Multimodal Transformer for Bearing Fault Diagnosis: A New Method Based on Frequency-Time Feature Decomposition

Kai Li (✉ 15842680735@163.com)  
Beijing University of Chemical Technology  
https://orcid.org/0009-0002-6655-5693

Chen Wang  
Beijing University of Chemical Technology

Haiyan Wu  
Beijing University of Chemical Technology

Research Article

Keywords: bearing fault diagnosis, feature extraction, multi-modal Transformer

Posted Date: September 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3258385/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Multimodal Transformer for Bearing Fault Diagnosis: A New Method Based on Frequency-Time Feature Decomposition

Kai Li†, Chen Wang†, Haiyan Wu*

Contributing authors: 15842680735@163.com; WC2020040064@163.com; wuhy@mail.buct.edu.cn;
†These authors contributed equally to this work.

Abstract
Bearing fault diagnosis technology enables early detection of faults and prevention of potential equipment failures. To fully utilize signal data, this paper introduces an innovative diagnostic model named Frequency-Time Multimodal Transformer (FTM-Transformer). It constructs two-dimensional frequency feature data and one-dimensional time-domain feature data, and incorporates a multimodal information fusion module comprehensively exploit the distinctive features of different modes of signals, thereby enhancing fault diagnosis accuracy. The FTM-Transformer model utilizes Multivariate Decomposition for Time and Frequency Features (MDTF) as the feature extraction module to analyze vibration signals from multiple sensors. MDTF employs Multivariate Variational Mode Decomposition (MVMD) to decompose the vibration signals into Intrinsic Mode Functions (IMFs) and applies the Discrete Wavelet Transform (DWT) to extract feature maps of these IMFs, while also analyzing statistical time-domain features of the IMFs to form feature vectors. Building upon this foundation, the FTM-Transformer module is designed to effectively integrates the frequency-domain feature maps and time-domain feature vectors using a deep Transformer model. Experimental results on the bearing fault dataset from Case Western Reserve University demonstrate that the FTM-Transformer model achieves remarkable diagnostic accuracy, significantly outperforming traditional models such as Vision Transformer (ViT) and ResNet. Furthermore, transfer testing on the gear dataset from Southeast University validates the adaptability of the diagnostic model to other application scenarios, showcasing its superior performance. The proposed MDTF and FTM-Transformer methods offer a novel and effective solution to bearing fault diagnosis, contributing to improved equipment reliability and maintenance efficiency in various industrial applications.
1 Introduction

Bearing fault diagnosis is a vital aspect of machinery condition monitoring, playing a crucial role in enhancing equipment reliability, safety, minimizing downtime, reducing maintenance costs, and preventing accidents [1–3]. Bearing fault detection methods can be broadly categorized into two main types: traditional signal processing-based methods and artificial intelligence-based methods.

Traditional signal processing-based methods utilize physical quantities such as bearing vibration signals or acoustic emission signals. Techniques like time-domain analysis [4, 5], frequency-domain analysis [6], and time-frequency analysis [7] are employed to extract characteristic parameters of bearing faults and determine the operational status. Researchers have employed various methods, including neural networks [8], support vector machines [9–11], fuzzy logic [12], expert systems [13], and principal component analysis [14], to establish diagnostic models for bearing faults, enabling automated identification and classification. For instance, Roy et al. utilized the autocorrelation function to process bearing fault signals and extracted nonlinear features from respective vibration correlograms, achieving fault recognition using a random forest classifier [15]. Konar et al. employed continuous wavelet transform (CWT) for processing bearing signals and applied Support Vector Machine for fault diagnosis [16]. Wang et al. introduced EMD manifold (EMDM) to improve fault classification accuracy in noisy environments [17]. Tian et al. extracted features of different fault categories using spectral kurtosis and cross-correlation, applying principal component analysis and semi-supervised k-nearest neighbor distance measure for classification [18]. Hamadache et al. improved fault signal classification’s signal-to-noise ratio using the modified PCA algorithm combined with a low-pass filter (LPF) [19]. Furthermore, Samanta et al. optimized the hidden layer of ANN and the radial basis function kernel parameter of SVM using a genetic algorithm, significantly enhancing the identification of bearing fault categories [20]. Lastly, Yu et al. utilized the Harmonic Significance Index (HSI) and Particle Swarm Optimization (PSO) to determine optimal parameters for wavelet transformation, particularly effective in handling strong non-Gaussian noise interference [21].

In recent years, with the development of computer technology, sensor technology, signal processing techniques, and artificial intelligence, deep learning methods have been widely adopted, continuously improving the accuracy and efficiency of bearing fault diagnosis. For example, Li et al. proposed the multipoint envelope L-kurtosis (MELkurt) method to extract time-domain features of bearing signals and transformed them into images using the Gramian Angular Difference Field (GADF) method. By employing the Conditional Super Token Transformer (CSTT), they achieved higher diagnostic accuracy and reliability [22]. Choudhary et al. extracted bearing signal features using a noninvasive thermal image-based method and employed CNN based on the LetNet-5 structure for bearing fault classification [23]. Ullah et al. achieved
96.65\% accuracy in the classification of non-uniform experimental data using a transfer learning-based pre-trained visual geometry group (VGG) [24]. He et al. effectively extracted features from high-dimensional space and achieved satisfactory accuracy on limited datasets by combining the Siamese network and Vision Transformer [25]. Moreover, Wang et al. innovatively proposed the Symmetrized Dot Pattern (SDP) method to represent bearing status and employed the Squeeze-and-Excitation enabled Convolutional Neural Network (SE-CNN) model for image feature extraction, achieving a high accuracy rate of 99\% in fault classification [26]. Additionally, Alexakos et al. utilized short-time Fourier transform (STFT) to extract time-frequency features of bearing signals and combined it with the Transformer model to improve the classifier’s ability to capture global information, achieving an accuracy rate of 98.3\% with fewer computational resources [27]. Furthermore, Zarei et al. employed two neural network models: the first one acting as a filter to remove non-bearing fault components (RNFC), and the second one used for bearing fault classification, requiring less computation and achieving better results compared to traditional methods [28]. Ma et al. proposed a new method for extracting early fault features in rolling bearings based on the modified Variational Mode Decomposition (MVMD) and Teager Energy Operator (TEO), effectively extracting valid Intrinsic Mode Functions (IMF) from correlation coefficients for fault identification and localization [29]. Moreover, Pang et al. proposed the M-SDP method, combining Multivariate Variational Mode Decomposition (MVMD) with Symmetrized Dot Pattern (SDP) to fuse the bearing drive end and fan end signals through MVMD and visualize the features using the SDP algorithm, employing the Transformer model for fault recognition [30]. Additionally, Georgoulas et al. used the Empirical Mode Decomposition (EMD) and the Hilbert Huang transform to extract signal features and utilized a hybrid ensemble detector to detect deviations of bearings from normal conditions [31].

The improvements in computational technology and artificial intelligence have opened new possibilities in bearing fault diagnosis. Nevertheless, even though deep learning techniques have demonstrated remarkable effectiveness in bearing fault diagnosis, the process involves converting one-dimensional vibration signals into two-dimensional feature maps, which are subsequently integrated with deep learning algorithms. These feature maps usually originate from the signal’s time-frequency combined characteristics. However, when working with simple and direct scalar time-domain features, integrating them into deep learning algorithms can pose certain challenges.

This gap in utilizing the full potential of sensor signals motivates our research to propose a novel multi-modal Transformer approach for advanced bearing fault diagnosis. Our main research goal is to develop an effective bearing fault diagnosis model that can achieve multi-domain generalization, integrate information from various sources, and deliver high accuracy and efficiency in the context of utilizing multi-channel sensor signals for intelligent diagnostics in factories.

Our proposed diagnostic method represents a significant advancement in the field of bearing fault diagnosis, with key features in the following aspects:
1. Multimodal Feature Utilization: The FTM-Transformer fully leverages both time-domain and frequency-domain features of rolling bearing signals. By integrating multivariate mode decomposition and statistical time feature analysis, it constructs one-dimensional time-domain features and two-dimensional frequency-domain maps. This holistic approach captures a more comprehensive representation of bearing health conditions, leading to improved diagnostic accuracy.

2. Information Fusion: Taking advantage of the advancements in Transformer models, the FTM-Transformer efficiently integrates information from different sources, including time-domain and frequency-domain features. This fusion of multi-modal information enhances the model’s ability to make more informed and accurate fault predictions.

3. Operating Condition Generalization: The FTM-Transformer is designed to perform effectively under different operating conditions. It can handle diverse data from various operating scenarios, making it more versatile and applicable in different industrial settings.

This paper proposes a novel approach that fully utilizes both time-domain and frequency-domain features of rolling bearing signals. The method combines multivariate mode decomposition, and statistical time feature analysis to construct one-dimensional time-domain features and two-dimensional frequency-domain maps. To leverage these multimodal features for accurate and efficient fault diagnosis while benefiting from the advancements in Transformer models, we introduce the Fusion ResT-based Multi-modal Transformer (FTM-Transformer) model. The proposed method is tested on the CWRU bearing dataset and the Southeast University gear dataset. To demonstrate the superiority of the FTM-Transformer in bearing fault diagnosis, we conduct a comparative analysis with classical models such as ViT and ResNet.

The main structure of this paper is as follows:

- In Section 2, we present related work, including fundamental methods such as MVMD, DWT, and Transformer.
- In Section 3, we propose the MDTF model, capable of extracting multi-modal features from rolling bearing signals. Besides, we propose the FTM-Transformer tailored to achieve diagnostic tasks. Then, we organize the specific implementation process of the fault diagnosis system proposed in this paper.
- In Section 4, we conduct accuracy experiments, comparative experiments, and ablation experiments on the CWRU bearing dataset, as well as transfer experiments on the Southeast University gear dataset. We in-depth discuss the experimental results.
- Finally, in Section 5, we summarize the experimental conclusions, highlighting the promising advancements in multimodal fault diagnosis of rolling bearings achieved by our proposed MDTF and FTM-Transformer method.

By addressing the limitations of current approaches and capitalizing on the potential of multi-modal features, our research contributes to the development of intelligent fault diagnosis systems, holding great promise for enhancing equipment reliability, safety, and maintenance efficiency in various industries.
2 Related Work

2.1 Multivariate Variational Mode Decomposition

The Multivariate Variational Mode Decomposition (MVMD) algorithm is an extension of the Variational Mode Decomposition (VMD) algorithm for multivariate or multi-channel datasets. The MVMD algorithm can decompose multivariate sequences into a set of Intrinsic Mode Functions (IMFs) with strong noise-robustness. It aims to extract an ensemble of band-limited modes that contain inherent multivariate modulated oscillations present in the multivariate input signal.

In real-life applications, the effectiveness and advantages of MVMD have been demonstrated through numerical examples and the analysis of signals such as EEG (Electroencephalogram) and vibration signals related to bearing fault diagnosis [32].

The main goal of MVMD is to extract K predefined intrinsic mode functions $u_k(t)$ from an input sequence $x(t) = [x_1(t), x_2(t), ..., x_C(t)]$ containing C data channels.

$$x(t) = \sum_{k=1}^{K} u_k(t)$$

where $u_k(t) = [u_{k1}(t), u_{k2}(t), ..., u_{kn}(t)]$.

To achieve the goal of extracting K pre-defined intrinsic mode functions $u_k(t)_{k=1}^{K}$ from the input sequence $x(t) = [x_1(t), x_2(t), ..., x_C(t)]$ containing C data channels, we seek to minimize the sum of the bandwidths of the extracted $u_k(t)_{k=1}^{K}$ while ensuring that the sum of the extracted IMFs can accurately recover the original signal $u_k(t)$. We represent $u_k(t)$ using its vector analytic representation $u_k^+(t)$, which yields:

$$u_+(t) = u(t) + jH u(t) = \begin{bmatrix} u_1^+(t) \\ u_2^+(t) \\ \vdots \\ u_n^+(t) \end{bmatrix} = \begin{bmatrix} a_1(t)e^{j\varphi_1(t)} \\ a_2(t)e^{j\varphi_2(t)} \\ \vdots \\ a_n(t)e^{j\varphi_n(t)} \end{bmatrix}$$

The bandwidth of $u_k(t)$ can be estimated by taking the L2 norm of the gradient function of the harmonic displacement $u_k^+(t)$. To extend the cost function of VMD, the cost function $f$ of MVMD is given by:

$$f = \sum_k \sum_c \left\| \partial_t \left[ e^{-j\omega_k t} u_k^+(t) \right] \right\|_2^2$$

The formula is the optimization objective of MVMD. Below, we will perform constrained optimization on Formula 4.

$$\begin{align*}
&\text{minimize} \quad \left\{ \sum_k \sum_c \left\| \partial_t \left[ e^{-j\omega_k t} u_k^+(t) \right] \right\|_2^2 \right\} \\
&\text{subject to} \quad \sum_k u_{kc}(t) = x_c(t), c = 1, 2, \ldots, C
\end{align*}$$
The intrinsic mode functions need to be distributed around an estimated center frequency \( \omega_k \), which can be obtained by solving the following optimization problem:

\[
L(\{u_{k,c}\}, \{\omega_k\}, \lambda_c) = \alpha \sum_k \sum_c \| \partial_t [e^{-j\omega_k t} u_k(t)] \|^2 + \\
\sum_c \left\| x_c(t) - \sum_k u_{k,c}(t) \right\|^2 + \sum_c \langle \lambda_c, x_c - \sum_k u_{k,c}(t) \rangle
\]

(5)

where \( u_{k,c}(t) \) represents the analytic signal representation of the kth mode for the cth channel, \( \alpha \) is the quadratic penalty factor, and \( \lambda_c(t) \) is the Lagrangian multiplier. The problem is solved using the alternating direction method of multipliers (ADMM), and the mode update relationships are obtained from Formula 6 and 7.

\[
\hat{u}_{k,c}^{n+1}(\omega) = \hat{x}_c(\omega) - \sum_{i \neq k} \hat{u}_{i,c}(\omega) + \frac{\lambda_c(\omega)}{1 + 2\alpha(\omega - \omega_k)^2}
\]

(6)

\[
\omega_k^{n+1} = \frac{\sum_c \int_0^\infty \omega |\hat{u}_{k,c}(\omega)|^2 d\omega}{\sum_c \int_0^\infty |\hat{u}_{k,c}(\omega)|^2 d\omega}
\]

(7)

Formula 6 and 7 can be used to calculate the intrinsic mode functions and their corresponding estimated center frequencies \( \omega_k \).

It can be seen that MVMD can decompose multiple sensor signals into multiple IMFs, making it an effective signal-processing technique. Its advantage lies in the ability to jointly use vibration signals from multiple channels of rolling bearings, providing a basis for time-frequency feature extraction.

### 2.2 Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a widely used transformation method in the field of signal processing [33, 34]. It has the ability to decompose signals into different frequency components, thereby providing effective and high-quality signal analysis and processing. Compared to the short-time Fourier transform, the wavelet transform possesses better time-frequency localization characteristics, enabling more precise capture of transient features in signals.

For bearing signals, fault characteristics are typically concentrated in the low-frequency region. Therefore, utilizing the wavelet transform to extract these sparse feature points is highly advantageous. Through the wavelet transform, we can sensitively detect subtle variations caused by bearing faults, thereby achieving more accurate fault detection and diagnosis.

DWT decomposes the signal into a series of low-frequency signals \( c_{j,k} \) and high-frequency signals \( d_{j,k} \) with different scales. Specifically, considering a one-dimensional signal \( x(t) \), where \( t \) represents time, for a given scale \( j \), we can decompose the signal \( x(t) \) using Formula 8.
\[ x(t) = \sum_k c_{j,k}\phi_{j,k}(t) + \sum_k d_{j,k}\psi_{j,k}(t) \]  

(8)

where \(\phi_{j,k}(t)\) and \(\psi_{j,k}(t)\) are the discrete wavelet basis functions at different scales, \(c_{j,k}\) and \(d_{j,k}\) are coefficients. When \(j = 0\), \(\phi_{j,k}(t)\) and \(\psi_{j,k}(t)\) are referred to as the lowest frequency wavelet functions. They represent the low-frequency and high-frequency components of the signal, respectively. The coefficients \(c_{0,k}\) and \(d_{0,k}\) correspond to the lowest frequency approximation coefficient and detail coefficient, respectively. These coefficients capture the essential information about the signal’s low-frequency content and high-frequency details during the discrete wavelet transform.

2.3 Transformer Framework

Transformer is a sequence model based on the attention mechanism, proposed by Vaswani et al. in 2017 [35]. Multiple research studies have demonstrated the superiority of the Transformer model in the field of fault detection. Fang et al. proposed a lightweight Transformer diagnostic model, significantly enhancing the diagnostic model’s robustness [36].

In this study, we use Transformer to address the bearing fault diagnosis task, which exhibits several key advantages over traditional models such as CNN and LSTM architectures, rendering it highly suitable for this specific task. Firstly, the Transformer excels in capturing long-term dependencies and global contextual information within the signals, facilitating the effective identification of complex fault patterns. Secondly, its self-attention mechanism enables the handling of variable-length signals, making it adaptable to inputs of different lengths. Moreover, the Transformer’s inherent parallel computing capabilities enhance its efficiency, making it well-suited for processing large-scale datasets. Lastly, the incorporation of positional encodings and attention visualization in the Transformer model provides enhanced interpretability, aiding in comprehending the model’s decision-making process. Taken together, these advantages showcase the Transformer’s clear applicability and performance superiority in bearing fault diagnosis.

The framework of the baseline Transformer consists of an Encoder and a Decoder, both of which are composed of multiple identical Transformer Blocks. The structure of a basic Transformer Block structure is presented in Figure 1.

The Transformer Block consists of three main components: Multi-Head Attention, Feedforward Network, and Layer Normalization. Among them, Multi-Head Attention is the core of the entire Transformer, used to learn the semantic information of the input sequence.

2.3.1 Multi-Head Attention

Assuming the input sequence is \(X = (x_1, x_2, \ldots, x_n)\), the Multi-Head Attention in Transformer Block first performs three linear transformations on the input sequence to obtain the Query, Key, and Value matrices as shown in Formula 15:

\[ Q = W_Q X, \quad K = W_K X, \quad V = W_V X \]  

(9)
where $W_Q, W_K,$ and $W_V$ are three different trainable parameter matrices. Then, the Query and Key are multiplied element-wise, followed by a softmax operation. Finally, the resulting attention weight matrix is multiplied by the Value to obtain a weighted sum as the Attention output, as shown in Formula 10:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$ (10)

where $d_k$ is the dimensionality of Query and Key, i.e., the number of columns in $Q$ and $K$. The softmax function in the above equation normalizes the dot product result, making it a weight matrix that represents the relevance of each position in the input sequence with respect to the other positions. This allows each position to be weighted by the information from other positions. In practice, multi-head attention is commonly used, which means that $Q$, $K$, and $V$ are transformed by $h$ different linear projections to obtain $h$ independent attention outputs. Then, these outputs are concatenated and transformed by another linear projection to obtain the final Attention output in Formula 11:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$ (11)

where $\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$, and $W_i^Q, W_i^K, W_i^V$ are the trainable parameter matrices for the $i$-th attention head, and $W^O$ is the final linear transformation parameter matrix.
2.3.2 Feedforward Network and Layer Normalization

The Feedforward Network takes the Attention output and processes it with two fully connected layers and a ReLU activation function, as shown in the following equation:

\[
FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2
\]  

(12)

Layer Normalization normalizes the input sequence along the last dimension so that the representation of each position has the same distribution in the same dimension, which is beneficial for model training and optimization. The mathematical expression of Layer Normalization is shown below:

\[
\text{LN}(x_i) = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \odot \gamma + \beta
\]  

where \( \mu \) and \( \sigma \) are the mean and variance of the input sequence on the last dimension, \( \gamma \) and \( \beta \) are trainable parameters, and \( \epsilon \) is a small number used to prevent the denominator from being zero.

By concatenating multiple Transformer Blocks, an Encoder or Decoder can be constructed to learn the input sequence and generate the output sequence.

2.4 Vision Transformer

The Vision Transformer (ViT) is a variant of the Transformer model that has been successfully applied to computer vision tasks. It is a visual representation learning method based on the self-attention mechanism that can directly learn features from image pixels without the need for traditional preprocessing with convolutional neural networks.

ViT regards image pixels as sequential data and feeds them into a Transformer Encoder to obtain the feature representation of the image. In the implementation of ViT, a trainable embedding layer is first used to convert the image pixels into a set of feature vectors. These feature vectors are treated as a sequence and fed into the Transformer Encoder for processing. The basic structure of ViT is shown in Figure 2.

The input image is first divided into \( N \) image patches of size \( p \times p \), and each patch \( x_i \) is fed into an embedding layer \( E \), which maps it to a \( d \)-dimensional feature vector \( z_i = E(x_i) \). These feature vectors form a sequence \( Z = z_1, z_2, \ldots, z_N \), which is then sent to a Transformer Encoder for processing.

The structure of the Transformer Encoder is similar to that of the Transformer Block used previously but with differences in the implementation of positional embedding and self-attention mechanism. In ViT, the positional embedding is learned through a trainable embedding layer, instead of using the positional encoding used in the Transformer Block.

ViT also employs a self-attention mechanism called Patch-based Self-Attention, which treats each element in the sequence as a block rather than a single token. This enables the Transformer Encoder to perform self-attention computation on the relationships between each block and generate a feature representation for the entire
image. In practice, this is achieved by partitioning the sequence into fixed-size blocks of size $p^2$.

3 Our Proposed Method

In practical applications, bearings usually operate under complex and varying conditions, which may lead to different types of faults in different scenarios. This places significant demands on the sensitivity of bearing fault detection models. Extracting sufficiently effective features from limited data is crucial for accurately classifying various types of faults. We propose a Map-Vector-based fault diagnosis model for bearing fault detection. The model combines frequency domain feature maps and time domain feature vectors, integrating information from both domains. This integration enhances its sensitivity in detecting faults under varying operating conditions. By combining these two types of information, we are able to comprehensively capture bearing fault characteristics, thereby improving the accuracy and reliability of fault detection.

3.1 MDTF Method

We propose a new method named Multivariate Decomposition for Time and Frequency Features (MDTF), which can simultaneously obtain frequency domain feature maps and time domain feature vectors for multivariate sequences.

The workflow of MDTF is illustrated in Figure 3.
In the MDTF method shown in Figure 3, the input is the vibration signals from \( n \) sensors, and the output is two-dimensional frequency domain feature maps and one-dimensional time domain feature vectors.

We utilized the MVMD algorithm, initialized with a normal distribution of center frequencies, to decompose the bearing signal at a sampling frequency of 48 kHz. The convergence error was less than 1e-10, resulting in 6 IMFs of each signal, an example outcome of the signal decomposition is shown in Figure 4, with the corresponding center-frequencies given in Table 1.

As shown in Table 1, after reaching six decomposition layers, the modal components exhibited similar center frequencies. The center frequency of IMF7 was 222.91, differing
by less than 10% from the center frequency of IMF5. Therefore, based on the center
frequency method, the appropriate number of decomposition layers should be six. This
choice effectively characterizes the signal’s features while avoiding the introduction of
high-frequency noise, thereby reducing the complexity and redundancy of the signal’s
characteristics.

3.1.1 Frequency Domain Feature Extraction

After we get intrinsic mode functions (IMFs), we then use a sliding window to divide
the IMFs into $x$ sets of IMF subsequences with each set consisting of $R$ IMFs. We
select a window length of $l_1$ and a sliding step of $l_2$ and apply $J$-level discrete wavelet
transform (DWT) to each set of IMF subsequences, resulting in $R \times J$ high-frequency
subbands of wavelet coefficient sequences denoted as $c_{IMF_{r,j}}$, where $r$ represents the
$r$-th set of IMF subsequences and $j$ represents the $j$-th level of DWT. To obtain the
$2D$ image features of each high-frequency wavelet coefficient sequence, we use bilinear
interpolation to transform the $1D$ coefficient sequence into an $m \times m$ image.

We can obtain $R \times J$ feature maps from each set of IMF sub-sequences from
Formula 11:

$$\mathcal{F}_{IMF} = \left[ g(c_{IMF_{1,1}}, \ldots, g(c_{IMF_{R,J}})) \right]$$

where $g(\cdot)$ denotes the interpolation mapping, and $\mathcal{F}_{IMF}$ is the combination of feature
maps of the wavelet coefficients obtained from the $j$-th level of the $i$-th set of IMF
sub-sequence, with each feature map having a dimension of $m \times m$. During usage, we
treat the combined feature maps as a whole and perform subsequent processing based
on this.

In this paper, taking into consideration the local nature and sparsity of bearing
signals, it is essential to utilize wavelet transformation with compact support and
a high compression ratio to extract signal details and features. Therefore, we adopt
the db4 wavelet as the basis function for wavelet transformation to extract frequency
features, resulting in the frequency domain feature map shown in Figure 5.

In Figure 5, the bright-colored region represents the frequency domain features
of the bearing signal. We can intuitively observe the sparsity characteristics of the
frequency domain features in bearing signals. The sparse nature of these features
imposes stringent requirements on the sensitivity of the model for fault detection.

3.1.2 Time Domain Feature Extraction

To ensure the model retains robust diagnostic capabilities even in the presence of
sparsely distributed feature points, we incorporate time-domain features from the
bearing signals. The continuity exhibited by these time-domain features helps reduce
the risk of the model overlooking essential characteristics during the detection process,
thereby effectively improving the overall diagnostic performance of the model.
We extracted statistical signal features (SSF) including mean, maximum value, root-mean-square (RMS), square-root-mean (SRM), standard deviation, variance, crest factor, kurtosis factor, impulse factor, skewness, kurtosis, normalized fifth moment, and normalized sixth moment [37], thus obtaining time-domain feature vectors for $R$ IMF, and concatenate these vectors into a feature vector of length $15x$, namely:

$$V_{\text{IMF}_i} = [V_{\text{IMF}_1}, V_{\text{IMF}_2}, \ldots, V_{\text{IMF}_{15}}]$$ (15)

where the Crest Factor is employed to characterize the peak characteristics of a signal, while the Kurtosis factor is utilized to measure the sharpness of the data distribution, describing the nature of signal peaks. The Impulse factor is applied to depict the characteristics of pulse signals, whereas Skewness quantifies the asymmetry of the data distribution, serving to describe the symmetry of the signal. The Kurtosis is utilized to gauge the sharpness of the data distribution, revealing the characteristics of signal peaks. Additionally, the Normalized fifth moment is employed to quantify the skewness of the signal, and the Normalized sixth moment is used to measure the nature of signal peaks.

According to the motion characteristics of bearing faults, the vibration of the inner ring relative to the rolling elements is usually characterized by high-frequency vibration, while the vibration of the rolling elements is more concentrated in the low-frequency range. In Table 3, we calculated the time-domain features of the inner race and ball fault signals under a 7 mils fault diameter. We used multiple indicators, including Crest Factor, Kurtosis, Impulse Factor, etc., to evaluate the effectiveness of these features for fault diagnosis.

By comparing the calculation results of time-domain features, we found that the vibration characteristics of the inner ring conform to the expected high-frequency vibration characteristics in multiple indicators such as Crest Factor, Kurtosis, Impulse Factor. These results indicate that the selected time-domain features are effective in fault diagnosis and can help us accurately determine the fault condition of the inner ring.
Table 2 Statistical indicators of time-domain characteristics

<table>
<thead>
<tr>
<th>Elements</th>
<th>Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Mean = ( \frac{1}{N} \sum_{i=1}^{N} x_i )</td>
</tr>
<tr>
<td>Maximum Value</td>
<td>Max = max (</td>
</tr>
<tr>
<td>Root Mean Square</td>
<td>RMS = ( \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} )</td>
</tr>
<tr>
<td>Square Root Mean</td>
<td>SRM = ( \frac{1}{N} \sqrt{\sum_{i=1}^{N}</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>SD = ( \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \text{Mean})^2} )</td>
</tr>
<tr>
<td>Variance</td>
<td>Variance = ( \frac{1}{N} \sum_{i=1}^{N} (x_i - \text{Mean})^2 )</td>
</tr>
<tr>
<td>Crest Factor</td>
<td>CF = ( \frac{\text{Maximum Value}}{\text{RMS}} )</td>
</tr>
<tr>
<td>Kurtosis Factor</td>
<td>KF = ( \frac{1}{3} \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^4}{\text{SD}^4} )</td>
</tr>
<tr>
<td>Impulse Factor</td>
<td>IF = ( \frac{1}{3} \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^3}{\text{SD}^3} )</td>
</tr>
<tr>
<td>Skewness</td>
<td>Skewness = ( \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^3}{\text{SD}^3} )</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Kurtosis = ( \frac{1}{3} \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^4}{\text{SD}^4} ) - 3</td>
</tr>
<tr>
<td>Normalized 5th Moment</td>
<td>N5 = ( \frac{1}{3} \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^5}{\text{SD}^5} )</td>
</tr>
<tr>
<td>Normalized 6th Moment</td>
<td>N6 = ( \frac{1}{3} \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^6}{\text{SD}^6} )</td>
</tr>
</tbody>
</table>

We applied the aforementioned process to each sampled sequence of the fault signal and concatenated the obtained frequency-domain features with time-domain features to construct a multi-modal data package of frequency-domain feature maps and time-domain feature vectors. Each data package consists of a feature map and a one-dimensional feature vector, namely:

\[ \text{Data}_i = \{ \text{Map}_i, \text{Vector}_i \} \]  \hspace{1cm} (16)

where \( \text{Data}_i \) serve as the data foundation for subsequent fault diagnosis tasks in this paper.

3.2 Frequency-Time Multimodal Transformer Method

Based on the implementation of multi-modal feature extraction in Section 2, this paper designs a network structure based on the transformer model to achieve deep fusion of multi-modal features and increase the interpretability of vibration signal feature extraction. Furthermore, a residual fusion module based on an improvement mechanism (Fusion ResT) is designed to make the network converge faster and achieve higher diagnostic accuracy.
Table 3 Examples of Time-domain Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>mean</th>
<th>max</th>
<th>rms</th>
<th>srm</th>
<th>std</th>
<th>var</th>
<th>CF</th>
<th>KF</th>
<th>IF</th>
<th>SK</th>
<th>KU</th>
<th>n5</th>
<th>n6</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMF1</td>
<td>0.07</td>
<td>0.33</td>
<td>0.11</td>
<td>0.11</td>
<td>0.08</td>
<td>0.01</td>
<td>2.91</td>
<td>3.00</td>
<td>2.91</td>
<td>0.13</td>
<td>3.00</td>
<td>0.88</td>
<td>13.42</td>
</tr>
<tr>
<td>IMF2</td>
<td>0.00</td>
<td>1.06</td>
<td>0.27</td>
<td>0.27</td>
<td>0.07</td>
<td>0.03</td>
<td>3.93</td>
<td>3.68</td>
<td>3.93</td>
<td>-0.02</td>
<td>3.68</td>
<td>0.01</td>
<td>22.73</td>
</tr>
<tr>
<td>IMF3</td>
<td>0.00</td>
<td>1.40</td>
<td>0.29</td>
<td>0.29</td>
<td>0.08</td>
<td>0.04</td>
<td>4.87</td>
<td>6.11</td>
<td>4.92</td>
<td>0.00</td>
<td>6.11</td>
<td>-0.15</td>
<td>69.61</td>
</tr>
<tr>
<td>IMF4</td>
<td>0.00</td>
<td>1.25</td>
<td>0.22</td>
<td>0.22</td>
<td>0.05</td>
<td>0.05</td>
<td>5.76</td>
<td>6.59</td>
<td>5.79</td>
<td>0.00</td>
<td>6.59</td>
<td>-0.14</td>
<td>104.53</td>
</tr>
<tr>
<td>IMF5</td>
<td>0.00</td>
<td>0.32</td>
<td>0.09</td>
<td>0.09</td>
<td>0.04</td>
<td>0.03</td>
<td>3.34</td>
<td>3.11</td>
<td>3.63</td>
<td>0.00</td>
<td>3.11</td>
<td>0.15</td>
<td>14.94</td>
</tr>
<tr>
<td>IMF6</td>
<td>0.00</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>3.02</td>
<td>6.92</td>
<td>5.58</td>
<td>0.01</td>
<td>6.92</td>
<td>-0.06</td>
<td>86.68</td>
</tr>
</tbody>
</table>

3.2.1 Design of Classifier

The classifier of ViT typically consists of a linear layer and a Softmax function. The linear layer maps the output vector of ViT to scores for each possible class, resulting in a $K$-dimensional vector for a classification problem with $K$ possible classes. After passing through the linear layer, a Softmax function is used to transform the vector into a probability distribution, providing the probability for each class.

The ViT classifier can be trained using the cross-entropy loss function. During the training process, the loss function value is calculated by comparing the model’s output with the actual label, and then the model’s parameters are updated using the backpropagation algorithm to minimize the loss function value. Specifically, the cross-entropy loss function for ViT can be used, with the mathematical formula as follows:

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij}$$  (17)

where $N$ is the number of training samples, $C$ is the number of classes, $y_{ij}$ is the true label for the $j$th class of the $i$th sample, and $\hat{y}_{ij}$ is the predicted probability of the $j$th class for the $i$th sample by the model. For a given sample, the true label is a one-hot vector with only one element equaling 1 and the rest equaling 0, while the predicted probability by the model is a probability distribution obtained by the Softmax function. Specifically, the calculation formula for $\hat{y}_{ij}$ is as follows:

$$\hat{y}_{ij} = \frac{\exp (z_{ij})}{\sum_{k=1}^{C} \exp (z_{ik})}$$  (18)

where $z_{ij}$ is the logit value of the $j$th class for the $i$th sample, indicating the confidence of the model in predicting the $i$th sample belonging to the $j$th class. During the training process, gradients are calculated using the backpropagation algorithm, and the optimization algorithm is used to update the model parameters. Generally,
stochastic gradient descent (SGD) or other optimization algorithms can be used to train the model.

### 3.2.2 Design of Residual Fusion Module Based on Attention Mechanism

In this section, we propose a residual fusion module based on an attention mechanism (Fusion ResT). This, together with other modules, forms FTM-Transformer, as shown in Figure 6, a novel bearing fault diagnosis model that integrates the time-domain and frequency-domain features of bearing faults, achieving a rapid diagnosis of different types of faults.

![Fig. 6 The Framework of FTM-Transformer](image)

In the context of prior research in the field of computer vision [38], it was emphasized that effectively extracting image features often requires the use of deep Transformer models. When performing multimodal fusion, the design of the fusion module is of paramount importance. Based on these ideas, this study addresses the issue of sparsity in feature points within bearing fault signals and selects a 32-layer
Transformer, ViT-B/32, as the image encoder, initializing it with parameters trained on the JFT-300M dataset.

The frequency domain feature map is encoded, resulting in a sequence of embeddings: \( \{v_1, v_2, ..., v_N\} \). On the other hand, the time domain feature vector is encoded using a 6-layer Transformer as the encoder, with parameters initialized using a Gaussian distribution. The time domain feature encoder is responsible for transforming the input time domain feature vector into a sequence of embeddings: \( \{t_1, t_2, ..., t_N\} \).

However, due to the sparsity of frequency domain features and the continuity of time domain features, feature fusion presents certain challenges. To address this problem, the study adopts a 12-layer Transformer module to increase the average attention span, enabling the model to comprehensively capture the interrelationships between the two types of features, thus extending the receptive field range.

With this deep learning framework, our aim is to effectively fuse frequency domain and time domain features to enhance the analysis performance of bearing fault signals. It is important to note that the specific evaluation of performance requires comprehensive consideration and validation based on actual datasets and specific application scenarios.

In addition, we introduce a structure called Fusion ResT, adding the fusion feature to the processed features by a weighted mechanism to obtain the comprehensive feature for inputting into the SoftMax module.

### 3.2.3 Fusion ResT

The study\[39\] has demonstrated that the weighted shortcut structure significantly improves the performance of transformer models. This structure introduces additional learnable parameters to the original shortcut path, enhancing and recombining the original features. This design prevents the gradual loss of information and diversity in features with increasing depth, effectively avoiding the phenomenon of feature collapse.

In the process of fault detection, frequency domain features play a crucial role. However, due to the relatively sparse nature of frequency domain feature points, there is a risk of feature loss in deep models. To address this challenge, the Fusion ResT structure is introduced to prevent the loss of frequency domain features. Subsequent experiments have validated this structure, leading to an average increase of 1% in diagnostic accuracy.

The Fusion ResT model optimizes feature propagation and gradient propagation by fully fusing time-domain features and frequency-domain features, and combining residual modules based on attention mechanisms.

### 3.3 Specific Steps of Our Proposed Method

We propose a novel fault diagnosis system based on MDTF and FTM-Transformer. The operational process of the system is illustrated in Figure 7.

The specific steps of our novel intelligent bearing fault diagnosis algorithm, which combines MDTF and mumethoTransformer models, are as follows:
Algorithm 1 Our Proposed Bearing Fault Diagnosis Method

Step 1: Use Multivariate Variational Mode Decomposition (MVMD) to fuse the fault signals from the drive end and fan end, obtaining the dominant intrinsic mode functions (IMFs).

\[
\text{IMFs} = \text{MVMD} (\text{drive\_end\_signal}, \text{fan\_end\_signal})
\]

Step 2: Compute time-domain features of the IMFs.

\[
\text{time\_domain\_features} = \text{StatisticTimeFeature} (\text{IMFs})
\]

Step 3: Use discrete wavelet transform to extract frequency-domain feature maps of the IMFs.

\[
\text{frequency\_domain\_feature\_maps} = \text{DWT} (\text{IMFs})
\]

Step 4: Combine frequency-domain and time-domain features to form the dataset.

\[
\text{Data} = \{ \text{Map}_{\text{IMF}} \cdot \text{Vector}_{\text{IMF}} \}
\]

Step 5: Divide the constructed time-domain and frequency-domain feature dataset into a training set, validation set, and test set.

\[
\text{training\_set, validation\_set, test\_set} = \text{SplitDataset} (\text{Data})
\]

Step 6: Use the FTM-Transformer to perform bearing fault classification.

\[
\text{trained\_model} = \text{FTM-Transformer} (\text{train\_data} = \text{training\_set}, \text{validation\_data} = \text{validation\_set})
\]

Step 7: Evaluate the diagnosis performance of our model on the test set.

\[
\text{loss, accuracy} = \text{EvaluatePerformance} (\text{diagnostic\_results}, \text{true\_labels})
\]
The proposed fault diagnosis system exhibits novelty and integration, primarily manifested in the following three aspects:

- **Multi-Modal Feature Utilization**: Signals collected from different sensors typically contain rich multi-modal information, including both time-domain and frequency-domain features in vibration signals. The proposed MDTF feature extraction method effectively harnesses these multi-modal features by employing MVMD to decompose the vibration signals into IMFs and then utilizing DWT to extract frequency-domain features. This comprehensive approach enables the construction of a more holistic feature representation.

- **Holistic Representation**: The uniqueness of MDTF feature extraction lies in its ability to decompose vibration signals into IMFs, which enables the combination of both time-domain and frequency-domain features. This comprehensive feature representation captures a more complete characterization of the bearing health status, surpassing the limitations of methods that rely solely on frequency-domain or time-domain features. By leveraging both domains, the MDTF approach enhances the model’s ability to discriminate subtle fault patterns and improves the overall diagnostic accuracy.

- **Information Fusion**: The strength of the FTM-Transformer lies in its efficient integration of the multi-modal features extracted by MDTF. The incorporation of the Fusion ResT structure in the FTM-Transformer facilitates the seamless fusion of time-domain and frequency-domain features. This fusion capability endows the model with more powerful information processing and representation abilities, enabling it to better exploit the correlations between different features. As a result, the FTM-Transformer achieves higher precision in fault diagnosis and demonstrates the synergistic benefits of combining multi-modal features.

By effectively combining these components, the proposed model offers a novel and integrated approach that capitalizes on the complementary nature of MDTF feature extraction and FTM-Transformer. The MDTF method extracts comprehensive multi-modal features, while the FTM-Transformer efficiently fuses and processes this information, resulting in a more accurate and robust bearing fault diagnosis system. This comprehensive and logical explanation emphasizes the model’s novelty and integration, providing a convincing argument for its effectiveness in addressing the challenges of bearing fault diagnosis.

4 Experimental Results and Analysis

4.1 CWRU Dataset

In this study, we use CWRU bearing dataset to evaluate the proposed model. The test rig used for data collection is shown in Figure 8.

The CWRU dataset consists of four common mechanical fault types: inner race defect, outer race defect, rolling element defect, and cage defect, as well as a normal state as a control group. These fault types were simulated by collecting vibration data from acceleration sensors under different speeds and load conditions. The dataset contains data from different test rigs, and each fault type has multiple severity levels.
The fault types are mainly divided into normal (N), inner race fault (IF), ball fault (BF), and outer race fault (OF). Each fault diameter can be divided into 0.007mm, 0.014mm, and 0.021mm.

In this experiment, we analyzed a total of 8 bearing working states including 7IR, 7BA, 7OR6, 14BA, 14OR6, 21IR, 21BA, and 21OR6, with a motor load of 0 hp using the drive-end and fan-end data sampled at a frequency of 48kHz.

4.2 Experiments

4.2.1 Dataset Establishment

We use 60% of the vibration signals to generate the training samples, 20% to generate the test set, and the remaining 20% to generate the validation set. As shown in Figure 9, the training set and validation set samples are obtained using a sliding window with a length of 2048 and a step size of 128. The testing set samples are also obtained using a sliding window with the same length, but the samples did not overlap with each other.
Finally, we calculate the time-domain feature vectors of the samples and extracted the frequency-domain feature maps using DWT. These feature vectors and maps are used to train and test the proposed model.

4.2.2 Model Performance

In this study, we use accuracy as the performance metric to evaluate the MDTF model. Accuracy is calculated using Formula 22.

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}
\]

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

The accuracy metric measures the overall performance of the model by calculating the proportion of correct predictions out of all predictions made by the model. It is a useful metric for evaluating binary classification models, such as the FTM-Transformer model used in this study.

In this study, the FTM-Transformer model was employed to classify 8 different bearing fault types, including 7IR, 7BA, 7OR6, 14BA, 14OR6, 21IR, 21BA, and 21OR6. The training process was monitored using loss and accuracy curves, as shown in Figure 10.

Figure 10 shows the performance of the FTM-Transformer model on a test dataset for classifying eight different bearing fault features. Figure 10(a) shows the convergence of the training loss of the FTM-Transformer model within 50 epochs, which approaches zero. The Figure 10(b) displays the variation trend of the validation accuracy of the FTM-Transformer model on the validation dataset within 50 epochs,

Fig. 9 Generate sample with overlap
achieving outstanding performance with 97% accuracy in the first epoch and reaching 100% accuracy by the 15th epoch.

We observed that the FTM-Transformer model achieved 100% accuracy for all bearing fault categories in the 48k dataset. This indicates that the proposed model can effectively separate different bearing faults in the feature space, with a large boundary area between different faults, thereby avoiding misdiagnosis. To further verify the diagnostic ability of the FTM-Transformer model, we compared it with the ViT, Efficient-Net, Resnet50, AlexNet, and VGG11 models. The comparative results are shown in Figure 11 and Table 4.
Figure 11 and Table 4 show the fast convergence and superior fault diagnosis ability of our model. Occasional misdiagnosis of bearing faults may have an important impact on industrial generation. Our proposed FTM-Transformer is able to amplify the differences between different fault classes and obtain highly accurate classification capability.

### 4.2.3 Ablation experiments

To verify the scientific and rationality of the model combination, we conducted ablation experiments to evaluate the impact of different modules on the diagnosis of bearing faults. We gradually removed the frequency domain features, residual structure, time domain features, and feature fusion module, and retested the model performance on the original data.

![Ablation of some modules in FTM-Transformer](image)

**Fig. 12** Ablation of some modules in FTM-Transformer
The experimental result, as shown in Figure 12 and Table 5, demonstrates that the frequency domain features have a significant impact on the fault diagnosis performance, leading to a significantly lower fault classification accuracy when removed. Additionally, the feature fusion module also has a significant impact on the classifier’s performance. When removed, the model’s recognition accuracy on the test set is approximately 98%, and the diagnostic stability is not as good as the original model’s performance. Apart from the two aforementioned modules, other small structures also have some impact on the final bearing diagnosis performance, causing some interference with the model’s convergence and diagnostic accuracy.

The experiment demonstrates that the residual module Fusion ResT, based on the attention mechanism, improved the model accuracy by approximately 2%. Moreover, the deep fusion of time-domain and frequency-domain features achieved a diagnostic accuracy of 100% by virtue of using the 12-layer Transformer module. The combined use of these two structures not only improved the convergence speed of the model in the early stage of training but also increased the upper limit of accuracy.

### 4.2.4 Transfer Testing

To evaluate the generalization ability of our model in different domains, we conducted transfer testing on the Southeast University gear dataset. The Southeast University gear dataset is collected from the Dynamic Drive Simulator (DDS) of transmission systems. The gear fault states include a gear bottom crack (Chipped), one missing tooth (Miss), a gear root crack (Root), and gear surface wear (Surface). It is a standard dataset for testing gear system faults and includes vibration data from multiple gear systems.

To be more specific, we applied the FTM-Transformer model again to the Southeast University gear dataset. Firstly, we fused the multi-scale sensor data and then extracted frequency-domain and time-domain features separately. These features were then fed into the ViT and Transformer modules, respectively, and fault diagnosis was performed using a classifier. The convergence of the loss function and the classification accuracy in the experiment are shown in Figure 13:
As shown in Figure 13 and Table 6, the model proposed in this paper has fast convergence and high diagnostic accuracy on the Southeast University gear dataset. These results indicate that the model proposed in this paper can achieve satisfactory results in other one-dimensional time-series signal analysis fields. This provides a good foundation for further research and application of the model.

5 Conclusion

In this study, we have made significant advancements in bearing fault diagnosis by introducing the novel FTM-Transformer method. By effectively fusing feature information from both time domain and frequency domain modalities, we designed the Fusion ResT module, enabling the construction of a powerful deep-learning fault
diagnosis model for multimodal features. Our proposed method not only achieves faster convergence on limited datasets but also demonstrates impressive generalizability to different objects. Compared to other advanced methods, our proposed FTM-Transformer exhibits superior performance in bearing fault diagnosis.

In future research, we envision applying the concept of multi-channel information fusion to other fields beyond bearing fault diagnosis. This approach has the potential to significantly improve the diagnostic capabilities of various systems and equipment. We will continue our efforts to enhance and optimize the FTM-Transformer model further. By exploring other potential improvement methods and incorporating domain-specific knowledge, we aim to continuously boost the diagnostic performance of our method.

The success of the FTM-Transformer method in this study has highlighted the value of leveraging multi-modal features for intelligent fault diagnosis. As technology continues to evolve, we expect that the FTM-Transformer will find broader applications in the industry. With continuous research and development, our proposed method is poised to become an essential tool for intelligent condition monitoring and fault diagnosis, contributing to enhanced equipment reliability, safety, and maintenance efficiency in industrial settings.

6 Compliance with Ethical Standards

The followings are essential statements and declarations of the manuscript.

- **Funding:** This study was funded by Beijing University of Chemical Technology 2023 Undergraduate Innovation and Entrepreneurship Training Plan (grant number 20230010262).
- **Conflict of Interest:** Author Kai Li declares that he has no conflict of interest. Author Chen Wang declares that he has no conflict of interest. Author Haiyan Wu declares that he has no conflict of interest.
- **Author Contributions:** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Kai Li, Chen Wang and Haiyan Wu. The first draft of the manuscript was written by Kai Li, Chen Wang and Haiyan Wu and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.
- **Data Availability:** The datasets analysed during the current study are available in the websites below.
  2. https://gitee.com/zhengkun110/Mechanical-datasets

References


