic of bearing faults. Since each band energy value is scalar, it must be projected into a higher-dimensional embedding space before being processed by the Transformer. This is achieved via a linear projection [44,45]:

$$e_{\tau,b} = W_b b_\tau[b] + c_b, W_b \in \mathbb{R}^{d_{\text{model}} \times 1}, c_b \in \mathbb{R}^{d_{\text{model}}} \tag{20}$$

This operation maps the scalar band energy into a d_{model}-dimensional representation. The projection allows the model to learn expressive embeddings for each frequency band while preserving numerical stability.

To retain information about the band index, a learnable band-position embedding p_b is added:

$$e_{\tau,b} \leftarrow e_{\tau,b} + p_b \tag{21}$$

This positional encoding ensures that the Transformer can distinguish between different frequency regions.

The resulting input matrix for window τ is

$$E_\tau = [e_{\tau,1}, \ldots, e_{\tau,B}]^T \in \mathbb{R}^{B \times d_{\text{model}}} \tag{22}$$

which serves as the input to the Transformer encoder.

3.4. Transformer Encoder

The Transformer encoder models dependencies among frequency bands using self-attention. The query, key, and value matrices are computed as [46]:

$$Q = E_\tau W_Q, K = E_\tau W_K, V = E_\tau W_V \tag{23}$$

These projections transform the embeddings into subspaces suitable for attention computation. The scaled dot product attention mechanism is defined as:

$$Attn(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{24}$$

Here, QK^T computes pairwise similarity scores between frequency bands. The scaling factor $\sqrt{d_k}$ prevents large dot-product values that could destabilize gradients. The softmax function converts similarity scores into normalized attention weights, enabling each band representation to aggregate information from all other bands.

For richer modeling capacity, multi-head attention is used [47–49]:

$$MSA(E_\tau) = \text{Concat}(head_1, \ldots, head_h)W_O \tag{25}$$

Each head learns relationships in a different subspace, allowing the encoder to capture multiple types of inter-band interactions. The encoder layer is defined as [50,51]:

$$Z = LN(E_\tau + MSA(E_\tau)) \tag{26}$$

$$H_\tau = LN(Z + FFN(Z)) \tag{27}$$

where LN denotes layer normalization and residual connections enhance training stability.

The feed-forward network is expressed as:

$$FFN(x) = W_2\phi(W_1x + b_1) + b_2 \tag{28}$$

where $\phi(\cdot)$ is a nonlinear activation function. This component introduces nonlinearity and increases representational capacity. After stacking L_e encoder layers, the contextualized window representation becomes:

$$H_\tau \in \mathbb{R}^{B \times d_{\text{model}}} \tag{29}$$

The Transformer output remains band-level. To obtain a single representation per window, learnable attention pooling is applied. The importance score for each band is computed as [50,51]:

$$a_{\tau,b} = u^T \tanh\left(W_p h_{\tau,b}\right) \tag{30}$$

This score quantifies the contribution of each band to the window representation. The normalized attention weights are:

$$\alpha_{\tau,b} = \frac{\exp(a_{\tau,b})}{\sum_{j=1}^{B} \exp\left(a_{\tau,j}\right)} \tag{31}$$

The final window representation is obtained as:

$$z_\tau = \sum_{b=1}^{B} \alpha_{\tau,b} h_{\tau,b} \in \mathbb{R}^{d_{\text{model}}} \tag{32}$$

Thus, z_τ represents an adaptive weighted aggregation of frequency bands. For each record, the window representations form the matrix [46–49]:

$$Z = \begin{bmatrix} z_1^T \\ \vdots \\ z_T^T \end{bmatrix} \in \mathbb{R}^{T \times d_z} \tag{33}$$

To achieve stable record-level summarization, both the mean and variance across windows are computed [47–49]:

$$\bar{z} = \frac{1}{T}\sum_{\tau=1}^{T} z_\tau \tag{34}$$

$$Var(z) = \frac{1}{T}\sum_{\tau=1}^{T}(z_\tau - \bar{z})\cdot(z_\tau - \bar{z}) \tag{35}$$

The same aggregation is applied to the spectral descriptors. The hybrid representation becomes:

$$h = [\bar{z}; Var(z); \bar{g}; Var(g)] \in \mathbb{R}^{D_h} \tag{36}$$

This vector encodes both magnitude (mean) and dispersion (variance) information from learned and analytical views.

3.5. Regularized Canonical Correlation Analysis (rCCA)

Canonical Correlation Analysis (CCA) seeks linear projections that maximize correlation between two views. The objective is [52–55]:

$$\max_{a,b} \frac{a^T \Sigma_{ZG} b}{\sqrt{a^T \Sigma_{ZZ} a}\sqrt{b^T \Sigma_{GG} b}} \tag{37}$$

where

$$\Sigma_{ZZ} = \frac{1}{T-1} Z^T Z, \Sigma_{GG} = \frac{1}{T-1} G^T G, \Sigma_{ZG} = \frac{1}{T-1} Z^T G \tag{38}$$

The solution reduces to the generalized eigenvalue problem:

$$\Sigma_{ZZ}^{-1}\Sigma_{ZG}\Sigma_{GG}^{-1}\Sigma_{GZ} a = \rho^2 a \tag{39}$$

However, when T is limited, covariance matrices may be ill-conditioned. Therefore, ridge regularization is introduced:

$$\Sigma_{ZZ}^{\lambda} = \Sigma_{ZZ} + \lambda_z I \tag{40}$$

$$\Sigma_{GG}^{\lambda} = \Sigma_{GG} + \lambda_g I \tag{41}$$

The regularized solution becomes:

$$\left(\Sigma_{ZZ}^{\lambda}\right)^{-1}\Sigma_{ZG}\left(\Sigma_{GG}^{\lambda}\right)^{-1}\Sigma_{GZ} a = \rho^2 a \tag{42}$$

The first q canonical correlation coefficients form the correlation signature:

$$c_{\mathrm{CCA}} = \left[\rho_1, \rho_2, \ldots, \rho_q\right] \in \mathbb{R}^q \tag{43}$$

In practice, the regularization coefficients are selected to stabilize the covariance estimates and to avoid overfitting when the dimensionality of the hybrid feature space is high relative to the number of training records. Accordingly, λ_x and λ_y are chosen from a small validation grid, and the final setting is determined by the best average validation accuracy together with numerical stability of the covariance inversion. This selection strategy has been clarified in the revised manuscript because the effectiveness of rCCA depends on balancing correlation preservation and regularization strength.

To further investigate the discriminative capability of the proposed rCCA-based feature integration, a t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization was performed on the hybrid feature space before and after correlation alignment. The effect of regularized canonical correlation analysis (rCCA) on the hybrid feature representation

is illustrated in Figure 3. The t-SNE projections demonstrate that the correlation-aligned hybrid features exhibit a more structured and compact distribution compared to the original hybrid features.

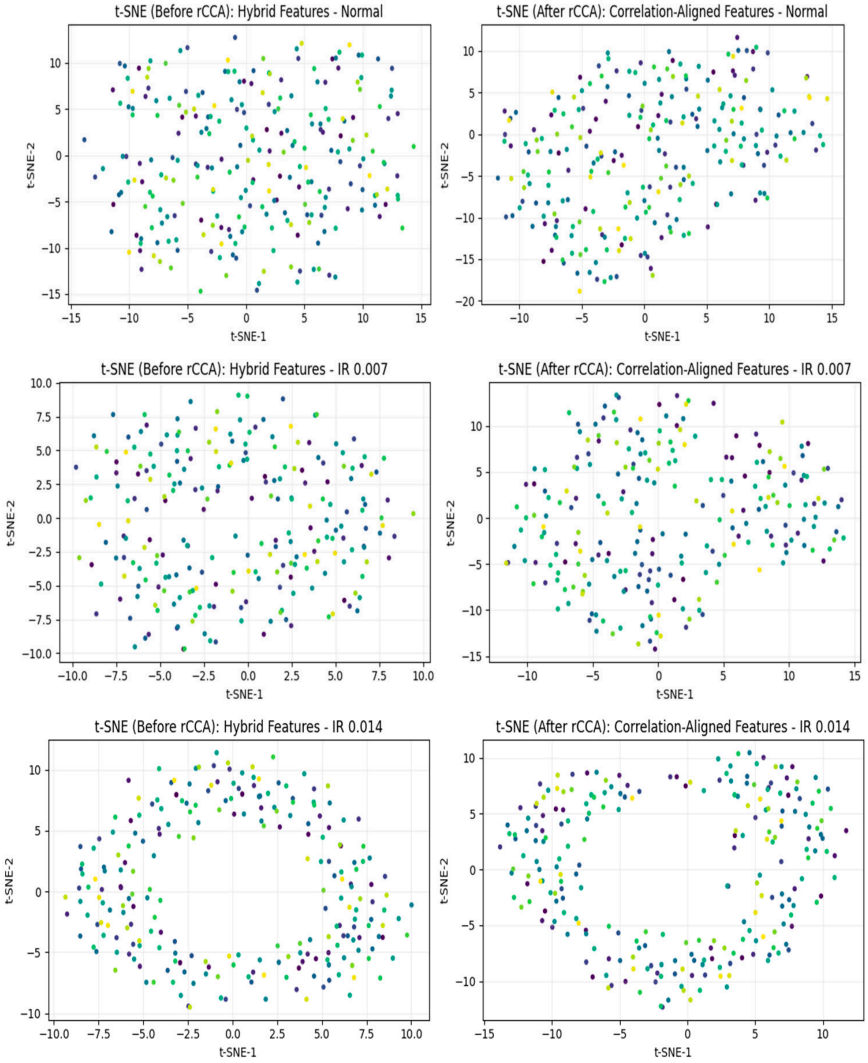


Figure 3. t-SNE projections of hybrid features before and after applying regularized canonical correlation analysis (rCCA). Colors represent different class labels, where each color corresponds to a specific category.

4. Results

4.1. Dataset

The dataset used for experimental validation in this study is the CWRU Bearing Dataset, developed by the Bearing Data Center at Case Western Reserve University and one of the most widely used open-access fault diagnostic datasets in the literature [56]. This dataset serves as a standard benchmark enabling vibration-based analysis of bearing failures in rotating machinery. It is particularly considered a reference dataset for performance comparisons of machine learning and deep learning-based fault classification approaches. The CWRU dataset was obtained through a controlled experimental setup in a laboratory environment. The experimental system primarily consists of a 2 HP three-phase induction motor, a torque transducer, a dynamometer (load application system), a data acquisition and control unit, and an accelerometer sensor. Vibration signals were measured from two different locations on the bearing: the Drive End (DE) and the Fan End (FE). High-precision vibration measurements were achieved by mounting the sensors directly onto the bearing

housing. This configuration allowed for the separation of the positional effects of the failures. The vibration signals obtained within the CWRU dataset consist of acceleration-based measurements. Sampling frequencies of 12 kHz and 48 kHz were used during the data collection process. Experiments were conducted at load levels of 0, 1, 2, and 3 HP to represent different operating conditions on the motor. Depending on these load levels, motor rotational speeds ranged from approximately 1730 to 1797 RPM. Vibration signals were measured via sensors placed at both the drive end (DE) and fan end (FE) of the bearing. All data are presented in MATLAB R2026a (.mat) format. The small variations in motor speed due to load changes allow the dataset to represent different operational conditions, thus enabling the evaluation of the generalization performance of the developed model under variable operating conditions. The classification structure used in the CWRU dataset is defined as 10 classes in this study, and the relevant distribution is given in Table 2. The dataset includes one healthy (normal) condition and three different failure types with three different failure diameter variations. Failure types were defined as inner race, ball, and outer race. Artificial damage with diameters of 0.007 inches, 0.014 inches, and 0.021 inches was created for each failure type. This structure allows not only the failure type but also the failure severity to be included in the classification problem. Thus, the model's ability to distinguish between different failure types and different damage levels can be evaluated. In the present study, the classification problem is organized as a 10-class task consisting of one healthy state and nine fault conditions obtained from three fault locations (inner race, ball, and outer race) and three defect diameters (0.007, 0.014, and 0.021 inches). Load conditions of 0–3 HP and speed variations of approximately 1730–1797 rpm are explicitly considered to reflect changing operating conditions within the public benchmark. Since the public CWRU release primarily provides vibration recordings and experimental metadata rather than high-resolution photographs of the induced defects, the present validation relies on the documented defect type, defect size, sensor position, sampling rate, and operating-load information reported for the dataset. This limitation has been acknowledged explicitly in the revised manuscript.

Table 2. Ten-class fault distribution used in the CWRU dataset.

Class No	Fault Type	Fault Diameter
0	Normal	—
1	Inner Race	0.007
2	Inner Race	0.014
3	Inner Race	0.021
4	Ball	0.007
5	Ball	0.014
6	Ball	0.021
7	Outer Race	0.007
8	Outer Race	0.014
9	Outer Race	0.021

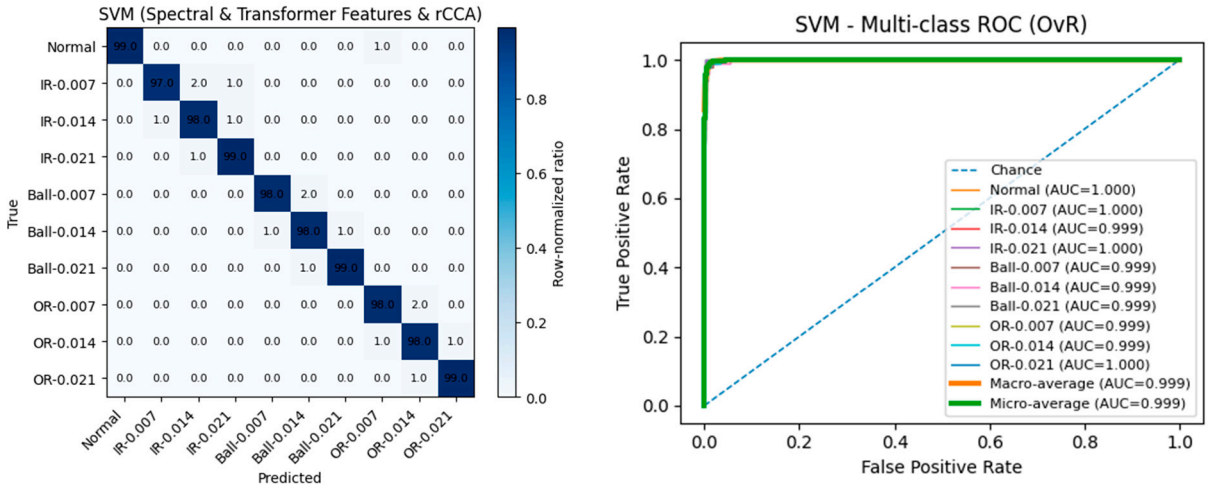
4.2. Results Analysis

This section presents the experimental findings obtained from the proposed feature extraction and fusion framework evaluated using the different classifier. The primary objective of this analysis is to investigate the discriminative capability of individual feature domains (spectral and transformer-based features), their direct hybridization, and the proposed rCCA-enhanced hybrid feature integration strategy. Performance is evaluated under two validation protocols: a conventional 70–30 train–test split and 10-fold cross-validation, ensuring both robustness and generalizability of the reported results. Standard classification metrics including accuracy, precision, recall, and F1-score are reported to

provide a comprehensive assessment of classification effectiveness across the ten bearing fault categories. As shown in Table 3, the experimental results demonstrate a clear and consistent performance improvement as the feature representation evolves from single-domain features to the proposed rCCA-based hybrid structure. When only spectral features are employed, the SVM classifier achieves an accuracy of 89.42% under the 70–30 split and 89.86% under 10-fold cross-validation. Although these results indicate reasonable discriminative power, they reveal limitations in capturing complex non-linear and temporal characteristics of bearing fault signals. The incorporation of transformer-based features significantly enhances classification performance, achieving 93.25% accuracy in the 70–30 setup and 93.98% under 10-fold validation. This improvement confirms that transformer-derived representations effectively model contextual and sequential dependencies within the vibration signals, yielding stronger class separability compared to purely frequency-domain features. A more substantial improvement is observed when spectral and transformer features are directly fused without correlation alignment. The hybrid feature representation without rCCA increases accuracy to 95.87% (70–30) and 96.48% (10-fold), indicating that complementary information from both domains strengthens the discriminative representation. Nevertheless, the absence of correlation-aware alignment may still allow redundant or weakly informative components to persist within the fused feature space. The highest performance is achieved when rCCA-based feature fusion is applied. Under the 70–30 split, the proposed hybrid model reaches 97.54% accuracy, while 10-fold cross-validation further improves accuracy to 98.21%. Similar trends are observed across precision, recall, and F1-score metrics, with the 10-fold rCCA-enhanced model achieving 97.90%, 97.42%, and 97.64%, respectively. These findings confirm that canonical correlation-based alignment effectively enhances inter-feature dependency modeling and reduces redundancy, leading to a more compact and discriminative representation. Figure 3 illustrates the classification performance of the SVM model using the proposed rCCA-based hybrid feature representation. Specifically, Figure 4a presents the normalized confusion matrix, while Figure 4b depicts the corresponding multi-class ROC curves obtained using the one-vs-rest (OvR) strategy. All class-specific ROC curves closely approach the upper-left corner of the ROC space, indicating near-perfect discrimination capability. The macro-average and micro-average AUC values are both approximately 0.999, demonstrating that the model maintains high sensitivity and specificity across all classes. The minimal deviation between macro and micro averages also suggests balanced classification performance without bias toward dominant classes.

Table 3. Comparative evaluation of SVM using spectral, Transformer, and rCCA-based hybrid features under 70–30 and 10-fold cross-validation.

Method	Spectral	Transformer	Hybrid Feature Without rCCA	Hybrid Feature with rCCA	Split	Acc (%)	Prec (%)	Recall (%)	F1 (%)
SVM	√				70–30%	89.42	88.98	88.11	88.53
		√			70–30%	93.25	92.84	92.61	92.73
			√		70–30%	95.87	95.41	95.09	95.22
				√	70–30%	97.54	97.02	96.81	96.92
	√				10-fold	89.86	89.33	89.02	89.17
		√			10-fold	93.98	93.44	93.22	93.31
			√		10-fold	96.48	96.03	95.66	95.79
				√	10-fold	98.21	97.90	97.42	97.64



(**a**) (**b**)

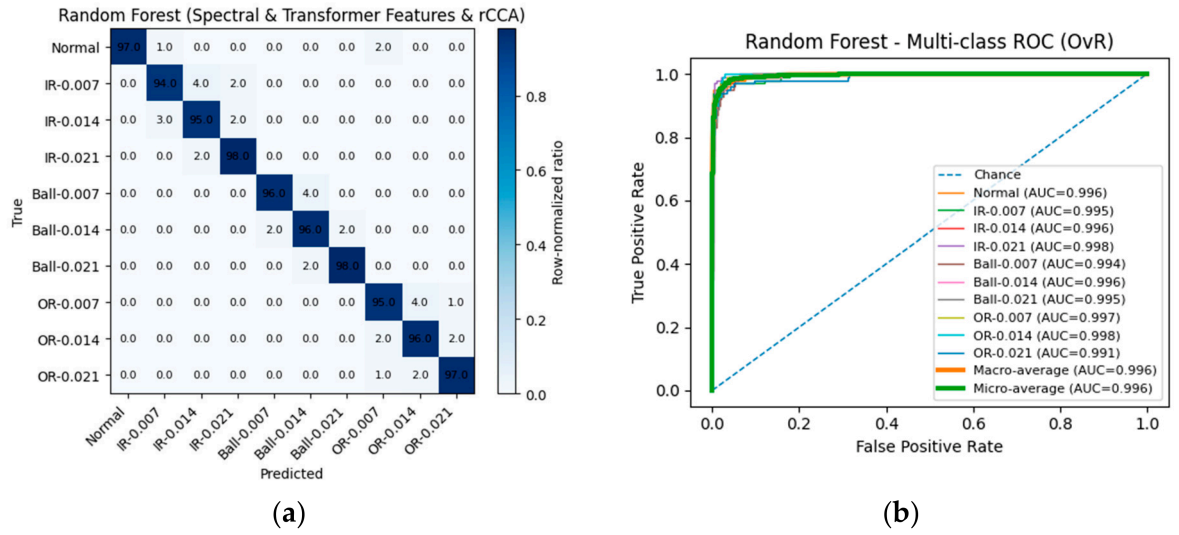
Figure 4. (**a**) Confusion matrix and (**b**) multi-class ROC curves of SVM with rCCA-based hybrid features.

The performance analysis of the Random Forest classifier under different feature extraction and fusion strategies is summarized in Table 4, while the corresponding classification behavior is illustrated in Figure 3 through the normalized confusion matrix and multi-class ROC curves. These results provide further insight into the contribution of hybrid feature modeling and correlation-aware integration within a tree-based ensemble framework. According to Table 3, the use of purely spectral descriptors yields the lowest performance among all tested configurations, with classification accuracy of 83.77% under the 70–30 split and 84.20% under 10-fold cross-validation. Although spectral information captures important frequency-domain characteristics of bearing faults, it appears insufficient for modeling more intricate signal variations when used alone within the Random Forest structure. When transformer-derived representations are introduced, the performance improves notably. The transformer-based configuration reaches 89.21% accuracy in the 70–30 setting and 89.98% under 10-fold validation. This improvement highlights the benefit of contextual and sequential feature encoding, suggesting that transformer representations effectively capture structural dependencies that are not fully represented in spectral analysis alone. A further performance increase is observed when spectral and transformer features are directly combined without applying correlation alignment. The hybrid representation without rCCA achieves 92.94% accuracy (70–30) and 93.54% (10-fold), demonstrating that complementary information from frequency and learned representations enhances discrimination capability. Nevertheless, some degree of feature redundancy may still persist due to the absence of explicit cross-domain correlation modeling. The most pronounced improvement is achieved when rCCA-based feature fusion is applied. Under the 70–30 split, the Random Forest classifier reaches 95.63% accuracy, while 10-fold cross-validation further increases accuracy to 96.74%. Similar improvements are consistently reflected in precision, recall, and F1-score metrics, confirming that correlation-guided integration refines the feature space and enhances separability across fault categories. The consistent gain across both validation protocols indicates that the proposed correlation-aware fusion strategy is not sensitive to the evaluation scheme and maintains stable generalization performance. The visual evidence provided in Figure 5a supports these findings. The normalized confusion matrix demonstrates strong concentration of predictions along the main diagonal, indicating accurate recognition of most fault categories. While minor confusion is visible between adjacent severity levels—particularly among low-amplitude inner race and ball faults—misclassification between fundamentally different fault types remains limited. This

suggests that the fused feature representation preserves class-specific structural information effectively. Further confirmation is provided by the ROC analysis in Figure 5b. The one-vs-rest curves for all ten classes approach the ideal upper-left region of the ROC space. The macro-average and micro-average AUC values are approximately 0.996, indicating high discriminative capability and balanced sensitivity-specificity trade-offs across classes. The close alignment between macro and micro averages implies that the model does not favor specific classes and maintains relatively uniform classification performance.

Table 4. Comparative evaluation of Random Forest using spectral, Transformer, and rCCA-based hybrid features under 70–30 and 10-fold cross-validation.

Method	Spectral	Transformer	Hybrid Feature Without rCCA	Hybrid Feature with rCCA	Split	Acc (%)	Prec (%)	Recall (%)	F1 (%)
Random Forest	√				70–30%	83.77	83.12	82.54	82.66
		√			70–30%	89.21	88.76	88.31	88.42
			√		70–30%	92.94	92.21	91.88	92.01
				√	70–30%	95.63	95.07	94.78	94.91
	√				10-fold	84.20	83.75	83.09	83.31
		√			10-fold	89.98	89.51	89.20	89.34
			√		10-fold	93.54	92.97	92.61	92.75
				√	10-fold	96.74	96.11	95.84	95.97



(**a**) (**b**)

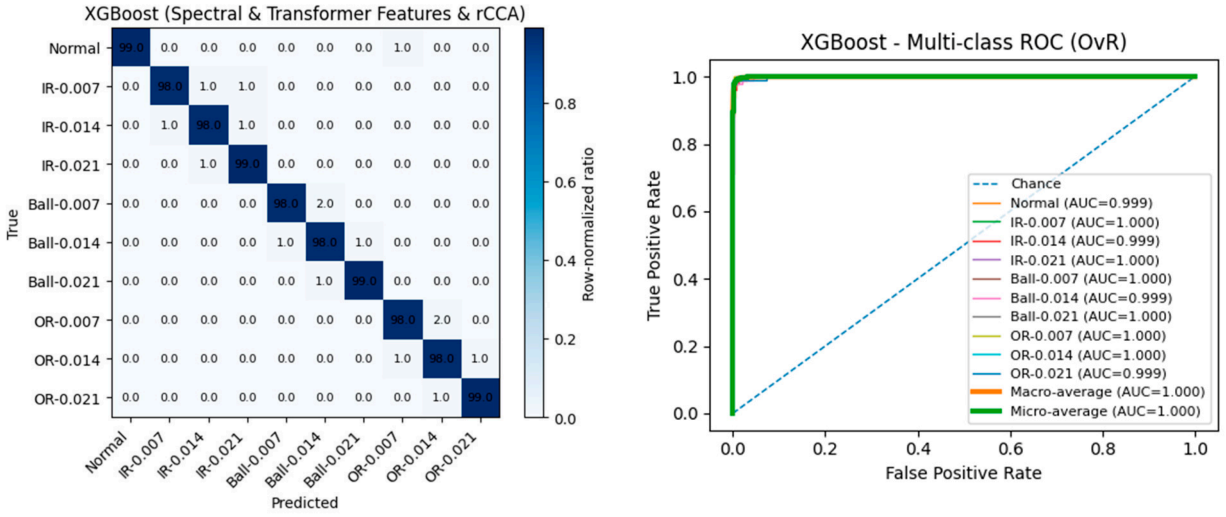
Figure 5. (**a**) Confusion matrix and (**b**) multi-class ROC curves of Random Forest with rCCA-based hybrid features.

The performance characteristics of the XGBoost classifier under different feature configurations are summarized in Table 5, and the corresponding classification behavior is illustrated in Figure 4 through the normalized confusion matrix and multi-class ROC analysis. These results further clarify the impact of transformer integration, hybrid feature modeling, and rCCA-based correlation alignment within a gradient boosting framework. From Table 5, it is evident that using spectral descriptors alone results in the lowest predictive performance, yielding 85.63% accuracy in the 70–30 split and 86.12% under 10-fold cross-validation. While spectral information captures frequency-domain signatures of bearing defects, it does not fully encode temporal dynamics or inter-feature dependencies required for highly discriminative modeling within boosting-based ensemble structures. A substantial improvement is observed when transformer-derived representa-

tions are utilized independently. The transformer-based model achieves 91.40% accuracy in the 70–30 evaluation and 92.04% in the 10-fold setup, demonstrating that contextual sequence modeling significantly enhances feature expressiveness. This confirms that transformer-encoded representations capture complex signal relationships that extend beyond conventional frequency-based descriptors. When spectral and transformer features are directly merged without correlation alignment, performance increases further. The hybrid configuration without rCCA reaches 94.88% accuracy under the 70–30 split and 95.61% with 10-fold validation. This improvement suggests that the complementary strengths of frequency-domain and deep contextual representations jointly enhance class discrimination. However, the absence of correlation-based refinement may still allow partially redundant components to influence the feature space. The highest performance is obtained when rCCA-based feature fusion is applied prior to classification. Under the 70–30 split, the XGBoost model achieves 97.42% accuracy, while 10-fold cross-validation increases the accuracy to 98.36%. Precision, recall, and F1-score follow the same improvement trend, reaching 97.92%, 97.41%, and 97.63%, respectively, under 10-fold validation. These results demonstrate that correlation-aware alignment effectively enhances inter-domain feature coherence, reduces redundancy, and strengthens class separability within the boosting framework. The confusion matrix in Figure 6a visually confirms the quantitative findings. Most class predictions are concentrated along the principal diagonal, indicating highly reliable recognition across all ten fault categories. Only marginal confusion is visible between neighboring fault severities, particularly among closely related inner race and ball defect sizes. Importantly, cross-type misclassification remains minimal, indicating that the fused representation preserves structural fault distinctions effectively. The ROC analysis in Figure 6b further supports these conclusions. The one-vs-rest ROC curves for all classes closely approach the ideal top-left corner of the ROC space. Both macro-average and micro-average AUC values are approximately 1.000, reflecting near-perfect discrimination capability across categories. The minimal separation between macro and micro AUC values suggests balanced predictive behavior without dominance of specific classes. Collectively, the results in Table 4 and Figure 6 demonstrate that XGBoost benefits substantially from the proposed multi-domain correlation-aware feature integration strategy. Compared to individual feature domains and non-aligned hybrid representations, the rCCA-enhanced fusion framework consistently delivers superior classification accuracy and stability under both validation schemes. These findings confirm that correlation-guided hybrid feature construction plays a critical role in achieving high-precision multi-class bearing fault diagnosis when coupled with gradient boosting techniques.

Table 5. Comparative evaluation of XGBoost using spectral, Transformer, and rCCA-based hybrid features under 70–30 and 10-fold cross-validation.

Method	Spectral	Transformer	Hybrid Feature Without rCCA	Hybrid Feature with rCCA	Split	Acc (%)	Prec (%)	Recall (%)	F1 (%)
XGBoost	√				70–30%	85.63	85.22	84.81	84.93
		√			70–30%	91.40	91.05	90.63	90.81
			√		70–30%	94.88	94.31	93.94	94.12
				√	70–30%	97.42	96.98	96.63	96.78
	√				10-fold	86.12	85.71	85.20	85.35
		√			10-fold	92.04	91.68	91.22	91.36
			√		10-fold	95.61	95.04	94.68	94.79
				√	10-fold	98.36	97.92	97.41	97.63



(**a**) (**b**)

Figure 6. (**a**) Confusion matrix and (**b**) multi-class ROC curves of XGBoost with rCCA-based hybrid features.

Table 6 provides a consolidated comparison of classification accuracy across three different learning algorithms under 10-fold cross-validation, highlighting the progressive impact of feature enhancement and correlation-aware fusion. Several important observations emerge from this comparative evaluation. First, for all classifiers, a consistent monotonic improvement in accuracy is observed as the feature representation evolves from purely spectral descriptors to transformer-based features, then to direct hybridization, and finally to rCCA-enhanced hybrid fusion. This uniform trend confirms that richer multi-domain representations significantly strengthen class discrimination regardless of the underlying classifier architecture. When examining the baseline spectral configuration, SVM achieves the highest starting performance (89.86%), followed by XGBoost (86.12%) and Random Forest (84.20%). This indicates that SVM appears more capable of exploiting frequency-domain features compared to tree-based ensemble methods. However, once transformer features are introduced, all models exhibit substantial improvement, suggesting that contextual and sequence-aware representations provide a more expressive feature space. The direct hybrid fusion without rCCA further improves accuracy across all classifiers, demonstrating the complementary nature of spectral and transformer representations. Nevertheless, the most significant performance gains occur after applying rCCA-based feature alignment. The rCCA-enhanced configuration yields the highest accuracies for each classifier: 98.21% for SVM, 96.74% for Random Forest, and 98.36% for XGBoost. An important observation concerns the magnitude of improvement introduced by rCCA. The largest relative gain is observed for Random Forest (+3.20%), followed by XGBoost (+2.75%) and SVM (+1.73%). This suggests that correlation-aware alignment plays a particularly crucial role for tree-based ensemble methods, likely because rCCA reduces redundancy and produces a more compact and discriminative feature subspace that benefits hierarchical decision boundaries. In contrast, SVM already demonstrates strong discrimination with direct hybrid features, resulting in a comparatively smaller but still meaningful increment. Overall, XGBoost achieves the highest final accuracy (98.36%), closely followed by SVM (98.21%), while Random Forest remains slightly lower but still highly competitive. The narrow performance gap between SVM and XGBoost indicates that both margin-based and gradient boosting approaches effectively exploit the correlation-enhanced hybrid representation. Although the numerical gains after rCCA may appear moderate in absolute percentage points, they are methodologically meaningful for three reasons. First, the im-

provement is consistent across all three classifiers, indicating that the gain is not tied to a single decision model. Second, the same trend is observed under both the 70–30 split and 10-fold cross-validation, which supports robustness rather than accidental overfitting. Third, the gain is obtained through feature-space refinement rather than by increasing classifier complexity, suggesting that the proposed correlation-aware fusion improves representation quality itself. From a diagnostic perspective, the stronger performance of the rCCA-enhanced representation indicates that analytical spectral descriptors and Transformer embeddings encode complementary fault evidence. When these views are aligned in a shared subspace, redundant information is suppressed and class-specific structures become more compact, which is also consistent with the t-SNE visualization. Therefore, the contribution of the proposed framework should be interpreted not only in terms of headline accuracy but also in terms of more structured and discriminative feature organization.

Table 6. Summary of key performance results across all classifiers.

Classifier	Spectral Acc (%)	Transformer Acc (%)	Hybrid Feature Without rCCA Acc (%)	Hybrid Feature with rCCA Acc (%)	rCCA Gain (%)
SVM	89.86	93.98	96.48	98.21	+1.73
Random Forest	84.20	89.98	93.54	96.74	+3.20
XGBoost	86.12	92.04	95.61	98.36	+2.75

5. Conclusions

This study presents a Transformer-based bearing fault classification framework enhanced by VMD-based noise suppression and rCCA-driven correlation-aware feature fusion. The proposed approach integrates analytical spectral descriptors and deep Transformer representations into a unified hybrid feature space, enabling the model to exploit both physically meaningful frequency-domain characteristics and data-driven global dependencies. The experimental results demonstrate that combining spectral features with Transformer-learned representations significantly improves classification performance compared to single-representation approaches. Furthermore, incorporating regularized Canonical Correlation Analysis (rCCA) strengthens the relationship between heterogeneous feature spaces, leading to more discriminative and compact hybrid feature vectors. Across all evaluated classifiers, the rCCA-enhanced hybrid representation consistently achieves the highest accuracy, precision, recall, and F1-score values. The findings confirm that modeling inter-feature correlations plays a critical role in improving robustness and generalization in vibration-based fault diagnosis. By integrating adaptive signal decomposition, deep representation learning, and correlation-aware fusion within a unified framework, the proposed method offers a reliable and scalable solution for intelligent bearing health monitoring systems. Future research may focus on extending the proposed framework to variable operating conditions, imbalanced fault scenarios, and real-time industrial deployment environments, as well as exploring advanced cross-domain transfer learning strategies for enhanced generalization performance. From a computational perspective, the total cost of the framework is dominated by three stages: VMD-based decomposition for each signal window, Transformer encoding of the band-token sequence and covariance estimation/eigendecomposition for rCCA. Therefore, the proposed method is more computationally demanding than using handcrafted spectral features alone; however, the added cost is compensated by the consistent improvement in representation quality and classification robustness. The study also has several limitations. The experiments are restricted to the controlled CWRU benchmark, the defect conditions are artificially induced, and the public release does not include high-resolution defect photographs for visual confirmation.

Moreover, the method has not yet been validated on variable-speed field measurements or strongly imbalanced real industrial datasets. Future work will focus on multi-dataset validation, external generalization tests, and real-time deployment analysis.

Author Contributions: Conceptualization, M.B.E. and T.K.; methodology, M.B.E. and A.Ç.; software, M.B.E.; validation, M.B.E. and T.K.; formal analysis, M.B.E.; investigation, M.B.E. and T.K.; resources, T.K.; data curation, T.K.; writing—original draft preparation, M.B.E. and A.Ç.; writing—review and editing, M.B.E. and T.K.; visualization, M.B.E. and A.Ç.; supervision, M.B.E.; project administration, M.B.E. All authors have read and agreed to the published version of the manuscript.

Funding: We would like to thank Inonu University Scientific Research Projects Unit for supporting this study with the project code FBA2025-4529.

Data Availability Statement: Data derived from public domain resources.

Acknowledgments: During the preparation of this work, the author used ChatGPT 5.2 to check the grammar, improve the readability of the text, and correct the spelling of the manuscript. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Randall, R.B.; Antoni, J. Rolling Element Bearing Diagnostics—A Tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520. [CrossRef]
2. Tian, Y.; Lu, C.; Wang, Z.L. Approach for Hydraulic Pump Fault Diagnosis Based on WPT-SVD and SVM. *Appl. Mech. Mater.* **2015**, *764*, 191–197. [CrossRef]
3. Tandon, N.; Choudhury, A. A Review of Vibration and Acoustic Measurement Methods for the Detection of Defects in Rolling Element Bearings. *Tribol. Int.* **1999**, *32*, 469–480. [CrossRef]
4. Li, Y.; Xu, M.; Wang, R.; Huang, W. A Fault Diagnosis Scheme for Rolling Bearing Based on Local Mean Decomposition and Improved Multiscale Fuzzy Entropy. *J. Sound Vib.* **2016**, *360*, 277–299. [CrossRef]
5. Shang, Y.; Tang, X.; Zhao, G.; Jiang, P.; Lin, T.R. A Remaining Life Prediction of Rolling Element Bearings Based on a Bidirectional GRU and CNN. *Measurement* **2022**, *202*, 111893. [CrossRef]
6. Lei, Y.; Li, N.; Guo, L.; Yan, T.; Lin, J. Machinery Health Prognostics: A Systematic Review from Data Acquisition to RUL Prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [CrossRef]
7. Guo, J.; Li, Z.; Li, M. A Review on Prognostics Methods for Engineering Systems. *IEEE Trans. Reliab.* **2019**, *69*, 1110–1129. [CrossRef]
8. Gao, Z.; Cecati, C.; Ding, S.X. A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3757–3767. [CrossRef]
9. Widodo, A.; Yang, B.-S. Support Vector Machine in Machine Condition Monitoring and Fault Diagnosis. *Mech. Syst. Signal Process.* **2007**, *21*, 2560–2574. [CrossRef]
10. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
11. Guo, J.; Yang, Y.; Li, H.; Wang, J.; Tang, A.; Shan, D.; Huang, B. A Hybrid Deep Learning Model towards Fault Diagnosis of Drilling Pump. *Appl. Energy* **2024**, *372*, 123773. [CrossRef]
12. Alonso-González, M.; Díaz, V.G.; Pérez, B.L.; G-Bustelo, B.C.P.; Anzola, J.P. Bearing Fault Diagnosis With Envelope Analysis and Machine Learning Approaches Using CWRU Dataset. *IEEE Access* **2023**, *11*, 57796–57805. [CrossRef]
13. Zhang, X.; Zhao, B.; Lin, Y. Machine Learning Based Bearing Fault Diagnosis Using the Case Western Reserve University Data: A Review. *IEEE Access* **2021**, *9*, 155598–155608. [CrossRef]
14. Hendriks, J.; Dumond, P.; Knox, D.A. Towards Better Benchmarking Using the CWRU Bearing Fault Dataset. *Mech. Syst. Signal Process.* **2022**, *169*, 108732. [CrossRef]
15. Neupane, D.; Seok, J. Bearing Fault Detection and Diagnosis Using Case Western Reserve University Dataset with Deep Learning Approaches: A Review. *IEEE Access* **2020**, *8*, 93155–93178. [CrossRef]
16. Prabhakar, S.; Mohanty, A.R.; Sekhar, A.S. Application of Discrete Wavelet Transform for Detection of Ball Bearing Race Faults. *Tribol. Int.* **2002**, *35*, 793–800. [CrossRef]
17. Malhi, A.; Gao, R.X. PCA-Based Feature Selection Scheme for Machine Defect Classification. *IEEE Trans. Instrum. Meas.* **2004**, *53*, 1517–1525. [CrossRef]

18. Nouri Khajavi, M.; Norouzi Keshtan, M. Intelligent Fault Classification of Rolling Bearings Using Neural Network and Discrete Wavelet Transform. *J. Vibroeng.* **2014**, *16*, 761–769.
19. Gupta, P.; Pradhan, M.K. Fault Detection Analysis in Rolling Element Bearing: A Review. *Mater. Today Proc.* **2017**, *4*, 2085–2094. [CrossRef]
20. Soleimani, A.; Khadem, S.E. Early Fault Detection of Rotating Machinery Through Chaotic Vibration Feature Extraction of Experimental Data Sets. *Chaos Solitons Fractals* **2015**, *78*, 61–75. [CrossRef]
21. Li, H.; Huang, J.; Ji, S. Bearing Fault Diagnosis with a Feature Fusion Method Based On an Ensemble Convolutional Neural Network and Deep Neural Network. *Sensors* **2019**, *19*, 2034. [CrossRef]
22. Magar, R.; Ghule, L.; Li, J.; Zhao, Y.; Farimani, A.B. FaultNet: A Deep Convolutional Neural Network for Bearing Fault Classification. *IEEE Access* **2021**, *9*, 25189–25199. [CrossRef]
23. 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]
24. Shao, H.; Jiang, H.; Wang, F.; Wang, Y. Rolling Bearing Fault Diagnosis Using Adaptive Deep Belief Network. *ISA Trans.* **2017**, *69*, 187–201. [CrossRef]
25. Luo, T.; Qiu, M.; Wu, Z.; Zhao, Z.; Zhang, D. Bearing Fault Diagnosis Based on Multi-Scale Spectral Images and Convolutional Neural Network. *arXiv* **2025**, arXiv:2503.21566. [CrossRef]
26. Alam, T.E.; Ahsan, M.M.; Raman, S. Multimodal Bearing Fault Classification under Variable Conditions: A 1D CNN with Transfer Learning. *Mach. Learn. Appl.* **2025**, *21*, 100682. [CrossRef]
27. Hatipoğlu, A.; Süpürtülü, M.; Yılmaz, E. Enhanced Fault Classification in Bearings: A Multi-Domain Feature Extraction Approach with LSTM-Attention and LASSO. *Arab. J. Sci. Eng.* **2024**, *50*, 10795–10812. [CrossRef]
28. Sinitsin, V.; Ibryaeva, O.; Sakovskaya, V.; Eremeeva, V. Intelligent Bearing Fault Diagnosis Method Combining Hybrid CNN-MLP Model. *Mech. Syst. Signal Process.* **2022**, *180*, 109454. [CrossRef]
29. Chen, J.; Lin, C.; Peng, D.; Ge, H. Fault Diagnosis of Rotating Machinery: A Review and Bibliometric Analysis. *IEEE Access* **2020**, *8*, 224985–225003. [CrossRef]
30. Sahu, D.; Dewangan, R.K.; Matharu, S.P.S. An Investigation of Fault Detection Techniques in Rolling Element Bearing. *J. Vib. Eng. Technol.* **2024**, *12*, 5585–5608. [CrossRef]
31. Jamil, M.A.; Khanam, S. Influence of One-Way ANOVA and Kruskal–Wallis Based Feature Ranking. *J. Vib. Eng. Technol.* **2024**, *12*, 3101–3132. [CrossRef]
32. Ali, U.; Ramzan, U.; Ali, W.; Al-Jaafari, K.A. An Improved Fault Diagnosis Strategy For Induction Motors Using Weighted Probability Ensemble Deep Learning. *IEEE Access* **2025**, *13*, 106958–106973. [CrossRef]
33. Abbasi, M.A.; Huang, S.; Khan, A.S. Fault Detection and Classification of Motor Bearings under Multiple Operating Conditions. *ISA Trans.* **2025**, *156*, 61–69. [CrossRef]
34. Liu, B.; Yan, C.; Liu, Y.; Wang, Z.; Huang, Y.; Wu, L. Multiscale Residual Antinoise Network via Interpretable Dynamic Recalibration Mechanism for Rolling Bearing Fault Diagnosis With Few Samples. *IEEE Sens. J.* **2023**, *23*, 31425–31439. [CrossRef]
35. Babiker, A.; Yan, C.; Li, Q.; Meng, J.; Wu, L. Initial Fault Time Estimation of Rolling Element Bearing by Backtracking Strategy, Improved VMD and Infogram. *J. Mech. Sci. Technol.* **2021**, *35*, 425–437. [CrossRef]
36. Li, F.; Zhang, B.; Verma, S.; Marfurt, K.J. Seismic Signal Denoising Using Thresholded Variational Mode Decomposition. *Explor. Geophys.* **2017**, *49*, 450–461. [CrossRef]
37. Zhang, L.; Tang, J.; Li, G.; Chen, W. Audio Magnetotelluric Denoising via Variational Mode Decomposition and Adaptive Dictionary Learning. *J. Appl. Geophy.* **2022**, *204*, 104748. [CrossRef]
38. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Process.* **2013**, *62*, 531–544. [CrossRef]
39. Irfan, M.; Alwadie, A.S.; AlThobiani, F.; Quraishi, K.S.; Jalalah, M.; Abbass, A.; Rahman, S.; Khan, M.K.A.; Alqhtani, S. A Comparison of Machine Learning Methods for the Diagnosis of Motor Faults Using Automated Spectral Feature Extraction Technique. *J. Nondestr. Eval.* **2022**, *41*, 31. [CrossRef]
40. Wang, K.; Guo, P.; Luo, A.-L. A New Automated Spectral Feature Extraction Method and Its Application in Spectral Classification and Defective Spectra Recovery. *Mon. Not. R. Astron. Soc.* **2017**, *465*, 4311–4324. [CrossRef]
41. Tian, J.; Morillo, C.; Azarian, M.H.; Pecht, M. Motor Bearing Fault Detection Using Spectral Kurtosis-Based Feature Extraction Coupled with K-Nearest Neighbor Distance Analysis. *IEEE Trans. Ind. Electron.* **2015**, *63*, 1793–1803. [CrossRef]
42. Li, P.; Lang, Z.; Zhao, L.; Tian, G.; Neasham, J.A.; Zhang, J.; Graham, D.J. System Identification-Based Frequency Domain Feature Extraction for Defect Detection and Characterization. *NDT E Int.* **2018**, *98*, 70–79. [CrossRef]
43. Al-Fahoum, A.S.; Al-Fraihat, A.A. Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time-frequency Domains. *Int. Sch. Res. Not.* **2014**, *2014*, 730218. [CrossRef]
44. Ma, B.; Zhang, W.; Jin, Z.; Li, J.; Zhang, P.; Song, X.; Jin, B. Frequency-Aware Token-Filtered Transformer for Fine-Grained Species Recognition. *Eng. Sci.* **2026**, *39*, 2040. [CrossRef]

45. Irani, H.; De, B.; Metsis, V. WaveFormer: Wavelet Embedding Transformer for Biomedical Signals. *arXiv* **2026**, arXiv:2602.12189. [CrossRef]
46. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
47. Zhang, X.; Liu, Y.; Gong, C.; Nie, Y.; Rodriguez, J. Electric Motor Bearing Fault Noise Detection via Mel-Spectrum-Based Contrastive Self-Supervised Transformer Model. *IEEE Trans. Ind. Appl.* **2024**, *60*, 8755–8765. [CrossRef]
48. Abdollah, M.A.F.; Scoccia, R.; Aprile, M. Transformer encoder based self-supervised learning for HVAC fault detection with unlabeled data. *Build. Environ.* **2024**, *258*, 111568. [CrossRef]
49. Li, J.; Bao, Y.; Liu, W.; Ji, P.; Wang, L.; Wang, Z. Twins Transformer: Cross-Attention Based Two-Branch Transformer Network for Rotating Bearing Fault Diagnosis. *Measurement* **2023**, *223*, 113687. [CrossRef]
50. Fu, Z.; Liu, Z.; Ping, S.; Li, W.; Liu, J. TRA-ACGAN: A Motor Bearing Fault Diagnosis Model Based on an Auxiliary Classifier Generative Adversarial Network and Transformer Network. *ISA Trans.* **2024**, *149*, 381–393. [CrossRef] [PubMed]
51. Raganato, A.; Tiedemann, J. An Analysis of Encoder Representations in Transformer-Based Machine Translation. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and İnterpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; pp. 287–297.
52. Huang, J.; Yuan, S.-J.; Li, D.; Li, H. A Kernel Canonical Correlation Analysis Approach for Removing Environmental and Operational Variations for Structural Damage Identification. *J. Sound Vib.* **2023**, *548*, 117516. [CrossRef]
53. Chen, L.; Wang, K.; Li, M.; Wu, M.; Pedrycz, W.; Hirota, K. K-Means Clustering-Based Kernel Canonical Correlation Analysis for Multimodal Emotion Recognition in Human–Robot Interaction. *IEEE Trans. Ind. Electron.* **2023**, *70*, 1016–1024. [CrossRef]
54. Zhou, X.; Shen, H. Regularized Canonical Correlation Analysis with Unlabeled Data. *J. Zhejiang Univ.-Sci. A* **2009**, *10*, 504–511. [CrossRef]
55. Tuzhilina, E.; Tozzi, L.; Hastie, T. Canonical Correlation Analysis in High Dimensions with Structured Regularization. *Stat. Model.* **2023**, *23*, 203–227. [CrossRef] [PubMed]
56. Case Western Reserve University. Bearing Data Center. Available online: https://engineering.case.edu/bearingdatacenter (accessed on 12 December 2025).

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.