machines

MDPI

Article

A Comparative Study of Time–Frequency Representations for Bearing and Rotating Fault Diagnosis Using Vision Transformer

Ahmet Orhan [1], Nikolay Yordanov [2], Merve Ertarğın [3,*], Marin Zhilevski [2] and Mikho Mikhov [2]

[1] Department of Electrical and Electronics Engineering, University of Firat, Elazig 23000, Türkiye; aorhan@firat.edu.tr

[2] Faculty of Automatics, Technical University of Sofia, 1000 Sofia, Bulgaria; nikolayyordanov@tu-sofia.bg (N.Y.); mzhilevski@tu-sofia.bg (M.Z.); mikhov@tu-sofia.bg (M.M.)

[3] Department of Electrical and Electronics Engineering, University of Munzur, Tunceli 62000, Türkiye

* Correspondence: merveboydak@munzur.edu.tr

Abstract

This paper presents a comparative analysis of bearing and rotating component fault classification based on different time–frequency representations using vision transformer (ViT). Four different time–frequency transformation techniques—short-time Fourier transform (STFT), continuous wavelet transform (CWT), Hilbert–Huang transform (HHT), and Wigner–Ville distribution (WVD)—were applied to convert the signals into 2D images. A pretrained ViT-Base architecture was fine-tuned on the resulting images for classification tasks. The model was evaluated on two separate scenarios: (i) eight-class rotating component fault classification and (ii) four-class bearing fault classification. Importantly, in each task, the samples were collected under varying conditions of the other component (i.e., different rotating conditions in bearing classification and vice versa). This design allowed for an independent assessment of the model's ability to generalize across fault domains. The experimental results demonstrate that the ViT-based approach achieves high classification performance across various time–frequency representations, highlighting its potential for mechanical fault diagnosis in rotating machinery. Notably, the model achieved higher accuracy in bearing fault classification compared to rotating component faults, suggesting higher sensitivity to bearing-related anomalies.

Keywords: bearing fault classification; rotating component fault classification; short-time Fourier transform; continuous wavelet transform; Hilbert–Huang transform; Wigner–Ville distribution; vision transformer

check for updates

Academic Editor: Hui Ma

Received: 1 July 2025
Revised: 12 August 2025
Accepted: 17 August 2025
Published: 19 August 2025

Citation: Orhan, A.; Yordanov, N.; Ertarğın, M.; Zhilevski, M.; Mikhov, M. A Comparative Study of Time–Frequency Representations for Bearing and Rotating Fault Diagnosis Using Vision Transformer. *Machines* **2025**, *13*, 737. https://doi.org/10.3390/machines13080737

1. Introduction

Rotating machinery plays a vital role in many industrial applications, from manufacturing systems to power generation facilities. Among its core components, bearings and rotating components are susceptible to faults due to prolonged operation and varying load conditions. Such faults can lead to reduced system performance, unexpected downtime, and high maintenance costs, making early and reliable fault detection critically important.

One of the most widely used approaches for this purpose is vibration-based condition monitoring. Vibration signals carry rich information about the operating condition of machinery and can be analyzed in different domains: the time domain, which reflects the raw waveform; the frequency domain, which reveals spectral content; and the time–frequency domain, which captures how frequency components evolve over time. In the literature, these representations have been utilized in various studies, each showing varying levels of effectiveness depending on the type of fault being detected.

In time–frequency-based studies, spectrograms obtained through STFT [1–7] and scalograms derived from CWT [8–13] are commonly used. Additionally, alternative methods such as HHT [14–16] and WVD [17–19] have also been employed in some work.

Recent advancements in deep learning have significantly enhanced the capability of fault diagnosis systems for rotating machinery and bearings. Convolutional neural networks (CNNs) have long been employed due to their strong feature extraction and classification abilities; however, they face challenges in capturing global dependencies within vibration signals. To address this, ViTs have emerged as a promising alternative, leveraging self-attention mechanisms to capture long-range dependencies. Several studies have explored hybrid approaches that combine CNNs and ViTs to harness both local and global feature representations. For example, Mo et al. [20] proposed a fusion model integrating CNN and ViT features (FCNVT) demonstrating efficient feature extraction using synchronized wavelet transform (SWT) images and achieving 100% classification accuracy with a lightweight architecture. Similarly, Ren and Lou [21] developed an enhanced ResCAA-ViT architecture that utilized SWT, residual convolutional modules, and adaptive attention mechanisms to improve diagnostic robustness under variable operating conditions, outperforming existing state-of-the-art models on both benchmark and real-world datasets. These studies highlight the effectiveness of hybrid CNN–ViT models in improving diagnostic performance while maintaining computational efficiency.

Beyond hybrid approaches, purely transformer-based methods have also been investigated for bearing and rotating machinery fault diagnosis. Tang et al. [22] proposed an integrated ViT framework combining discrete and continuous wavelet transforms with soft voting, whichimproved classification accuracy and generalization across multiple datasets. Zhang et al. [23], to better capture temporal dynamics directly from vibration signals, proposed the time series vision transformer (TSViT), which incorporated convolutional layers with transformer encoders, achieving nearly perfect accuracy across diverse operating conditions. He et al. [24], addressing data scarcity and complex working environments, proposed a Siamese vision transformer model to efficiently extract discriminative features under limited-training-data scenarios. This model introduced a novel bidirectional Kullback–Liebler divergence-based loss function and a random mask training strategy, demonstrating strong cross-domain generalization. Collectively, these studies underline the growing importance of ViT-based architectures in intelligent fault diagnosis and motivate further exploration of transformer models for robust, accurate, and practical machinery health monitoring.

In this study, the classification of bearing and rotating component faults was comparatively evaluated using various time–frequency representations. The raw signals were transformed into visual formats using four different time–frequency transformation techniques: STFT, CWT, HHT, and WVD. These images were then fed into a ViT, a cutting-edge deep learning architecture originally developed for image classification tasks. The ViT-Base model was specifically employed, as it has not yet been widely explored in the context of bearing and rotating component fault classification. To the best of our knowledge, this is one of the first studies to systematically evaluate fault classification across varying bearing and rotating component conditions, integrating four classical time–frequency transformations with a ViT framework rarely explored in prior literature.

The model's performance was evaluated in two separate classification tasks: (i) a four-class bearing fault classification, including healthy (H), ball fault (B), inner race fault (IR), and outer race fault (OR); and (ii) an eight-class rotating component fault classification, covering healthy (H), looseness (L), three severity levels of misalignment (M1, M2, M3), and three severity levels of imbalance (U1, U2, U3). In the bearing fault task, each sample corresponded to a specific bearing condition, but the data were collected under various rotating component states (e.g., L, U1, etc.). Similarly, in the rotating component fault task, each sample represented a specific rotating fault type while being recorded under different bearing fault conditions (e.g., B, IR, etc.). This experimental design enabled the assessment of the model's ability to classify one type of fault independently of the other and provided a comprehensive comparison of representation–model interactions for robust fault classification under complex operating scenarios.

2. Time–Frequency Representation

A signal is typically analyzed within two fundamental domains: the time domain and the frequency domain. In the time domain, the focus is on examining how the amplitude of the signal varies over time, providing insights into the signal's temporal behavior. However, this approach is insufficient for identifying the specific frequency components present within the signal. Conversely, frequency domain analysis, commonly performed using the Fourier transform (FT), reveals the complete spectral content of the signal, allowing for a detailed assessment of its frequency characteristics. Despite its effectiveness in uncovering frequency information, Fourier analysis inherently loses the temporal localization of these components, making it challenging to determine when certain frequencies occur within the signal. The FT is defined as [25]:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \tag{1}$$

where $X(f)$ represents the Fourier transform of the signal, indicating its frequency spectrum, f denotes the frequency variable in Hz, t represents time, j is the imaginary unit, and dt indicates integration over the entire time domain from $-\infty$ to $+\infty$.

This transformation reveals the frequency components present in the signal over its entire time span. However, while it provides a comprehensive view of the signal's spectral content, it lacks the ability to capture how these frequency characteristics evolve over time. This limitation makes it insufficient for analyzing non-stationary signals where frequency components change dynamically.

To address this limitation, the concept of time–frequency representation was introduced, aiming to analyze signals in both the time and frequency domains simultaneously. This idea was initially explored in the 1940s through the pioneering work of Gabor [26] and Ville [27]. Gabor approached the time–frequency plane by interpreting signals as composed of discrete units of information, while Ville focused on capturing the energy distribution within this domain. Their contributions laid the foundation for modern time–frequency analysis techniques, which are now essential in understanding complex, time-varying signals.

2.1. Spectrogram

In time series analysis, it is often essential to examine both the time and frequency characteristics of a signal simultaneously. This is particularly important in fields such as vibration analysis, audio processing, biomedical signal interpretation, and fault detection. One of the most widely used tools for such analysis is the spectrogram, which visually represents how the frequency content of a signal evolves over time.

Spectrograms are generated using the STFT, which analyzes non-stationary signals by dividing them into short, overlapping time segments and applying the Fourier transform to each segment. This method provides a time–frequency representation of the signal. Mathematically, the STFT is defined as [25]:

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

where $X(t,f)$ is the STFT of the signal, representing its time–frequency representation, $x(\tau)$ is the original time-domain signal to be analyzed, and $w(\tau - t)$ is the window function that localizes the signal in the time domain, centered on time t. t denotes the time-shift parameter, indicating the current position of the window, and τ is the integration variable corresponding to time. Formally, the spectrogram is computed as the squared magnitude of the STFT, representing the signal's energy distribution in the time–frequency domain.

One of the key parameters in STFT is the window length, which determines the trade-off between time and frequency resolution. A shorter window offers better time resolution, making it ideal for detecting sudden changes or transient events. However, it reduces frequency resolution, making it difficult to distinguish closely spaced frequencies. Conversely, a longer window improves frequency resolution, but limits the ability to detect rapid changes. This trade-off reflects the uncertainty principle in time–frequency analysis. The choice of window length is problem-specific and depends on factors such as the signal duration, sampling frequency, and the spectral characteristics of the signal. In addition to window length, the overlap between adjacent windows also affects the quality of the spectrogram. Higher overlap helps maintain continuity and reduces artifacts.

A comprehensive search was conducted on the Web of Science (WOS) database using a keyword combination of ("bearing" OR "bearings" OR "rotating machinery") AND ("machine learning" OR "deep learning" OR "neural network") AND ("fault diagnosis" OR "fault detection" OR "fault classification") AND ("spectrogram" OR "short-time Fourier transform" OR "STFT"), with the aim of examining how the STFT technique has been applied to motor fault diagnosis in the existing literature. The search was limited to the title, abstract, and keywords of the publications, covering the period from 2013 to 2025. As a result, 67 journal articles and 13 conference proceedings were identified. Some of these studies are presented in Table 1.

Table 1. Summary of studies utilizing spectrogram images for bearing fault classification.

Ref.	Dataset	Faults	Method	Results

[1]	CWRU [28]
	MFPT [29]
		CWRU: H (healthy), IR (inner race), OR (outer race), B (ball)
		MFPT: H, IR, OR
			CNN	CWRU = 100%
				MFPT = 99.96%

[2]	Paderborn [30]	H, IR, OR	CNN	97.48%
[3]	CWRU	H, IR, OR, B	DRNN	99.86%, 99.91%, 99.88%
[4]	Private	H, IR, OR, B, C (cage)	LAMSTAR Neural Network	96% to 100%
[5]	CWRU	H, three different damage diameters for B, IR and OR	CatGAN	91.89%
[6]	CWRU Yanshan University dataset (YSU)	CWRU: H, three different damage diameters for B, IR and OR YSU: H, IR, OR, B	CNN	CWRU: 92.67 ± 4.28–99.32 ± 0.55% YSU: 97.81 ± 0.71%
[7]	Private	H, IR, OR, B	Stacked sparse autoencoder	%96.29

In these studies, datasets such as the Case Western Reserve University (CWRU) [28], Machine Failure Prevention Technology (MFPT) [29], and Paderborn [30] have been utilized. For input data, vibration signals [1–3,5,6], acoustic emission [4], and sound [7] were commonly preferred. Various deep learning models have been employed, including CNNs [1,2,6], deep residual neural networks (DRNNs) [3], large memory storage and retrieval (LAMSTAR) neural networks [4], categorical generative adversarial networks (CatGANs) [5], and stacked sparse autoencoders [7]. These studies collectively highlight the diversity of data sources and model architectures used in spectrogram-based fault classification.

2.2. Scalogram

While STFT provides a fixed-resolution time–frequency analysis, it faces limitations in balancing time and frequency resolution due to its constant window size. In contrast, CWT offers a flexible, multi-resolution approach by adapting the analysis window size according to frequency. This makes CWT particularly well-suited for analyzing non-stationary signals that contain both transient and long-duration components [31,32].

CWT operates by convolving the signal with scaled and shifted versions of a selected mother wavelet function. The CWT of a signal $x(t)$ is defined as [32]:

$$\text{CWT}(a,b) = \int_{-\infty}^{\infty} x(t) \cdot \frac{1}{\sqrt{|a|}} \Psi\left(\frac{t-b}{a}\right) dt \tag{3}$$

where a is the scale parameter controlling frequency content, b is the translation parameter controlling time localization, and $\psi(t)$ is the mother wavelet. CWT provides a multi-resolution analysis that allows for simultaneous localization in time and frequency, making it particularly effective for non-stationary signals such as vibration data, used in this study. Smaller scales (low a) capture high-frequency, short-duration events, while larger scales (high a) reveal low-frequency, long-duration patterns.

The output of CWT is a complex-value matrix of coefficients that describe the signal's similarity to the wavelet at various scales and time points. The scalogram is obtained by plotting the squared magnitude of these coefficients in a two-dimensional image, with time on the x-axis, scale on the y-axis, and color intensity representing amplitude. This representation enables capturing localized transient features that may not be visible using time-only or frequency-only analysis, providing an enhanced representation for fault diagnosis.

A comprehensive search was conducted on the WOS database using a keyword combination of ("bearing" OR "bearings" OR "rotating machinery") AND ("machine learning" OR "deep learning" OR "neural network") AND ("fault diagnosis" OR "fault detection" OR "fault classification") AND ("scalogram" OR "continuous wavelet transform" OR "CWT"), with the aim of examining how the CWT technique has been applied to motor fault diagnosis in the existing literature. The search was limited to the title, abstract, and keywords of the publications, covering the period from 1997 to 2025. As a result, 118 journal articles and nine conference proceedings were identified. Some of these studies are presented in Table 2.

Table 2. Summary of studies utilizing scalogram images for bearing fault classification.

Ref.	Dataset	Faults	Method	Results
[8]	MFPT	H, IR, OR	Local binary convolutional neural network (LBCNN)	99.56 ± 0.97
[9]	CWRU XJTU-SY	CWRU: H, IR, OR, B XJTU-SY: C, IR, OR	CNN-gcForest	CWRU: 98.24% to 99.79% XJTU-SY: 99.8%

Table 2. *Cont.*

Ref.	Dataset	Faults	Method	Results
[10]	Private	H, IR, OR, B, LB (lack of lubrication), dual faults, multiple faults	CNN	99.39% to 99.97%
[11]	CWRU	H, IR, OR, B	Gauss convolutional deep belief network (CDBN)	Four classes:99.579% Ten classes: 99.028%
[12]	CWRU	H, IR, OR, B	LeNet-5-LSTM	99.6%
[13]	CWRU MFPT	CWRU: H, IR, OR, B MFPT: H, IR, OR	CWMS-GAN	CWRU:99.83% MFPT: 97.94%

2.3. Hilbert Spectrum

The HHT [33] is a time–frequency analysis method designed specifically for non-linear and non-stationary signals. It consists of two main stages: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA).

In the first stage, the input signal is decomposed into a finite set of intrinsic mode functions (IMFs) using EMD. Each IMF is a simple oscillatory mode that satisfies specific mathematical criteria, allowing it to represent a single frequency component with well-behaved amplitude and frequency variations over time. Once the signal is decomposed into its IMFs, each component is subjected to Hilbert transform to obtain its instantaneous frequency and amplitude.

The Hilbert transform of an IMF $c(\tau)$ is defined as:

$$\hat{c}(t) = \frac{1}{\pi} \mathrm{P} \cdot \mathrm{V} \cdot \int_{-\infty}^{\infty} \frac{c(\tau)}{t-\tau} d\tau \quad (4)$$

where $\hat{c}(t)$ is the Hilbert transform of $c(\tau)$ and P·V denotes the Cauchy principal value. The analytic signal is then formed as:

$$z(t) = c(t) + j\hat{c}(t) = A(t)e^{j\theta(t)} \quad (5)$$

Here, $A(t)$ is the instantaneous amplitude and $\theta(t)$ is the instantaneous phase, from which the instantaneous frequency can be derived as:

$$\omega(t) = \frac{d\theta(t)}{dt} \quad (6)$$

By combining the instantaneous amplitudes and frequencies of all IMFs, the Hilbert spectrum is constructed. It provides a detailed time–frequency–amplitude representation of the signal and is especially valuable for characterizing transient and non-linear behaviors.

A comprehensive search was conducted on the WOS database using a keyword combination of (“bearing” OR “bearings” OR “rotating machinery”) AND (“machine learning” OR “deep learning” OR “neural network”) AND (“fault diagnosis” OR “fault detection” OR “fault classification”) AND (“Hilbert spectrum” OR “Hilbert-Huang transform” OR “HHT”). The search was limited to the title, abstract, and keywords of the publications, covering the period from 2014 to 2025. As a result, 16 journal articles and four conference proceedings were identified. Some of these studies are presented in Table 3.

Table 3. Summary of studies utilizing Hilbert–Huang transform for bearing fault classification.

Ref.	Dataset	Faults	Method	Results
[14]	CWRU	H, IR, OR, B	Elman neural network	%100
[15]	Private	H, three different fault severities for each of the B, IR, and OR	Extreme learning machine	99.04% to 100%
[16]	University of Ottawa [34]	H, IR, OR, B, combined	VMD-DenseNet	92%

2.4. Wigner–Ville Spectrum

The WVD is a quadratic time–frequency representation that offers high resolution in both time and frequency domains. Unlike linear transforms, WVD is a member of the Cohen class of distributions and provides an energy-preserving representation of a signal's instantaneous frequency content. It is particularly effective for analyzing non-stationary signals with rapid frequency variations or multi-component structures.

Mathematically, the Wigner–Ville distribution of a signal $x(t)$ is defined as [35]:

$$W_x(t,f) = \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau \tag{7}$$

Here, $x^*(t)$ denotes the complex conjugate of the signal $x(t)$, t is time, f is frequency, and τ is the lag variable. The result is a real-valued function $W_x(t,f)$ that describes the energy distribution of the signal over the time–frequency plane.

The WVD has the ability to localize signal components with great precision, even in the presence of rapid frequency modulations. It directly reflects the signal's instantaneous power, making it suitable for detailed signal characterization. However, because of its quadratic nature, the WVD may also produce cross-terms when analyzing multi-component signals, which can complicate the interpretation of the resulting time–frequency distribution. These cross-terms are mathematical interference artifacts that arise due to the bilinear structure of the transform.

A comprehensive search was conducted on the WOS database using a keyword combination of ("bearing" OR "bearings" OR "rotating machinery") AND ("machine learning" OR "deep learning" OR "neural network") AND ("fault diagnosis" OR "fault detection" OR "fault classification") AND ("Wigner-Ville distribution" OR "WVD"). The search was limited to the title, abstract, and keywords of the publications, covering the period from 2013 to 2025. As a result, seven journal articles and three conference proceedings were identified. Some of these studies are presented in Table 4.

Table 4. Summary of studies utilizing Wigner–Ville distribution for bearing fault classification.

Ref.	Dataset	Faults	Method	Results
[17]	CWRU	H, three different fault severities for each of the B, IR, and OR	Meta-transfer-learning and original relational network (MTLRN-AM)	98%
[18]	CWRU	H, IR, OR, B	QIM-NWNN	97.5%
[19]	CWRU	H, two different fault severities for each of the B, IR, and OR	Deep echo state network based on fixed convolution kernels: FCK-DESN	95.43%

3. Materials and Methods

3.1. Dataset

In this study, a multi-domain vibration dataset under compound machine fault scenarios was utilized [36]. This dataset provides a comprehensive collection of vibration signals obtained using a deep groove ball bearing (MOCHU 6204) under various fault conditions for fault diagnosis in rotating machinery. The dataset includes three different singular bearing faults, seven different singular rotating component faults, and 21 combined fault scenarios [37]. Data were collected at rotational speeds of 600, 800, 1000, 1200, 1400, and 1600 RPM, with sampling rates of 8 kHz and 16 kHz, and for different bearing types. Each vibration signal was recorded for 160 s at an 8 kHz sampling rate and 80 s at a 16 kHz sampling rate, with each recording containing a total of 1,280,000 samples. The dataset was structured hierarchically based on sampling rate and rotational speed, with 32 data files available for each speed category.

The dataset defines various conditions that allow for the examination of rotating components and bearings under different fault scenarios. For rotating components, the system categorizes faults as H, M, U, and L. Misalignment faults include three severity levels (M1, M2, M3), corresponding to shaft displacements of 0.6 mm, 0.8 mm and 1.0 mm, respectively. Similarly, imbalance faults have three severity levels (U1, U2, U3), corresponding to additional masses of 3 g, 4 g and 5 g attached to the rotor disk, respectively. Bearing conditions are classified into H, B, IR, and OR.

In this study, data collected at a 16 kHz sampling frequency and 1000 RPM rotational speed were utilized. The higher sampling frequency of 16 kHz compared to 8 kHz was chosen to ensure that high-frequency components associated with faults could be accurately captured without aliasing, thereby preserving critical fault signatures. The rotational speed of 1000 RPM was selected as it represents an average operating condition within the dataset, providing a balanced scenario where fault-related vibration amplitudes are sufficiently pronounced for reliable detection while avoiding the excessive noise and harmonic distortions observed at higher speeds and the weak fault signatures typical of lower speeds.

Rotating component and bearing faults were identified separately. For the classification of rotating component faults, all data files containing the same type of rotating component fault were combined. This dataset included both different healthy bearing data and data with various bearing faults. In other words, while detecting rotating component faults, the bearing condition in the dataset varies: some data contain healthy bearings, while others include ball faults, inner ring faults, or outer ring faults. Similarly, for the classification of bearing faults, all data files containing the same type of bearing fault were combined. This dataset included records with different rotating component faults or healthy rotating components. That is, while identifying bearing faults, the condition of rotating components varied, with some data containing entirely healthy rotating components, while others included different faults such as misalignment, imbalance, or mechanical looseness.

Subsequently, various transformation algorithms were applied to the dataset to obtain different time–frequency representations for feature extraction based on time–frequency analysis. First, STFT was employed to analyze the frequency components of signals within specific time intervals, generating spectrogram images. Then, CWT was utilized to produce scalogram images, offering an adaptive frequency resolution. Additionally, WVD transformation was applied to visualize the signal's autocorrelation-based analysis, resulting in the Wigner–Ville spectrum. Finally, the HHT was used to determine the instantaneous frequency components of the signals, producing the Hilbert spectrum. The time–frequency images obtained through these transformations were used to train a deep learning-based model for machine fault diagnosis and condition monitoring.

In the time–frequency analysis conducted using STFT, a sampling frequency of 16,000 Hz was utilized. The Hann window, which is provided as the default option in the scipy.signal.spectrogram function and a built-in method in the SciPy library used to compute a spectrogram via STFT [38], was employed, with the window length set to

256 samples. To maintain temporal continuity between successive frames, an overlap of 32 samples—equivalent to one-eighth of the window length—was applied. The length of the fast Fourier transform (FFT) was configured to match the window length, specifically 256 points, in order to achieve a compromise between computational efficiency and frequency resolution. The spectrograms that were generated through this process were subsequently log-scaled and visualized using Gouraud shading, which was applied to enhance the smoothness and clarity of the time–frequency representation.

In the scalogram-based time–frequency analysis, CWT was carried out using the Morlet wavelet, which is known for providing a favorable trade-off between time and frequency localization. Each segment of the signal was composed of 16,000 samples. A range of scales, corresponding to wavelet widths varying from 1 to 30, was employed in the analysis. This range of scales was selected to enable the effective extraction of both low-frequency and high-frequency components from the signal, thereby ensuring a comprehensive representation of its spectral content.

For the time–frequency analysis based on the HHT, EMD was applied using the mask sift approach, with the decomposition limited to a maximum of five intrinsic IMFs. Each input segment was composed of 16,000 data samples, thereby maintaining consistency with the analyses conducted using the STFT and CWT methods. After the decomposition process was completed, the normalized Hilbert transform was subsequently applied to each extracted IMF in order to derive instantaneous frequency and amplitude information. The resulting time–frequency representations were then constructed using 150 frequency bins that were logarithmically spaced across a range from 1 Hz to 8000 Hz. This configuration was chosen to enable a detailed and comprehensive characterization of both low-frequency and high-frequency components present in the signal.

For the WVD analysis, each original signal segment consisting of 16,000 samples was divided into smaller subsegments of 2000 samples in order to reduce the computational and memory demands typically associated with generating the full WVD matrix. If the WVD were to be computed over the entire segment, a matrix of size 16,000 × 16,000—containing 256 million elements—would be required, which is considered impractical due to significant memory constraints. By employing subsegments of 2000 samples, the matrix size was effectively reduced to 2000 × 2000, resulting in only 4 million elements and thereby allowing for more efficient and feasible computation. The time–frequency representations that were obtained through this method were subsequently normalized by scaling the absolute values to the [0, 1] range, ensuring consistency, comparability, and interpretability across different signal samples.

The image samples obtained are presented in Figure 1. Among these, the WVD representation initially resulted in a much larger dataset due to the subsegmentation strategy applied to handle its quadratic time and memory complexity. To ensure a fair comparison with the STFT, CWT, and HHT representations, an equal number of samples were randomly selected from the WVD dataset to create the training, validation, and testing sets. Consequently, for all four methods, a total of 2560 samples were used for each representation, distributed as 1792 for training, 384 for validation, and 384 for testing.

3.2. Method

The overall structure of the ViT-based fault classification approach used in this study is illustrated in Figure 2. Vibration signals are first transformed into time–frequency representations using one of four methods: STFT, CWT, HHT or WVD. The resulting 2D image is resized to 224 × 224 pixels and fed into a ViT architecture. The input image is divided into non-overlapping patches of size 16 × 16 pixels, flattened, and linearly projected. Positional embeddings are then added, and the embedded patches are processed through transformer encoder blocks consisting of multi-head self-attention and feed-forward (MLP) layers.

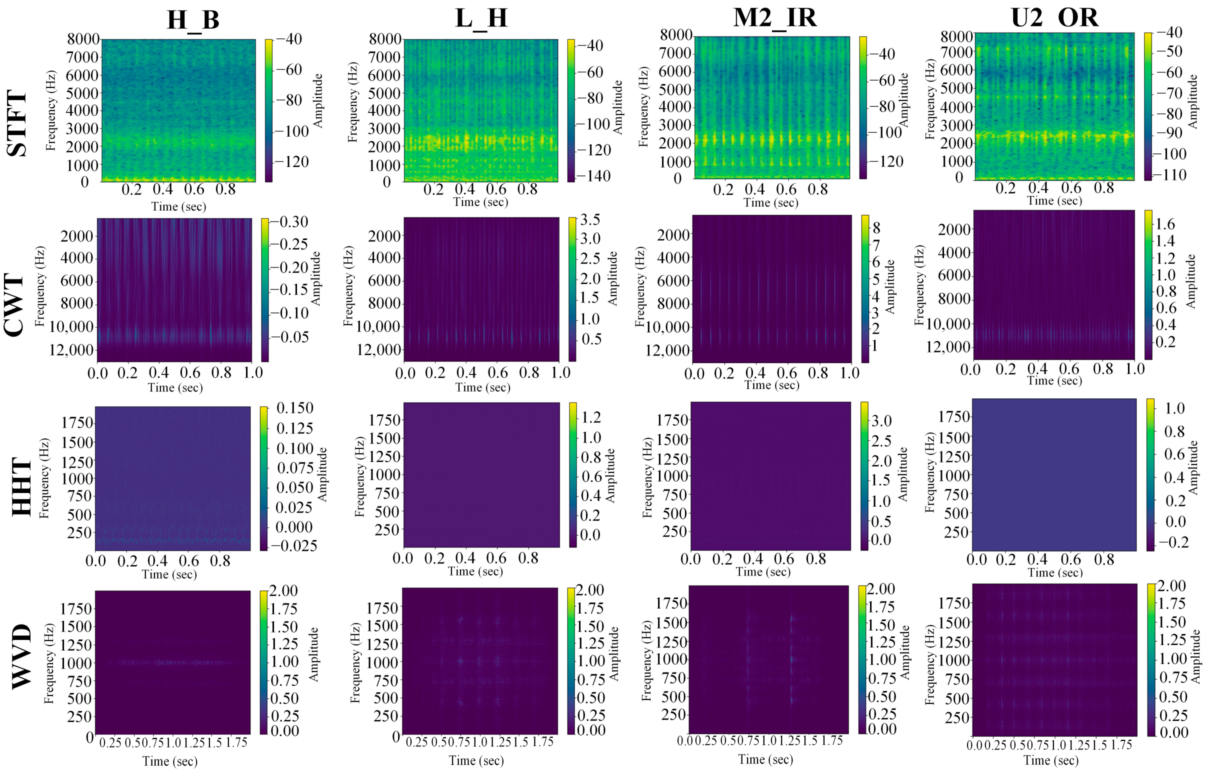

Figure 1. Samples of images in the dataset.

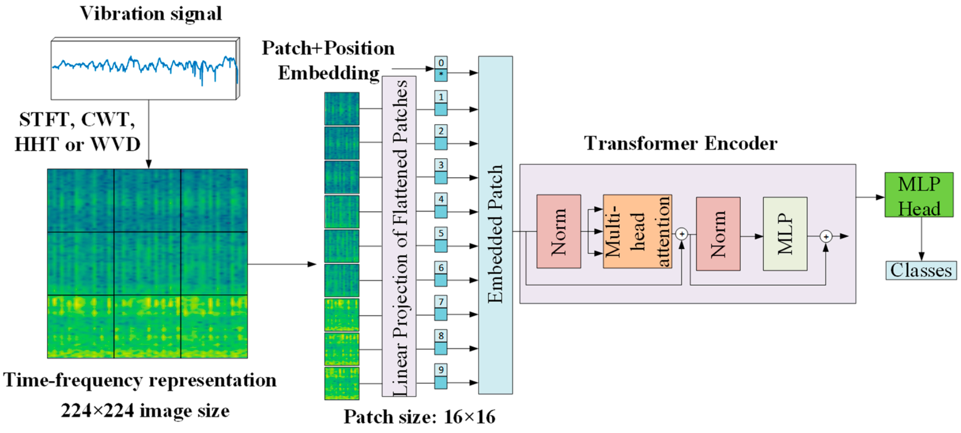

Figure 2. Overall structure of the ViT-based fault classification approach.

To provide additional clarity, the detailed internal mechanism of the multi-head self-attention block is illustrated in Figure 3. In this mechanism, each embedded patch is first linearly projected into three distinct representations: queries (Q), keys (K), and values (V). The dot product between Q and K determines the similarity (attention score) between patches, indicating how much focus each patch should receive relative to others. These scores are scaled and passed through a softmax function to obtain attention weights, which are then used to combine the V vectors. Multiple attention heads perform this operation in parallel, capturing diverse relationships among patches. The outputs from all heads are concatenated and passed through a linear transformation to generate the final attention output. This process allows the model to effectively learn global dependencies between different fault-related features in the time–frequency representations.

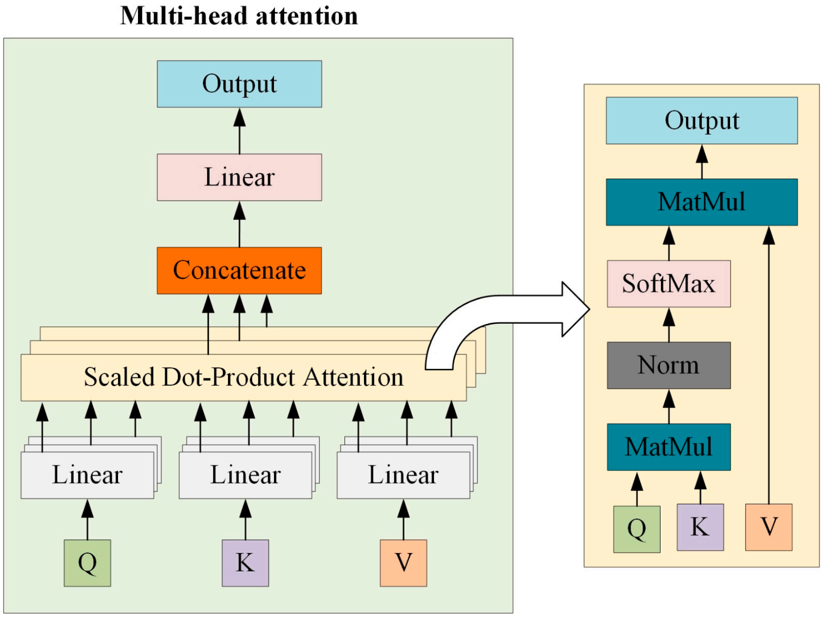

Figure 3. Detailed structure of the multi-head self-attention mechanism within the ViT-based model.

The output token is finally passed to a classification head (MLP head) to predict the corresponding fault class. A pre-trained ViT-Base model was employed, and a transfer learning strategy was applied. To reduce computational cost and training time, all layers were frozen except for the final classification head, which was fine-tuned on the target dataset. For visualization clarity, only 9 patches are shown in the figure. In practice, the input image is divided into 14 × 14 = 196 patches.

To evaluate the model's performance, two independent classification tasks were conducted. In the first task, an eight-class rotating component fault classification (H, L, M1, M2, M3, U1, U2, U3) was performed, and in the second a four-class bearing fault classification (H, B, IR, OR) was carried out. In both tasks, the samples were intentionally constructed to include varying conditions of the other component. That is, bearing fault samples were collected under different rotating component states, while rotating component fault samples were collected under varying bearing conditions. This structure ensured that the model learned to identify fault types independently of variations in other mechanical components, allowing for a more realistic and robust evaluation.

4. Experimental Results

ViT-based architecture was utilized in this study for image-based fault classification. The pretrained "vit-base-patch16-224" model from the Hugging Face library was employed, with only the final classification layer fine-tuned. The optimizer used was Adam with a learning rate of 0.001, and cross-entropy was selected as the loss function. The batch size was set to 32. An "EarlyStopping" mechanism with a patience of five epochs was implemented to stop the training process once overfitting was detected. Although the primary cause of overfitting is the complexity of the model rather than the training duration, the inclusion of this mechanism aimed to prevent further degradation in validation performance. The dataset was split in a balanced manner across classes, with 70% of the data used for training, 15% for validation, and 15% for testing.

4.1. STFT Results

Performance metrics on the test dataset using spectrograms are presented in Figure 4. The figure reports precision, recall, F1-score, and accuracy values for two main fault types:

rotating component faults and bearing faults. For rotating component faults, all metrics are reported as 88%, whereas for bearing faults, these metrics reach a notably high level of 99%. This indicates that the spectrogram-based approach yields more successful results in detecting bearing faults.

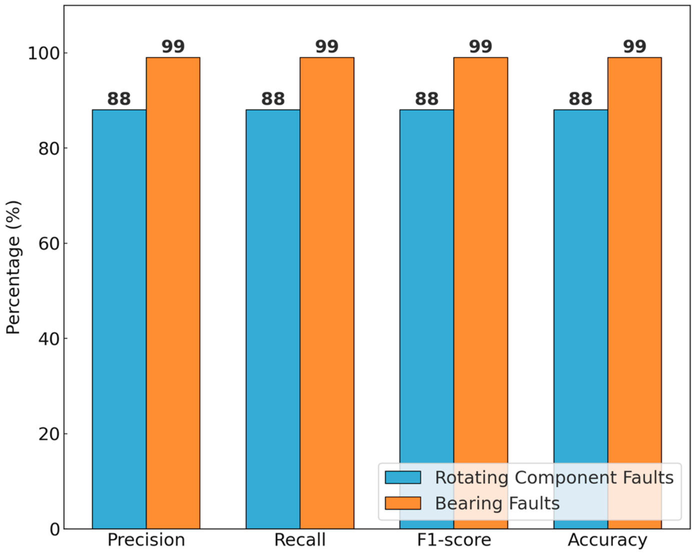

Figure 4. Performance metrics on test dataset for spectrogram.

Table 5 displays the confusion matrix for the classification of rotating component faults. A closer examination reveals that different severity levels of the same fault type (e.g., M1, M2, M3 or U1, U2, U3) are more frequently confused with one another. For instance, six instances of the U1 class were misclassified as U2 and four as U3. Similarly, four instances from the U3 class were predicted as U2. This suggests that the model struggles to distinguish between different severity levels within the same fault type. In contrast, different fault types (e.g., H and L, or the M-series and U-series faults) are generally well classified, indicating a clearer separation between these categories in the feature space.

Table 5. Confusion matrix for rotating component fault classification and spectrogram.

		Predicted Label							
		H	**L**	**M1**	**M2**	**M3**	**U1**	**U2**	**U3**
Actual label	**H**	45	0	0	0	0	0	3	0
	L	0	48	0	0	0	0	0	0
	M1	1	0	44	2	0	0	0	1
	M2	0	0	0	40	6	0	2	0
	M3	0	0	0	4	44	0	0	0
	U1	3	0	0	0	0	35	6	4
	U2	3	1	1	0	0	1	40	2
	U3	3	0	0	0	0	0	4	41

Table 6 presents the confusion matrix for the classification of bearing faults. All classes exhibit a high number of correctly classified instances. This demonstrates that bearing faults present more distinct characteristics compared to other classes, allowing the model to differentiate them more easily.

Table 6. Confusion matrix for bearing fault classification and spectrogram.

		Predicted Label			
		B	**H**	**IR**	**OR**
Actual label	**B**	95	1	0	0
	H	1	95	0	0
	IR	0	0	96	0
	OR	0	0	0	96

4.2. CWT Results

In Figure 5, the performance metrics on the test dataset obtained using scalograms are presented. Precision, recall, F1-score, and accuracy values are reported for two main fault types: rotating component faults and bearing faults. For rotating component faults, the metrics are around 86%–87%, whereas for bearing faults, all values reach 97%. These results indicate that the scalogram-based method is particularly effective in detecting bearing faults, while its performance on rotating component faults is comparatively lower.

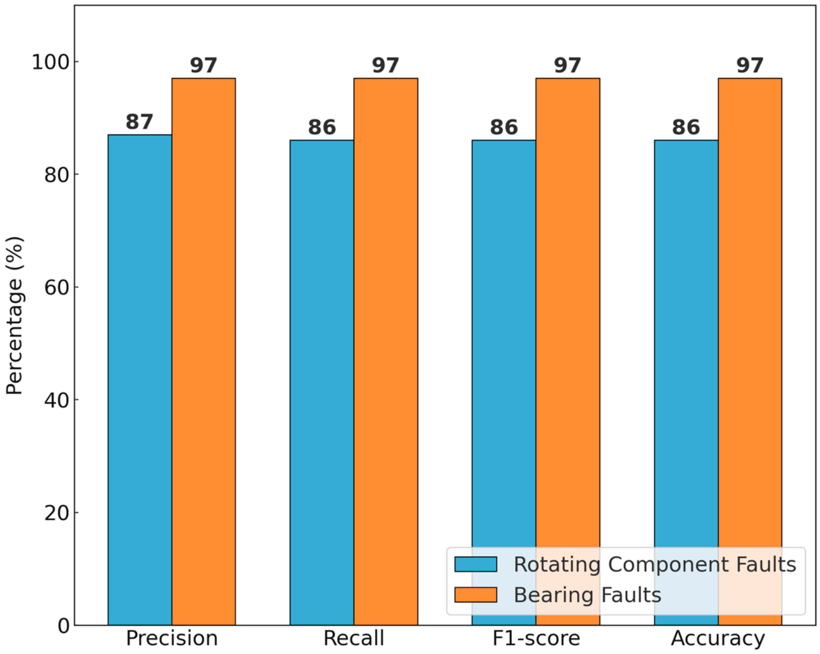

Figure 5. Performance metrics on test dataset for scalogram.

Table 7 presents the confusion matrix generated by the scalogram-based model for the classification of rotating component faults. Similar to previous findings, different severity levels of the same fault type tend to be misclassified among each other. For instance, six samples from class M2 were misclassified as M3, and 4 as M1. Likewise, 4 samples from class U1 were predicted as U2 and six as U3. In the case of class U3, there is notable confusion with U2, with 10 samples incorrectly classified. These results highlight the model's ongoing difficulty in distinguishing between varying severity levels within the same fault type. In contrast, clearly distinct fault types such as H and L were classified with high accuracy.

Table 7. Confusion matrix for rotating component fault classification and scalogram.

		Predicted Label							
		H	**L**	**M1**	**M2**	**M3**	**U1**	**U2**	**U3**
Actual label	**H**	48	0	0	0	0	0	0	0
	L	0	48	0	0	0	0	0	0
	M1	0	0	46	1	0	0	1	0
	M2	0	0	4	35	6	0	1	2
	M3	0	0	0	4	42	0	0	2
	U1	0	0	0	0	1	37	4	6
	U2	0	2	0	1	1	1	41	2
	U3	0	1	0	0	0	2	10	35

Table 8 displays the confusion matrix for bearing fault classification using the scalogram-based approach. Although three samples of class B were misclassified as H, and seven samples of H were classified as B. Overall, the scalogram-based method still achieved high accuracy in detecting bearing faults.

Table 8. Confusion matrix for bearing fault classification and scalogram.

		Predicted Label			
		B	**H**	**IR**	**OR**
Actual label	**B**	92	3	0	1
	H	7	89	0	0
	IR	0	0	96	0
	OR	0	0	0	96

4.3. HHT Results

In Figure 6, the performance metrics on the test dataset obtained using the Hilbert spectrum are presented. For rotating component faults, the accuracy, precision, recall, and F1 score remain within the range of 69%–71%, while for bearing faults, these metrics reach 90%. Similarly to the other methods, the HHT-based approach demonstrates higher effectiveness in detecting bearing faults, but shows limited capability in identifying rotating component faults. This limitation can be attributed to the EMD process used in HHT, which is susceptible to mode mixing and noise sensitivity [39–41]. Consequently, these shortcomings hinder its ability to extract sufficiently discriminative features, leading to lower classification accuracy, particularly for rotating component faults.

Table 9 presents the confusion matrix for the classification of rotating component faults using the HHT-based model. The matrix reveals a high level of confusion between classes. In particular, for classes M2 and U1, only 26 out of 48 samples were correctly classified, highlighting the model's difficulty in distinguishing between these fault types.

Table 9. Confusion matrix for rotating component fault classification and HHT.

		Predicted Label							
		H	**L**	**M1**	**M2**	**M3**	**U1**	**U2**	**U3**
Actual label	**H**	31	4	3	1	0	1	5	3
	L	0	37	0	1	2	1	5	2
	M1	1	0	38	5	0	0	2	2
	M2	1	2	4	26	8	1	5	1

Table 9. *Cont.*

		Predicted Label							
		H	**L**	**M1**	**M2**	**M3**	**U1**	**U2**	**U3**
Actual label	**M3**	1	5	2	7	31	0	0	2
	U1	1	5	2	1	0	26	10	3
	U2	0	3	4	0	0	1	38	2
	U3	1	0	2	0	1	3	3	38

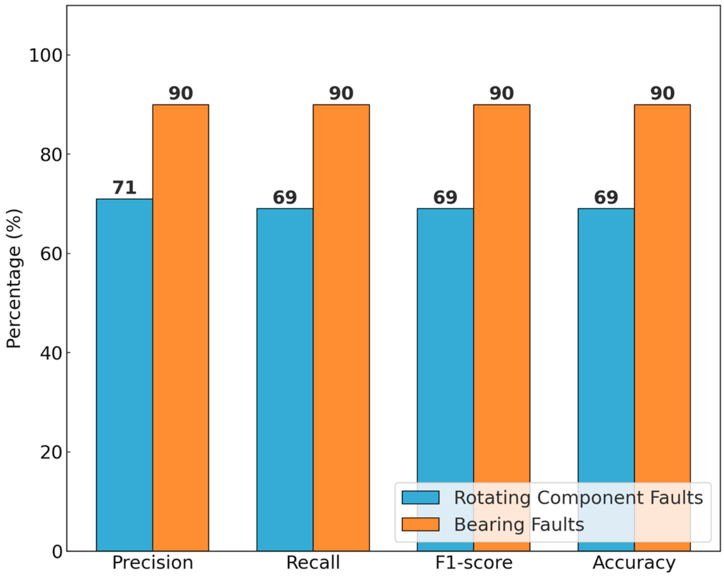

Figure 6. Performance metrics on test dataset for HHT.

Table 10 shows the confusion matrix for bearing fault classification. In this case, the model exhibited its poorest performance when classifying samples from class H. Specifically, 12 samples from B were classified as H, 2 as IR, and 3 as OR.

Table 10. Confusion matrix for bearing fault classification and HHT.

		Predicted Label			
		B	**H**	**IR**	**OR**
Actual label	**B**	79	12	2	3
	H	6	88	1	1
	IR	2	0	85	9
	OR	2	0	2	92

4.4. WVD Results

Figure 7 presents the performance metrics of the WVD-based approach in the classification of rotating component and bearing faults. The results indicate that the model achieves remarkably high performance in classifying bearing faults, with 98% accuracy, precision, recall, and F1 score. This demonstrates the model's effectiveness in correctly identifying positive samples of bearing faults while minimizing false positives. In contrast, the accuracy for rotating component faults remains at 76%, suggesting that the model is relatively less successful in distinguishing between these fault types.

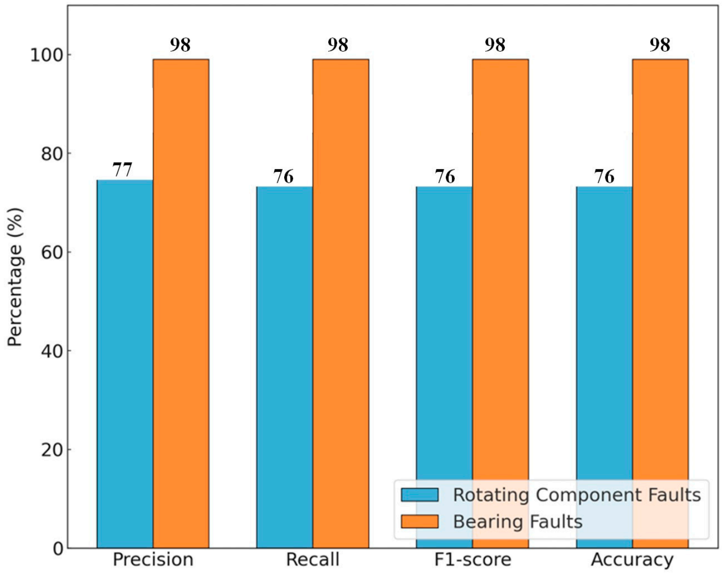

Figure 7. Performance metrics on test dataset for WVD.

As shown in Table 11 the confusion matrix for rotating component faults reveals a concentration of misclassifications within specific classes. The model demonstrates strong classification capability for the L and M1 classes, achieving 45 and 43 correctly identified samples, respectively, indicating a high level of discriminative performance for these fault types. However, significant misclassifications are observed in the M2 and U3 classes. The fact that six samples from the U3 class were predicted as U2 indicates that the model struggles to clearly distinguish between these two classes. Additionally, the misclassification of eight samples from the U3 class as H suggests that the model has not sufficiently learned the distinctive features of this class and tends to overpredict the healthy condition.

Table 11. Confusion matrix for rotating component fault classification and WVD.

		Predicted Label							
		H	**L**	**M1**	**M2**	**M3**	**U1**	**U2**	**U3**
Actual label	**H**	37	0	2	1	0	5	1	2
	L	1	45	0	0	0	1	1	0
	M1	1	1	43	1	0	1	1	0
	M2	2	1	3	28	10	2	2	0
	M3	1	0	1	2	39	3	2	0
	U1	2	0	0	0	1	43	2	0
	U2	4	0	2	0	0	7	31	4
	U3	8	1	0	1	4	1	6	27

Table 12 provides the confusion matrix for bearing faults and shows minimal classification errors among the classes. The absence of misclassifications for the OR class highlights the model's robust discriminative ability in identifying this fault type. These findings suggest that bearing faults exhibit more distinctive features, allowing the WVD-based method to perform highly effectively in this context.

Table 12. Confusion matrix for bearing fault classification and WVD.

		Predicted Label			
		B	**H**	**IR**	**OR**
Actual label	**B**	92	0	1	3
	H	2	94	0	0
	IR	2	0	94	0
	OR	0	0	0	96

5. Discussion

Table 13 presents a comparative evaluation of classification accuracies achieved by different deep learning models (ViT-Base, ResNet101, and DenseNet121) using four time–frequency representations: spectrogram, scalogram, HHT and WVD. The analysis considers two fault types separately, namely rotating component faults and bearing faults, to comprehensively assess the discriminative capability of each representation-model pair.

Table 13. Classification accuracy on test dataset for different representations.

Model	Fault Type	Spectrogram	Scalogram	HHT	WVD
ViT-Base	Rotating component faults	88%	86%	69%	76%
	Bearing faults	99%	97%	90%	98%
Res-Net101	Rotating component faults	89%	85%	77%	70%
	Bearing faults	99%	96%	94%	97%
Dense-Net121	Rotating component faults	84%	85%	74%	71%
	Bearing faults	98%	95%	94%	98%

When bearing fault classification was considered, all time–frequency representations and model architectures achieved consistently high accuracy levels. The ViT-Base model demonstrated strong and stable performance with spectrogram (99%), scalogram (97%), WVD (98%), and HHT (90%) representations. Similarly, the ResNet101 and DenseNet121 architectures also achieved accuracies above 94% across all representations, with WVD and spectrogram representations delivering the highest results, reaching 97%–99% accuracy.

When classifying rotating component faults, the overall performance was significantly lower compared to the bearing fault scenario, regardless of the representation or model employed. The observed differences in classification performance can be explained by the physical fault factors (type, location, severity, and time) and the properties of the time–frequency transformations. Bearing faults produce distinct, localized frequency components [42,43], which facilitate their separation in the time–frequency domain. In contrast, rotating component faults, particularly different severity levels of misalignment and imbalance, exhibit similar low-frequency vibration patterns and overlapping spectral signatures, leading to misclassifications among these classes.

In the classification of rotating component faults, the spectrogram and scalogram representations consistently achieved higher classification performance across all deep learning models compared to the other two time–frequency representations, HHT and WVD. In the ViT-Base model, the spectrogram and scalogram achieved accuracies of 88% and 86%, respectively, whereas HHT and WVD yielded lower accuracies of 69% and 76%. Similarly, with the ResNet101 architecture, the spectrogram and scalogram attained accuracies of 89% and 85%, while HHT and WVD achieved 77% and 70%, respectively. A similar trend was observed for DenseNet121, where the spectrogram and scalogram provided the highest accuracies of 84% and 85%, while HHT and WVD reached only 74% and 71%, respectively. These findings indicate that spectrogram and scalogram representations offer more discriminative features for the classification of rotating component faults and exhibit greater compatibility with different deep learning architectures.

In the classification of rotating component faults, WVD demonstrated higher performance than HHT in the ViT-Base model, achieving an accuracy of 76% compared to HHT's 69%. However, in the ResNet101 and DenseNet121 models, WVD achieved accuracies of 70% and 71%, respectively, while HHT delivered higher accuracies of 77% and 74%. These results indicate that WVD is relatively more advantageous than HHT when using the transformer-based ViT architecture, whereas HHT outperforms WVD in convolution-based architectures such as ResNet and DenseNet.

In a related study [44] utilizing the MFPT dataset, which involved the classification of H, IR, and OR conditions, scalogram representations achieved the highest classification accuracy of 99.9%, followed by HHT with 95.5% and spectrogram with 91.7%. Similarly, in the CWRU dataset comprising H, BF, IR, and OR classes, both spectrogram and scalogram representations achieved 99.5% accuracy, while HHT reached 97.6%. These results suggest that scalogram representations can deliver highly accurate fault diagnosis performance across various datasets.

In comparison, the results obtained in the present study using the same time–frequency representations demonstrated comparable, and in some cases even superior, classification performance. Specifically, when classifying bearing faults using the dataset employed in this study, the scalogram achieved up to 97% accuracy with ViT-Base, while the spectrogram and WVD reached 99% and 98%, respectively. HHT, on the other hand, yielded slightly lower results, ranging from 90% to 94% depending on the model used. These findings confirm the robustness of spectrogram and scalogram representations, as highlighted in the previous study, while also emphasizing the strong performance of the WVD as an alternative representation capable of delivering top-tier accuracy levels.

One limitation of this study is that the data used were collected at a sampling frequency of 16 kHz and a rotational speed of 1000 RPM. Although the dataset itself includes a variety of operating conditions, only a specific subset was utilized in this work. This choice allowed for a consistent comparative analysis of the time–frequency representations under controlled conditions, but it also restricted our ability to directly assess the model's performance under different speeds and sampling rates. In future studies, utilizing a broader portion of the dataset would enable a more comprehensive evaluation of the method's generalizability across varying operating conditions.

Another important practical consideration is the detectability of faults in real engineering systems, where early-stage or incipient defects often exhibit subtle signatures that conventional approaches may fail to recognize. In our study, the dataset included multiple severity levels for rotating component faults, which allowed us to partially assess how the ViT-based model differentiates between defect magnitudes. However, for bearing faults, the dataset did not contain varying defect sizes, which limited our ability to fully analyze model behavior for early-stage bearing anomalies. Future research should therefore incorporate datasets with progressive life-cycle data for bearings to better evaluate the robustness of time–frequency representations and transformer-based models in detecting small and evolving faults.

Additionally, the current study did not include visual interpretability analyses such as attention maps. Incorporating such analyses would contribute to a better understanding of the model's decision-making process and more clearly highlight the differences among the transformation techniques. This aspect will be addressed in future research.

6. Conclusions

This paper presented a comprehensive comparison of four time–frequency transformation techniques, STFT (spectrogram), CWT (scalogram), HHT, and WVD, for the purpose of classifying faults in rotating and bearing components using ViT-Base architecture. By converting vibration signals into two-dimensional time–frequency representations and fine-tuning a pre-trained ViT-base model on these images, the research explored the model's effectiveness in two distinct diagnostic scenarios: rotating component fault classification and bearing fault classification. A key aspect of the experimental design was the introduction of cross-condition variability where the non-target component was subjected to different fault states, thereby providing a more realistic and challenging evaluation of the model's generalization ability.

The results indicate that the ViT-Base model can successfully classify mechanical faults with high accuracy across different time–frequency representations, with particularly strong performance in the bearing fault classification task. Among the tested representations, spectrogram and scalogram consistently delivered high accuracies, while HHT yielded comparatively lower results in both fault types. Additionally, the study demonstrates that bearing faults are more readily detected than rotating component faults, likely due to their more distinctive signal characteristics. Overall, the findings suggest that ViT-based architectures, when combined with appropriate time–frequency representations, offer a powerful and flexible framework for fault diagnosis in rotating machinery.

Author Contributions: Conceptualization, A.O., M.E., and M.Z.; methodology, A.O., M.Z., and M.M.; software, N.Y. and M.E.; validation, A.O., M.Z., and M.M.; formal analysis, N.Y. and M.E.; investigation, N.Y. and M.E.; writing—review and editing, N.Y., M.E., M.Z., and M.M.; visualization, N.Y. and M.E.; supervision, A.O., M.Z., and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was accomplished with financial support by the European Regional Development Fund within the operational program "Bulgarian national recovery and resilience plan," procedure for direct provision of grants ("Establishing a network of research higher education institutions in Bulgaria"), and under Project BG-RRP-2.004-0005: "Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia (IDEAS)."

Data Availability Statement: The data used in this study are available online: Multi-domain vibration dataset with various bearing types under compound machine fault scenarios—subset 1 (deep groove ball bearing).

Acknowledgments: During the preparation of this manuscript, the authors used GenIA tools such as Google Gemini (gemini-2.0-flash-001), Microsoft Copilot (GPT-4), ChatGPT (GPT-4o), DeepSeek (DeepSeek-R1), and Grok (Grok-3) to improve the clarity and fluency of certain sentences and to enhance the descriptions of signal-to-image transformation techniques. These tools were used solely for linguistic refinement and improving expression clarity. The authors have carefully reviewed and edited the outputs from these tools and take full responsibility for the entire content of this publication.

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

References

1. Zhang, Q.; Deng, L. An Intelligent Fault Diagnosis Method of Rolling Bearings Based on Short-Time Fourier Transform and Convolutional Neural Network. *J. Fail. Anal. Prev.* **2023**, *23*, 795–811. [CrossRef]
2. Zhong, D.; Guo, W.; He, D. An Intelligent Fault Diagnosis Method based on STFT and Convolutional Neural Network for Bearings Under Variable Working Conditions. In Proceedings of the 2019 Prognostics and System Health Management Conference (PHM-Qingdao), Qingdao, China, 25–27 October 2019; pp. 1–6. [CrossRef]

3. Peng, B.; Xia, H.; Lv, X.; Annor-Nyarko, M.; Zhu, S.; Liu, Y.; Zhang, J. An intelligent fault diagnosis method for rotating machinery based on data fusion and deep residual neural network. *Appl. Intell.* **2022**, *52*, 3051–3065. [CrossRef]
4. He, M.; He, D. Deep learning based approach for bearing fault diagnosis. *IEEE Trans. Ind. Appl.* **2017**, *53*, 3057–3065. [CrossRef]
5. Tao, H.; Wang, P.; Chen, Y.; Stojanovic, V.; Yang, H. An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks. *J. Frankl. Inst.* **2020**, *357*, 7286–7307. [CrossRef]
6. Zhang, Y.; Xing, K.; Bai, R.; Sun, D.; Meng, Z. An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image. *Measurement* **2020**, *157*, 107667. [CrossRef]
7. Liu, H.; Li, L.; Ma, J. Rolling bearing fault diagnosis based on STFT-deep learning and sound signals. *Shock Vib.* **2016**, *2016*, 6127479. [CrossRef]
8. Cheng, Y.; Lin, M.; Wu, J.; Zhu, H.; Shao, X. Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform-local binary convolutional neural network. *Knowl.-Based Syst.* **2021**, *216*, 106796. [CrossRef]
9. Xu, Y.; Li, Z.; Wang, S.; Li, W.; Sarkodie-Gyan, T.; Feng, S. A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurement* **2021**, *169*, 108502. [CrossRef]
10. Mian, T.; Choudhary, A.; Fatima, S. Vibration and infrared thermography based multiple fault diagnosis of bearing using deep learning. *Nondestruct. Test. Eval.* **2023**, *38*, 275–296. [CrossRef]
11. Zhao, H.; Liu, J.; Chen, H.; Chen, J.; Li, Y.; Xu, J.; Deng, W. Intelligent diagnosis using continuous wavelet transform and gauss convolutional deep belief network. *IEEE Trans. Reliab.* **2022**, *72*, 692–702. [CrossRef]
12. Ahsan, M.; Hassan, M.W.; Rodriguez, J.; Abdelrahem, M. Enhanced Fault Diagnosis in Rotating Machinery Using a Hybrid CWT-LeNet-5-LSTM Model: Performance Across Various Load Conditions. *IEEE Access* **2024**, *13*, 1026–1045. [CrossRef]
13. Yu, S.; Li, Z.; Gu, J.; Wang, R.; Liu, X.; Li, L.; Guo, F.; Ren, Y. CWMS-GAN: A small-sample bearing fault diagnosis method based on continuous wavelet transform and multi-size kernel attention mechanism. *PLoS ONE* **2025**, *20*, e0319202. [CrossRef]
14. Liu, H.; Wang, X.; Lu, C. Rolling Bearing Fault Diagnosis under Variable Conditions Using Hilbert-Huang Transform and Singular Value Decomposition. *Math. Probl. Eng.* **2014**, *2014*, 765621. [CrossRef]
15. Suthar, V.; Vakharia, V.; Patel, V.K.; Shah, M. Detection of compound faults in ball bearings using multiscale-SinGAN, heat transfer search optimization, and extreme learning machine. *Machines* **2022**, *11*, 29. [CrossRef]
16. Lin, S.L. Intelligent fault diagnosis and forecast of time-varying bearing based on deep learning VMD-DenseNet. *Sensors* **2021**, *21*, 7467. [CrossRef]
17. Wei, P.; Liu, M.; Wang, X. Few-shot bearing fault diagnosis using GAVMD–PWVD time–frequency image based on meta-transfer learning. *J. Braz. Soc. Mech. Sci. Eng.* **2023**, *45*, 277. [CrossRef]
18. Hua, L.; Qiang, Y.; Gu, J.; Chen, L.; Zhang, X.; Zhu, H. Mechanical fault diagnosis using color image recognition of vibration spectrogram based on quaternion invariable moment. *Math. Probl. Eng.* **2015**, *2015*, 702760. [CrossRef]
19. Li, X.; Bi, F.; Zhang, L.; Lin, J.; Bi, X.; Yang, X. Rotating machinery faults detection method based on deep echo state network. *Appl. Soft Comput.* **2022**, *127*, 109335. [CrossRef]
20. Mo, C.; Huang, K.; Ji, H.; Li, W.; Xu, K. Efficient intelligent fault diagnosis method and graphical user interface development based on fusion of convolutional networks and vision transformers characteristics. *Sci. Rep.* **2025**, *15*, 7110. [CrossRef] [PubMed]
21. Ren, S.; Lou, X. Rolling Bearing Fault Diagnosis Method Based on SWT and Improved Vision Transformer. *Sensors* **2025**, *25*, 2090. [CrossRef] [PubMed]
22. Tang, X.; Xu, Z.; Wang, Z. A Novel Fault Diagnosis Method of Rolling Bearing Based on Integrated Vision Transformer Model. *Sensors* **2022**, *22*, 3878. [CrossRef]
23. Zhang, S.; Zhou, J.; Ma, X.; Pirttikangas, S.; Yang, C. TSViT: A Time Series Vision Transformer for Fault Diagnosis of Rotating Machinery. *Appl. Sci.* **2024**, *14*, 10781. [CrossRef]
24. He, Q.; Li, S.; Bai, Q.; Zhang, A.; Yang, J.; Shen, M. A Siamese Vision Transformer for Bearings Fault Diagnosis. *Micromachines* **2022**, *13*, 1656. [CrossRef]
25. Allen, J. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1977**, *25*, 235–238. [CrossRef]
26. Gabor, D. Theory of communication. Part 1: The analysis of information. *J. Inst. Electr. Eng.-Part III Radio Commun. Eng.* **1946**, *93*, 429–441. [CrossRef]
27. Ville, J. Theorie et Applications de la Notion de Signal Analytique. *Cables Transm.* **1948**, *2A*, 61–74.
28. "Apparatus & Procedures | Case School of Engineering". Available online: https://engineering.case.edu/bearingdatacenter/apparatus-and-procedures (accessed on 1 July 2025).
29. Bechhoefer, E. A Quick Introduction to Bearing Envelope Analysis, MFPT Data. Available online: www.mfpt.org/fault-data-sets (accessed on 1 July 2025).
30. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. *PHM Soc. Eur. Conf.* **2016**, *3*, 1. [CrossRef]

31. Graps, A. An introduction to wavelets. *IEEE Comput. Sci. Eng.* **1995**, *2*, 50–61. [CrossRef]
32. Polikar, R.; Mastorakis, N. The story of wavelets inPhysics and modern topics in mechanical and electrical engineering. *World Sci. Eng. Soc. Press* **1999**, 192–197.
33. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.-C.; Tungand, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
34. Huang, H.; Baddour, N. Bearing vibration data collected under time-varying rotational speed conditions. *Data Brief* **2018**, *21*, 1745–1749. [CrossRef] [PubMed]
35. Boashash, B. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*; Academic Press: Cambridge, MA, USA, 2015.
36. Lee, S.; Kim, T.; Kim, T. Multi-domain vibration dataset with various bearing types under compound machine fault scenarios: Subset 1 (deep groove ball bearing). *Mendeley Data* **2024**, *57*, 110640. [CrossRef]
37. Lee, S.; Kim, T.; Kim, T. Multi-domain vibration dataset with various bearing types under compound machine fault scenarios. *Data Brief* **2024**, *57*, 110940. [CrossRef]
38. SciPy Documentation—Scipy.signal.spectrogram. Available online: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.spectrogram.html (accessed on 12 August 2025).
39. Jabloun, M. Empirical mode decomposition revisited using ordinal pattern concepts. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 2186–2190.
40. Tsai, C.W.; Hsiao, Y.R.; Lin, M.L.; Hsu, Y. Development of a noise-assisted multivariate empirical mode decomposition framework for characterizing PM 2.5 air pollution in Taiwan and its relation to hydro-meteorological factors. *Environ. Int.* **2020**, *139*, 105669. [CrossRef]
41. Stallone, A.; Cicone, A.; Materassi, M. New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms. *Sci. Rep.* **2020**, *10*, 15161. [CrossRef]
42. Yu, L. Bearing fault diagnosis using time-frequency synchrosqueezing transform. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 4260–4264.
43. Wang, Z.; Xu, Z.; Cai, C.; Wang, X.; Xu, J.; Shi, K.; Zhong, X.; Liao, Z.; Li, Q. Rolling bearing fault diagnosis method using time-frequency information integration and multi-scale TransFusion network. *Knowl.-Based Syst.* **2024**, *284*, 111344. [CrossRef]
44. Verstraete, D.; Ferrada, A.; Droguett, E.L.; Meruane, V.; Modarres, M. Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings. *Shock Vib.* **2017**, *2017*, 5067651. [CrossRef]

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