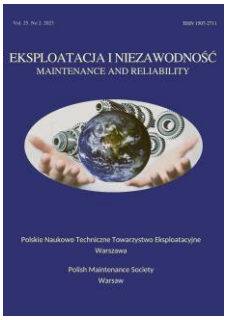

Eksploatacja i Niezawodnosc – Maintenance and Reliability

Volume 28 (2026), Issue 4

journal homepage: http://www.ein.org.pl

Article citation info:
Xue H, Wang X, Zhang L, Xu Y, Ouyang X, Jiang Z, Bao X, Yue H, Chen P, A multi-scale depth-wise separable convolution swin transformer for fault diagnosis of in-wheel motor bearings, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2026: 28(4) http://doi.org/10.17531/ein/220213

A multi-scale depth-wise separable convolution swin transformer for fault diagnosis of in-wheel motor bearings

Indexed by: Web of Science Group

Hongtao Xue[a,*], Xuan Wang[a], Liang Zhang[a], Yanlong Xu[a], Xupeng Ouyang[a], Xiaoyi Bao[a], Zongkang Jiang[a], Huiyu Yue[a], Peng Chen[b]

[a] School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China
[b] Graduate School of Bioresources, Mie University, Tsu 514-8507, Japan

Highlights

- A multi-scale Swin Transformer framework for in-wheel motor fault diagnosis.
- MSCF-EFFN for enhanced shallow multi-scale feature representation.
- Deformable window self-attention for adaptive modeling of transient impacts.
- Depth-wise convolution attention for efficient deep feature refinement.
- High diagnostic accuracy under dynamic disturbances and multiple conditions.

Abstract

Fault diagnosis of in-wheel motor bearings is challenging due to weak fault features and non-stationary vibration signals under complex operating conditions. To address the limitations of conventional models in transient impact modeling and feature representation, this study proposes an enhanced Swin Transformer–based fault diagnosis framework. The proposed method integrates a multi-scale convolutional feature-enhanced feed-forward network (MSCF-EFFN) to improve shallow cross-scale representation, a unified deformable shifted window multi-head self-attention (DSW-MSA) framework to adaptively capture irregular transient impact features, and a depth-wise convolution attention module (DWCAM) to refine deep feature selection. The model is validated on a self-built dynamic test bench covering nine bearing health states and 28 operating conditions, achieving an average accuracy of 98.7% and a peak accuracy of 99.12%. Comparative and ablation studies demonstrate superior accuracy, robustness, and convergence performance over existing models.

Keywords
in-wheel motor, fault diagnosis, swin transformer, attention mechanism, depth-wise separable convolution

1. Introduction

With the increasing societal focus on sustainable transportation and green energy, the development of new energy vehicles has emerged as a national strategic priority. Among these, pure electric drives have garnered widespread attention due to their high efficiency and environmental benefits. As an advanced drive configuration, in-wheel direct-drive electric vehicles integrate motors directly into their wheels, facilitating multi-motor cooperative control and precise power distribution. These systems offer advantages such as high efficiency, rapid response, and superior space utilization, rendering them ideal architectures for electric vehicles [1–3]. However, due to their special installation locations, in-wheel motors are more susceptible to suspension and road impacts under complex operating conditions. Coupled with limited heat dissipation and sealing conditions, this often leads to failures, such as bearing damage and rotor eccentricity, subsequently causing torque fluctuations and power imbalances, which can seriously compromise driving safety [4,5]. Therefore, the development of efficient and robust fault diagnosis methods for in-wheel motors is of significant importance.

(*) Corresponding author.
E-mail addresses: H. Xue (ORCID: 0000-0003-0912-3413) xueht@ujs.edu.cn, X. Wang (ORCID: 0009-0006-3987-265X) wangx@stmail.ujs.edu.cn, L. Zhang, zhangliang@stmail.ujs.edu.cn, Y. Xu, xuyanlong@stmail.ujs.edu.cn, X. Ouyang, oyxp@stmai.ujs.edu.cn, X. Bao, baoxiaoyi@stmail.ujs.edu.cn, Z. Jiang, jzk@stmail.ujs.edu.cn, H. Yue, yuehuiyu@ujs.edu.cn, P. Chen, chen@bio.mie-u.ac.jp

Traditional methods primarily rely on signal processing and manual feature extraction, such as time-domain [6–8], frequency-domain [9–11], and time-frequency domain analyses [12–14]. Recent studies have proposed advanced signal processing–based methods to enhance fault feature extraction accuracy. Representative approaches include methods employing multi-scale entropy or energy-based features together with optimization-enhanced SVM classifiers, such as GWO-SVM, and have exhibited improved fault-identification accuracy and robustness under varying operating conditions [15]. Furthermore, approaches that integrate VMD-based feature extraction with optimization-enhanced machine-learning models, such as the improved GWO-LS-SVM framework, have also demonstrated superior diagnostic accuracy and stability across the full life cycle of high-speed bearings under complex and variable operating conditions [16]. Despite these promising outcomes, these methods still heavily rely on prior knowledge, exhibit limited feature representation capacity, and face challenges when adapting to varying operational environments. Beyond signal-processing-based approaches, classical modeling and classification strategies have also demonstrated successful applications across various domains. For example, associative Petri net (APN) models have been employed for arrhythmic beat classification [17], highlighting the effectiveness of structured logic-based models in complex signal classification tasks. Similarly, structural optimization strategies have been explored in traditional clustering algorithms, such as improved centroid update approaches for k-means [18], which enhance computational efficiency and convergence behavior. Furthermore, neural network-based pattern recognition frameworks developed for edge computing environments [19] emphasize the importance of architecture-specific adaptation and structural refinement to meet resource and performance constraints. Collectively, these studies indicate that explicit structural design and optimization play a critical role in improving modeling effectiveness beyond simple parameter adjustment or model scaling. This perspective is central to the field of evolving fuzzy and neuro-fuzzy systems, where real-time structural adaptation from data streams is a key characteristic [20].

To further improve the time-frequency representation capability of signal features, researchers have proposed various time-frequency analysis methods. The short-time Fourier transform (STFT) [21] analyzes non-stationary signals using a fixed window for local spectrum analysis; however, its time and frequency resolutions are limited by the choice of window length. In contrast, the wavelet transform (WT) [22] achieves multi-scale decomposition by scaling and translating the mother wavelet, allowing for adaptive adjustment of time-frequency resolution based on signal frequency, thus offering significant advantages for non-stationary signal analysis. The continuous wavelet transform (CWT) further extends this concept and has been extensively utilized in research to map one-dimensional vibration signals into two-dimensional time-frequency images for deep network learning. Recent studies have explored the deep integration of wavelet principles into neural network architectures for intelligent fault diagnosis under conditions of small-sample sizes and low signal-to-noise ratios [23]. Other studies have integrated the wavelet concept directly into the convolutional kernel design to construct multi-wavelet kernel CNNs, thereby enhancing the sensitivity of deep models to non-stationary transient features [24]. Additionally, methods combining the CWT with a CNN have been successfully applied to fault identification in complex systems, such as wind turbines, demonstrating strong robustness and generalization capability [25]. These studies collectively indicate that CWT provides more discriminative two-dimensional time-frequency feature inputs for deep neural networks.

In recent years, deep learning has become a prevalent approach for fault diagnosis in rotating machinery. Convolutional neural networks (CNNs) [26,27] are adept at efficiently extracting local features; however, pooling operations may result in a loss of detail. Long short-term memory networks (LSTMs) [28,29] are well-suited for temporal modeling but exhibit low training efficiencies. In contrast, transformer architectures [30,31] utilize multi-head self-attention (MHSA) to facilitate global dependency modeling and parallel computation, demonstrating robust performance on medium- and long-sequence tasks. Nevertheless, as transformers were originally designed for natural language processing, their direct application to vibration spectrograms presents challenges such as structural flattening and high computational cost. To mitigate these limitations, recent studies have explored vision transformer (ViT) architectures for time-

frequency image analysis, which have shown improved long-range spatial modeling and enhanced diagnostic performance under noisy and variable operating conditions [32].To further augment the feature focusing and representation capabilities of deep models, researchers have introduced attention mechanisms [33]. The early squeeze-and-excitation (SE) module [34] achieved adaptive channel weight adjustment using global average pooling. Subsequently, the convolutional block attention module (CBAM) [35] combined channel and spatial attention to collectively enhance the important feature regions. In recent years, self-attention mechanisms have been widely incorporated into fault diagnosis models to improve the modeling of long-range dependencies and to focus on key features. Recent studies have explored the application of transformer architectures in mechanical fault diagnosis. For instance, a lightweight multi-scale convolutional sparse attention transformer (MCSAT) has been developed for rotating machinery, which integrates multi-scale feature extraction and dynamic sparse attention mechanisms to achieve high diagnostic accuracy under limited sample conditions while maintaining low computational cost [36].

The swin transformer introduces shifted windows and a hierarchical architecture, significantly reducing computational complexity while maintaining global modeling capability. It has been widely adopted in visual tasks, demonstrating remarkable performance [37]. Motor vibration signal analysis methods based on locality-sensitive hashing (LSH) and swin transformer have been proposed to refine attention distribution under sparse features, thereby enhancing key feature learning and convergence efficiency [38]. These studies underscore the substantial potential of the swin transformer in modeling complex visual features. However, most existing swin transformer-based fault diagnosis methods are directly adapted from vision architectures without sufficient structural adjustment for non-stationary mechanical signals. First, conventional frameworks rely primarily on hierarchical patch embedding and standard Transformer blocks, lacking a dedicated front-end mechanism to explicitly enhance cross-scale transient features at the input stage. This limits the early representation of weak and multi-scale coupled fault characteristics. Specifically, weak impact features in vibration spectrograms are often sparsely distributed and exhibit relatively low energy compared with dominant background components. During the hierarchical patch embedding process, local regions are aggregated into coarse tokens at early stages, which may smooth subtle transient variations and weaken the representation of fine-grained impact details. Consequently, the discriminative information carried by weak fault signatures may be partially suppressed before entering deeper attention layers. Second, the fixed window-based attention mechanism constrains spatial adaptivity, which may reduce sensitivity to irregular transient impact distributions under varying operating conditions. Third, channel interaction across hierarchical stages remains relatively implicit, which may weaken robustness when load and speed fluctuate.

To alleviate similar structural limitations observed in deep architectures, researchers have explored various structural enhancement strategies. Recent studies have demonstrated that structural enhancement through multi-scale convolutional embedding, lightweight depth-wise separable convolution, and adaptive attention refinement can effectively improve feature representation capability in vision-based and vibration-based diagnostic tasks. Multi-scale convolutional front-end designs have been shown to strengthen cross-scale transient feature extraction before transformer encoding [39], while depth-wise separable convolution offers a computationally efficient means of enhancing local modeling capacity. In addition, deformable or adaptive attention mechanisms have been introduced to improve spatial flexibility and better capture irregular feature distributions [40]. Furthermore, channel attention refinement strategies have been widely adopted to strengthen inter-channel interaction and enhance feature discrimination [41]. These developments provide important structural inspirations for the present work.

These limitations become more pronounced in in-wheel motor fault diagnosis, where fault signatures are typically weak, non-stationary, and strongly coupled across scales. Therefore, a problem-oriented structural enhancement of the swin transformer is necessary rather than simple architectural transplantation or empirical stacking of modules.

To address the aforementioned issues, this study proposes a fault diagnosis method for in-wheel motors based on a multi-scale depth-wise separable convolution swin transformer. The proposed framework aims to overcome the structural limitations

of conventional swin transformer architectures in modeling non-stationary and transient fault signals, rather than merely improving empirical performance. The main contributions are summarized as follows:

1) A multi-scale convolution feature-enhanced feed-forward network (MSCF-EFFN) is designed as a front-end feature extraction module prior to the Swin backbone. By employing parallel depth-wise separable convolution branches, MSCF-EFFN enhances cross-scale transient feature representation and compensates for the limited multi-scale modeling capability of conventional Transformer structures.
2) A deformable shifted window multi-head self-attention (DSW-MSA) mechanism is proposed by incorporating deformable sampling into both regular-window and shifted-window attention forms, thereby improving adaptive perception of irregular transient impact features.
3) A depth-wise convolution attention module (DWCAM) is introduced to strengthen cross-layer channel interaction modeling through dual-weight channel refinement, enhancing robustness under varying operating conditions.
4) A systematic ablation and convergence validation framework is established to quantitatively evaluate the contribution of MSCF-EFFN, DSW-MSA, and DWCAM, and to verify the optimization stability of the overall multi-scale depth-wise separable convolution swin transformer architecture.

Through the coordinated integration of these modules, the proposed model achieves a balance between shallow local feature enhancement and deep global representation learning, thereby improving the stability, adaptability, and diagnostic reliability of the swin transformer in in-wheel motor fault diagnosis tasks. It is worth noting that alternative structural schemes could also be considered for each proposed module. For example, conventional multi-scale convolution blocks, standard deformable attention mechanisms, or other channel attention variants may serve as potential substitutes. However, the present design is selected by jointly considering computational efficiency, structural compatibility with the swin backbone, and the characteristics of non-stationary in-wheel motor fault signals. The MSCF-EFFN adopts lightweight depth-wise separable branches to balance multi-scale feature enhancement and parameter efficiency. The DSW-MSA introduces deformable sampling while preserving the shifted-window mechanism to maintain hierarchical modeling consistency. The DWCAM is designed to strengthen cross-layer channel interaction without significantly increasing model complexity. Therefore, the proposed modules represent a task-oriented structural optimization rather than arbitrary architectural stacking.

2. Multi-scale depth-wise separable convolution swin transform network

The swin transformer, a hierarchical vision transformer architecture, employs the shifted window multi-head self-attention (SW-MSA) mechanism to maintain efficiency while balancing local and global modeling. It has demonstrated promising results in tasks such as computer vision and signal processing, and has been increasingly adopted in state classification tasks due to its ability to extract discriminative features that provide a solid foundation for subsequent classification. However, its direct application to bearing fault diagnosis under complex operating conditions encounters three major bottlenecks: the front-end feature extraction relies solely on linear projection, resulting in insufficient cross-scale representation capability; the window partitioning is rigidly fixed, making adaptation to irregular transient impacts difficult; and as the network depth increases, the feature screening efficiency at each stage is inadequate, leading to a gradual loss of spatial details and underutilization of channel dependencies. To achieve more accurate fault state classification, this study proposes a network architecture based on a multi-scale depth-wise separable convolution swin transformer, as illustrated in Fig. 1, which is specifically designed to address the aforementioned limitations while enhancing the discriminative power of learned features for classification purposes. Building upon the hierarchical design and computational efficiency advantages of the swin transformer, this method innovatively introduces and integrates three key modules: 1) The multi-scale convolution feature-enhanced feed forward network (MSCF-EFFN) is designed to extract cross-scale features at the input through parallel convolutional branches, thereby enriching shallow-level representations; 2) The deformable shifted

window multi-head self-attention (DSW-MSA) framework incorporates a deformable sampling mechanism into the window attention module of the swin transformer block. This enhancement improves both regular-window and shifted-window instantiations within the unified DSW-MSA framework by endowing them with dynamic learnable sampling capability during local window modeling and cross-window interaction, thus augmenting the model's adaptive modeling ability for irregular transient impact features; 3) The depth wise convolution attention module (DWCAM) is embedded at the end of each stage to perform cross-layer screening and channel dependency enhancement on the feature output by the multiple swin transformer blocks. Through the complementary and synergistic effects of these three components, the proposed model achieves precise capture and robust modeling of weak fault features in bearings under complex operating conditions, thereby providing a more efficient and reliable solution for intelligent fault diagnosis of in-wheel motors. From a feature learning perspective, this design promotes the formation of a discriminative feature space with enhanced intra-class compactness and inter-class separability, which is beneficial for classification and for promoting clustering-like discriminative organization of fault patterns in the learned feature space.

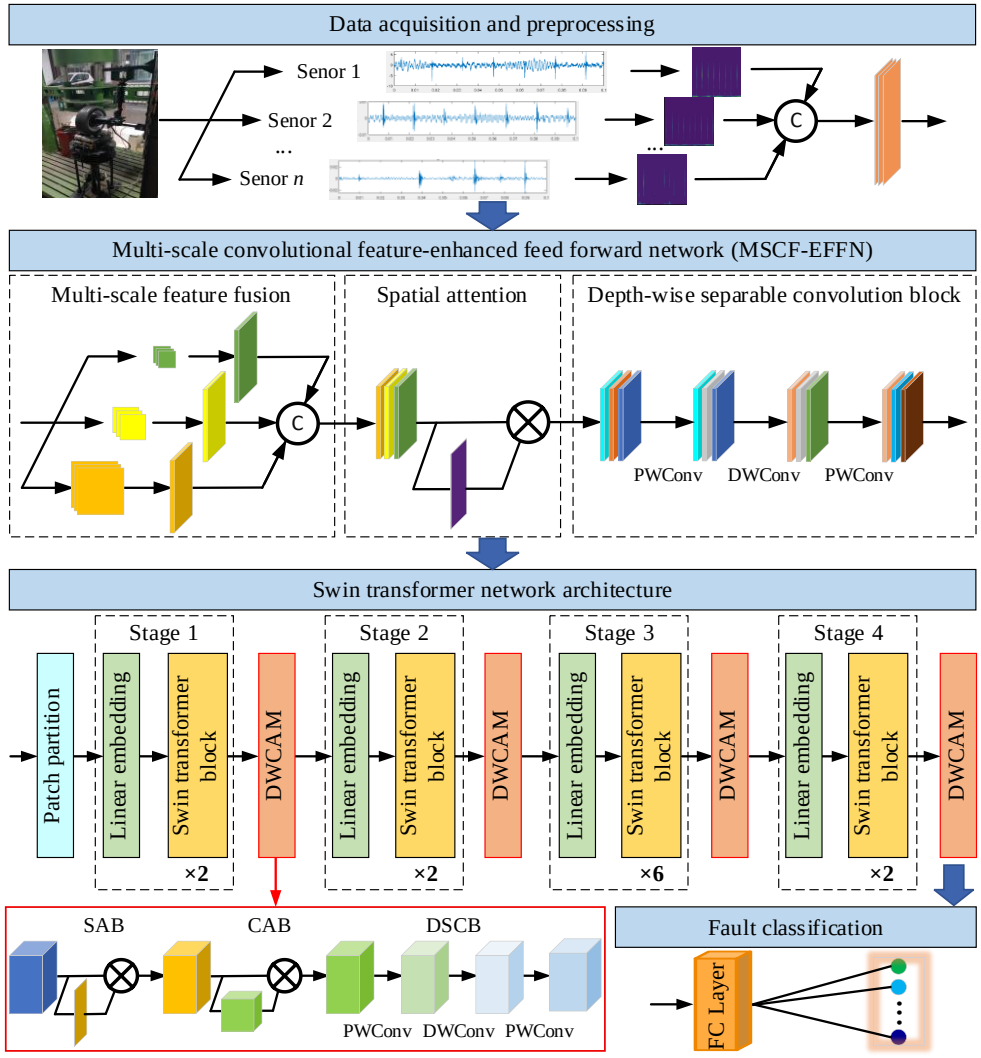

Figure 1. Multi-scale depth-wise separable convolution swin transformer network architecture.

2.1. Multi-scale convolutional feature-enhanced feed forward network

A multi-scale convolutional feature-enhanced feed forward network (MSCF-EFFN), as shown in Fig. 2, represents a feature extraction method designed based on the integration of multi-scale modeling and convolutional feature extraction. It is tailored to handle multiscale components under complex operating conditions, including long-term trends, mid-frequency resonances, and transient impacts. The MSCF-EFFN

effectively overcomes the limitations of the standard swin transformer, which relies solely on a single linear projection at the input stage, by capturing cross-scale features to augment fault discrimination capabilities. The MSCF-EFFN comprises three parallel convolutional branches that utilize large, medium, and small receptive field kernels to extract features at varying scales, as shown in Fig. 2(A). The large kernel is dedicated to capturing long-term trends, the medium kernel focuses on the mid-frequency components, and the small kernel is responsible for capturing transient impacts. The selection of these kernel sizes (7×7, 3×3, and 1×1) is based on the characteristic time-frequency scales of vibration components and was validated through preliminary experiments to achieve an optimal balance between feature richness and computational efficiency. Let the input signal tensor be denoted as X, and the convolutional operation of the i-th branch is defined as:

$$F_{msc}^{(i)} = Conv(X, k_i \times k_i), k_i \in \{k_{large}, k_{medium}, k_{small}\} \quad (1)$$

Where k_{large}, k_{medium}, and k_{small} denote the sizes of convolutional kernel with large, medium, and small receptive fields, respectively. F_i represents the feature map extracted by the i-th convolutional branch (large, medium, or small kernel). In this study, the kernel sizes for large, medium and small receptive fields are specified as 7×7, 3×3 and 1×1, respectively. Each convolutional branch is configured with a stride of 2 and appropriate padding to ensure $F_{msc}^{(i)} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$.

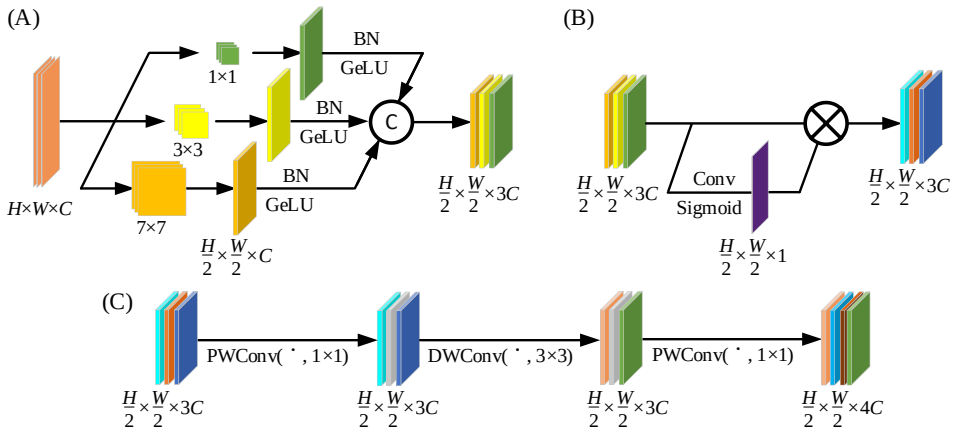

Figure 2. Multi-scale convolutional feature-enhanced feed-forward network: (A) Multi-scale feature fusion, (B) Spatial attention, (C) Depth-wise separable convolution block.

The features derived from various scales produced by each convolutional branch are subjected to batch normalization and GeLU operations respectively. Subsequently, these features are concatenated along the channel dimension to constitute the multi-scale feature set, which can be expressed as:

$$F_{cat} = Concat\left[F_{msc}^{(1)}, F_{msc}^{(2)}, F_{msc}^{(3)}\right] \quad (2)$$

Where $Concat[\cdot]$ represents the concatenation operation along the channel dimension. The concatenated feature $F_{cat} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3C}$ forms a 3C-channel structure that integrates large, medium, and small-scale semantics, thereby offering enriched feature mappings for subsequent attention fusion.

The spatial attention mechanism is employed to assign weights to the multi-scale feature set, as shown in Fig. 2(B), obtaining the re-calibrated feature set G:

$$G = \alpha F_{cat} \quad (3)$$

Where α represents the set of attention weights corresponding to the output $F_{mas}(i)$ of the i-th convolutional branch, α=[α_1, α_2, α_3]. The weight α_i can be determined based on the parameter ω_i as:

$$\alpha_i = \frac{exp(\omega_j)}{\sum_{j=1}^{3} exp(\omega_j)} \quad (4)$$

Where ω_i serves as the weight generation input, reflecting the spatial importance of each scale feature. Following Sigmoid normalization, it is mapped to the attention weight α_i, facilitating the dynamic adjustment of the importance across various scale features. This mechanism enables the model to automatically prioritize critical-scale features while minimizing the influence of redundant and noisy components, thereby achieving a recalibrated feature representation with an enhanced discriminative capability. The effectiveness of this attention-based fusion was verified through ablation studies comparing it with simple concatenation alternatives.

Next, it is the depth-wise separable convolution block, as shown in Fig. 2(C), which structure is sequentially composed of

a pointwise convolution, depthwise convolution, and another pointwise convolution connected in series.

$$F' = PWConv(DWConv(PWConv(G, k_4 \times k_4)), k_5 \times k_5), k_6 \times k_6) \quad (5)$$

Where $PWConv(\cdot)$ denotes pointwise convolution and $DWConv(\cdot)$ represents depthwise convolution. Depthwise convolution operates by assigning independent convolutional kernels to each input channel and performing spatial convolution exclusively within individual channels. This "channel-wise convolution" mechanism avoids parameter sharing between channels, enabling comprehensive capture of local spatial textures and detailed features without increasing the parameter count. However, depthwise convolution is confined to intra-channel modeling and lacks cross-channel information interaction capability. To mitigate this limitation, pointwise convolution linearly combines local features from various channels, achieving a weighted fusion and redistribution of cross-channel information. In the depth-wise separable convolution block of this study, the convolutional kernels are sequentially configured as 1×1, 3×3, and 1×1, with all convolutions employing a stride of 1 and appropriate padding. This specific configuration was determined through comparative experiments to balance representational capacity and computational cost. It is important to note that the channel dimensionality within this block evolves according to the number of filters used in the pointwise convolutions. Specifically, the initial 1×1 convolution preserves the original channel dimension (3*C*), while the intermediate depthwise convolution also maintains this dimensionality due to its one-kernel-per-channel design. The final 1×1 convolution expands the channel dimension from 3*C* to 4*C* by employing 4*C* output filters, thereby enabling richer feature projection and enhanced representational capacity. Batch normalization and GeLU operations are applied after each pointwise and depthwise convolution, ultimately yielding the depthwise separable convolution feature set $F' \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 4C}$.

The process first enters the patch partition and linear embedding stage to achieve the transition from the 2D convolutional feature space to the high-dimensional transformer representation space. During the patch-partition stage, the input features are divided into several nonoverlapping P×P image patches. Each patch is flattened into a vector of length $P^2 \times 4C$, thereby reducing the spatial resolution while aggregating contextual information from local regions. Subsequently, the linear projection layer maps each patch vector to a unified high-dimensional feature space, thereby achieving dimensional mapping from $P^2 \times 4C$ to C'. In this study, the patch size was set to 2×2, and the channel number C' was twice that before the projection. This process not only accomplishes channel dimension expansion and feature domain transformation but also ensures consistent representation dimensions across different patches, providing a unified input foundation for subsequent hierarchical feature modeling and window-based multihead self-attention mechanisms.

2.2. Deformable Shifted window multi-head self-attention

To enhance the model's capability in representing irregular transient impact features, this study introduces the concept of deformable sampling into the window self-attention mechanism and proposes a unified deformable shifted window multi-head self-attention (DSW-MSA) framework, which can be instantiated in both regular-window and shifted-window forms within the swin transformer. Here, the term DSW-MSA framework is used as an umbrella term for the proposed deformable window-attention mechanism, which includes both a regular-window variant and a shifted-window variant. Inspired by deformable convolution, the proposed framework enables flexible receptive field adaptation for spatially non-uniform features. By incorporating dynamic offset sampling after window partitioning, the attention mechanism can adaptively adjust the sampling positions within local windows, thereby improving sensitivity to irregular transient impacts in time-frequency representations. An overview of the improved swin transformer blocks integrating the proposed deformable attention mechanism is shown in Fig. 3.

In the standard swin transformer multi-head self-attention mechanism, each attention head independently performs a weighted computation, and the output of the *i*-th attention head can be expressed as:

$$Attention_i(Q, K, V) = Softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_k}} + B\right) V_i \quad (6)$$

Where Q_i, K_i and V_i represent the query, key and value matrices of the *i*-th attention head, respectively; d_k denotes the dimensionality of the key vectors in each attention head; B is the relative position bias, used to encode relative spatial position

information within the window; and the softmax function normalizes all attention scores.

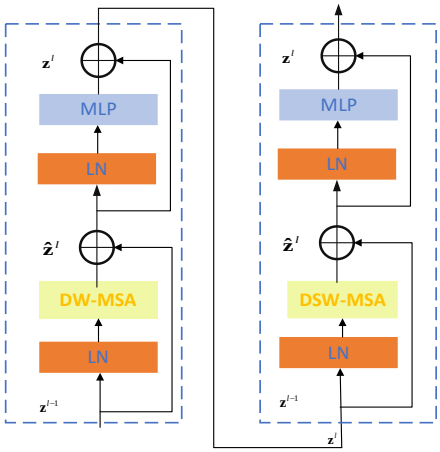

Figure 3. Improved structure of swin transformer blocks.

Unlike the W-MSA, which relies solely on fixed window partitioning, the SW-MSA achieves cross-window interaction by shifting window positions in adjacent layers. However, the window shape and partitioning strategy remain fixed and the attention sampling locations are predefined, making it difficult to adapt to the irregular and discretely distributed transient impact features in the time-frequency spectrograms of vibration signals. The proposed deformable attention module effectively overcomes the limitations in capturing discrete and irregular impact features in time-frequency representations, thereby enhancing the perception of local transient patterns and achieving superior global and local dependency modeling under complex operating conditions. Its core steps are as follows. The deformable offset prediction mechanism was designed based on the observation that transient impact features in vibration spectrograms are often distributed irregularly and sparsely. A lightweight convolutional layer with 3×3 kernels was selected for offset prediction after comparative experiments, as it provides sufficient spatial context for offset learning while maintaining computational efficiency.

1) For each query point Q, a lightweight convolutional layer is used to predict the 2D offset Δp:

$$\Delta p = s \cdot tanh\left(\theta_{offset}(Q)\right) \tag{7}$$

Where the reference point p represents the initial localization center of the query on the feature map, defining the starting position for attention sampling; $\theta_{\text{offset}}(\cdot)$ denotes a lightweight convolutional or linear layer used to learn the offset distribution based on Q; tanh(·) represents the hyperbolic tangent function, which constrains the predicted offset range to [-1, 1]; and s is a scale factor that controls the maximum offset range, empirically set to 2 based on preliminary experiments to balance sampling flexibility and feature alignment, preventing the model from generating excessively large sampling displacements that could cause feature misalignment.

2) Add the offset Δp to the reference point p, and perform bilinear interpolation sampling on the key and value features at the offset position:

$$K'(p) = S(K, p + \Delta p) \tag{8}$$

$$V'(p) = S(V, p + \Delta p) \tag{9}$$

Where $S(\cdot)$ is the bilinear interpolation sampling function used to compute continuous feature values at non-integer coordinate positions; $K'(p)$ and $V'(p)$ represent the key and value features after offset sampling, respectively.

3) Compute the final attention output using the resampled key and value features K' and V', which follows the same formulation as Eq. (6).

Building upon this foundation, the proposed framework is implemented in two variants, namely deformable window multi-head self-attention (DW-MSA) and deformable shifted-window multi-head self-attention (DSW-MSA). As the regular-window variant, DW-MSA extends W-MSA by introducing deformable sampling offsets within local windows, thereby enabling the attention mechanism to adaptively focus on the most responsive time-frequency features in local regions. This enhancement improves the model's ability to capture irregular transient impact signals. In contrast, as the shifted-window variant, DSW-MSA incorporates a shifted-window strategy to establish cross-region feature interactions between adjacent windows, thereby achieving dynamic fusion of local modeling and broader contextual dependencies. These two variants are alternately integrated into the swin transformer blocks, forming a hierarchical architecture characterized by local-global interaction. The alternating use of DW-MSA and DSW-MSA in successive swin transformer blocks was determined through ablation studies, which confirmed that this local-global interaction pattern achieves superior feature representation compared with using either variant alone. This synergistic mechanism not only significantly enhances the model's ability to represent non-stationary vibration signals under complex

operating conditions, but also improves the flexibility of time-frequency feature extraction and diagnostic robustness.

2.3. Depth wise convolution attention module

The attention mechanism significantly enhances the representation capability of deep learning models by focusing on key information. Rahman et al [42] introduced the multi-scale convolutional attention module (MSCAM), which combines channel and spatial attention with multi-scale convolution to enhance feature discriminability while reducing computational costs. However, the efficiency and adaptability of this module in focusing on features within fault diagnosis contexts necessitate further enhancement. Building upon this inspiration, we propose the depth-wise convolution attention module (DWCAM), which introduces two key innovations tailored for fault diagnosis: (1) a spatial-first channel-later attention sequence that aligns with the logical process of localizing abnormal regions before identifying fault types, and (2) a simplified single-branch depthwise separable convolution design that matches the progressive resolution reduction in swin transformer stages.

Fig. 4 illustrates the depth wise convolution attention module (DWCAM), which comprises three core units arranged sequentially: the spatial attention block (SAB), the channel attention block (CAB), and the depth-wise separable convolution block (DSCB). Unlike channel-prioritized attention structures, the DWCAM introduces an innovative sequence of attention mechanisms by adopting a spatial-first channel-later processing flow. This design is more congruent with the logical process of fault diagnosis, wherein abnormal regions are initially localized before determining the fault type. Empirical evidence indicates that this sequence more effectively targets and enhances critical local features. Furthermore, to accommodate the hierarchical structure of the swin transformer, where feature resolution decreases progressively across stages, the DSCB abandons complex multibranch designs and employs only a single-branch depthwise separable convolution. This strategy reduces the parameter count while ensuring that the receptive field matches the feature-map scale, thereby augmenting the capability for spatial detail extraction. It should be noted that the term “depth wise convolution” in DWCAM emphasizes the dominant spatial modeling operation within the attention module. In practice, the module is implemented using a depth-wise separable convolution structure to achieve efficient spatial–channel decoupling and computational efficiency.

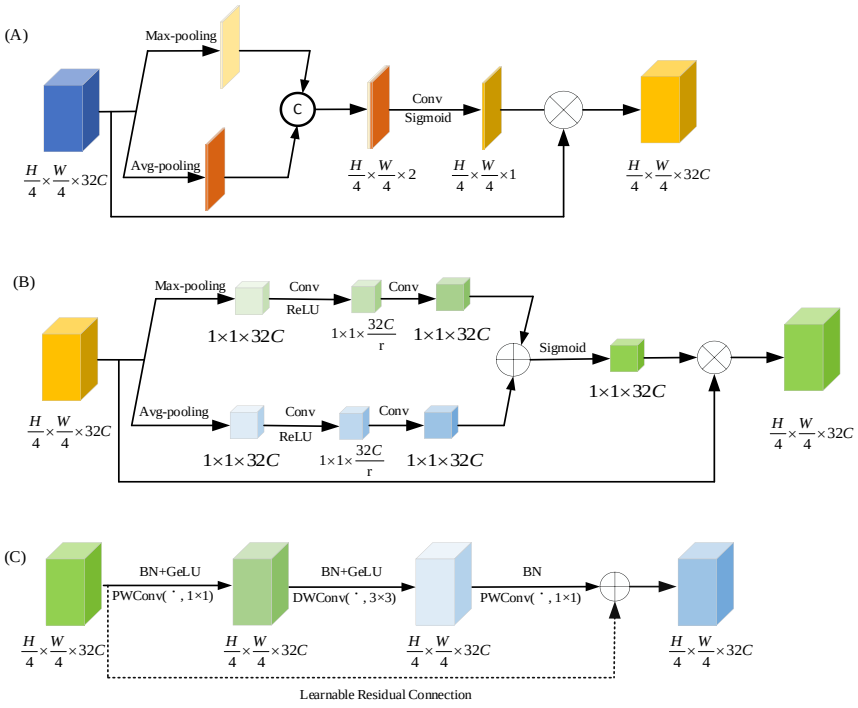

Figure 4. Depth wise convolution attention module: (A) spatial attention block (B) channel attention block (C) depth-wise separable convolution block.

To illustrate the parameter computation process of DWCAM, the DWCAM module embedded at the end of Stage 1 is considered as an example for the derivation. Given an input tensor$Z \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 32C}$, the comprehensive operation of DWCAM can be expressed as follows:

$$DWCAM(Z) = DSCB(CAB(SAB(Z))) \quad (10)$$

1) The DSCB constitutes the central component of this module, facilitating efficient computation. Its structure has been specifically optimized based on classical design principles, and its computational flow is shown as follows:

$$Y_1 = GU(BN(PWConv(Z', k_7 \times k_7))) \quad (11)$$

$$Y\_2 = GU(BN(DWConv(Y\_1, k\_8 \times k\_8))) \quad (12)$$

$$Y_3 = BN(PWConv(Y_2, k_9 \times k_9))) \quad (13)$$

Where Z' denotes the feature tensor derived following the processing of the input tensor through the SAB and CAB. The operations $PWConv(\cdot)$ and $DWConv(\cdot)$ represent pointwise convolution and depthwise convolution operations, respectively. The structural parameters for these operations were configured as previously described, utilizing kernel sizes of 1×1, 3×3, and 1×1, without spatial downsampling and with appropriate padding. This specific kernel configuration was selected through comparative experiments evaluating different depthwise separable convolution designs. The 1×1 pointwise convolutions enable efficient cross-channel fusion, while the 3×3 depthwise convolution provides sufficient spatial context for local feature refinement. The choice of 3×3 for depthwise convolution is particularly motivated by the feature map resolutions at different stages—it offers an optimal receptive field for capturing local spatial details without introducing excessive parameters, as validated through ablation studies comparing 3×3 with 5×5 and 7×7 alternatives. Batch normalization and GeLU activation were applied subsequent to each convolution to stabilize the feature distribution and enhance the nonlinear representation capability.

This module introduces a learnable residual weight λ, which is a scalar with value range from 0 to 1, to replace traditional fixed-weight connections, as depicted in Equation 14. This design enables the network to dynamically and adaptively adjust the strength of the residual skip connections, thereby improving gradient flow and regularization effects.

$$Z'' = \lambda * Y_3 + Z' \quad (14)$$

2) The spatial and channel attention blocks (SAB and CAB) are engineered to enhance responses in critical spatial regions and important feature channels. This study adjusts their sequence of application. Firstly, the SAB focuses on the most relevant spatial locations in the feature map, followed by the CAB, which recalibrates the importance of the channel features at these locations. This sequence more closely aligns with the cognitive process of fault diagnosis, and empirical results indicate that it achieves superior performance improvement.

In summary, the proposed DWCAM effectively integrates lightweight convolution with attention mechanisms, thereby significantly reducing computational costs while enhancing the cross-layer feature selection and spatial detail modeling capabilities. Unlike MSCF-EFFN, which performs shallow multi-scale feature extraction at the input stage, DWCAM is embedded at the end of each stage, functioning as an inter-stage feature recalibration module to further strengthen the semantic expression and structural consistency of deep features. Moreover, the reintroduction of the depth-wise separable convolution structure in this module is not intended for parameter compression but rather to compensate for the limitations of attention mechanisms in local detail modeling. Depthwise convolution refines spatial responses, whereas pointwise convolution achieves cross-channel fusion, thereby effectively enhancing the representation of key regions while maintaining lightweight characteristics. The two modules complement each other functionally: MSCF-EFFN focuses on shallow multi-scale feature enhancement, whereas DWCAM strengthens spatial detail representation and cross-channel dependencies in deeper stages, enabling comprehensive modeling of fault features and improving discriminative capability.

Within this framework, the model implements hierarchical feature downsampling and global information aggregation through the patch merging module between Stages 2 and 4. This module effectively performs the tasks of patch partitioning and linear embedding at the onset of each stage, maintaining feature spatial consistency during dimensionality reduction. For example, in the patch merging operation of Stage 2, the process encompasses three principal steps: spatial downsampling to reduce spatial resolution, channel concatenation to aggregate neighborhood contextual information, and linear projection to

align with the input dimensions of subsequent swin transformer blocks, as shown in Fig. 5.

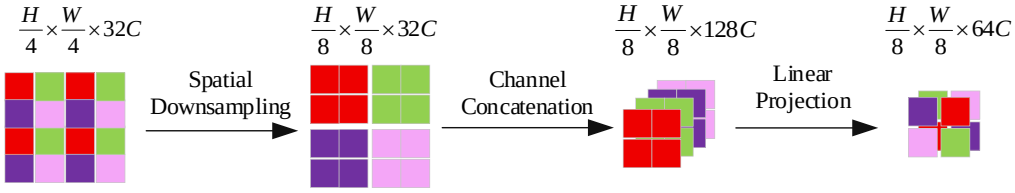

Figure 5. Patch merging.

To further clarify the execution flow of the proposed method and its correspondence with the mathematical formulations presented in Eqs. (1)–(14), the overall training procedure of the network is summarized in Algorithm 1.

Algorithm 1 Training Procedure of the Proposed Network

Input: Time-frequency spectrogram X, corresponding label Y
Output: Updated network parameters θ

Initialize network parameters θ
Extract shallow multi-scale features using MSCF-EFFN
  Compute multi-branch convolution using Eq. (1)
  Concatenate multi-scale features using Eq. (2)
  Apply spatial attention recalibration using Eqs. (3)–(4)
  Apply depth-wise separable convolution using Eq. (5)
Perform patch partition and linear embedding
for stage s=1 to 4 **do**
  for each swin transformer block in stage s **do**
    Compute Q, K, V using Eq. (6)
    Predict deformable offsets using Eq. (7)
    Perform bilinear interpolation sampling using Eqs. (8)–(9)
    Compute deformable attention output using Eq. (6)
  end for
  Apply DWCAM for feature recalibration
  Apply spatial attention block (SAB) and channel attention block (CAB)
  Apply depth-wise separable convolution block using Eqs. (11)–(13)
  Apply learnable residual weighting using Eq. (14)
if s<4 **then**
  Perform patch merging
  end if
end for
Apply global average pooling
Compute classification logits via fully connected layer
Compute cross-entropy loss
Update parameters θ using SGD optimizer

Algorithm 1 systematically illustrates the sequential operations of MSCF-EFFN, DSW-MSA, the DWCAM module, and the hierarchical patch merging strategy, thereby establishing a clear mapping between theoretical derivations and implementation steps.

3. In-wheel motor bearing fault diagnosis method

The fault diagnosis of in-wheel motor bearings is formulated as a supervised multi-class classification problem. Given an input time-frequency spectrogram X, the proposed network learns a nonlinear mapping f(X; θ) to predict the corresponding fault label Y$\in${1,⋯,K} by minimizing the cross-entropy loss. The training process is formulated as an optimization problem [43], where the objective is to minimize the cross-entropy loss with respect to the network parameters θ. The optimization is performed using the stochastic gradient descent (SGD) algorithm. Although the model is trained in a supervised manner, the learned feature representations exhibit an implicit clustering property in the feature space. Specifically, samples belonging to the same fault category tend to form compact clusters, while samples from different categories are well separated. This behavior is consistent with the clustering-like property desired in discriminative feature learning, namely intra-class compactness and inter-class separability. The following sections detail the architecture of f(·), which is designed to extract highly discriminative features for accurate classification under complex operating conditions.

To address the challenges posed by weak fault characteristics and non-stationary signals in in-wheel motor bearings under such conditions, an intelligent fault diagnosis method based on a multi-scale depth-wise separable convolution swin transformer is proposed. This method consists of three primary components: data preprocessing, diagnostic network design, and fault classification. First, during the data preprocessing stage, the original vibration signals are segmented into periodic intervals and analyzed using the CWT for time-frequency analysis, thereby constructing a unified time-frequency feature dataset across multiple sampling rates. Second, to overcome the limitations of the traditional swin transformer in modeling local impacts and preserving spatial details, this study introduces the MSCF-EFFN, DSW-MSA, and

DWCAM to develop a hierarchical feature extraction network. This facilitates the multi-scale information fusion and adaptive modeling of non-stationary features. Finally, the intelligent identification of different bearing faults is accomplished using global average pooling and a softmax classifier.

3.1. Data preprocessing

In the prolonged operation of in-wheel motors, bearings represent some of the most susceptible components to damage. The typical fault conditions include outer race faults, inner race faults, rolling element faults, and normal operating conditions. This study focused on normal state and three fault states of in-wheel motor bearings to construct a dataset for fault diagnosis.

In the process of constructing samples, this study adheres to the principle of complete cycle coverage, ensuring that each signal segment contains at least one full rotation cycle of the motor to preserve periodic characteristics [44]. The sample length was determined based on the relationship between the sampling frequency and rotational speed. For instance, with a sampling frequency of 12.8 kHz and a rotational speed of approximately 750 r/min, one mechanical cycle corresponds to approximately 1024 sampling points. The sample length was not fixed but flexibly adjusted according to the sampling rate and operating conditions to ensure the completeness and representativeness of the extracted signal features. A sliding window segmentation strategy was adopted, and an appropriate overlap rate was set to balance the sample independence and data utilization efficiency.

In terms of feature construction, this study employs a CWT to map time-domain signals into the time-frequency domain, obtaining more discriminative two-dimensional spectrogram inputs. Compared with the short-time Fourier transform (STFT), the CWT offers multi-resolution analysis advantages, allowing for adaptive adjustment of time and frequency resolution based to the signal frequency. Therefore, it can reliably extract rich time-frequency features under different sampling conditions, ranging from 12.8 to 100 kHz. This approach effectively enhances the consistency and robustness of feature representation across different sampling rates, thereby providing unified and high-quality input representations for subsequent deep learning networks.

During the data preprocessing phase, this study established an in-wheel motor bearing fault dataset covering multiple sampling rates through the implementation of sample construction and CWT feature extraction methods. This methodological framework is applicable to four typical fault conditions and demonstrates extensibility, allowing for its adaptation to more complex fault types and a wider range of operating conditions.

3.2. Fault diagnosis network based on multi-scale depth-wise separable convolution swin transformer

Following the completion of data preprocessing and the construction of time-frequency features, the processed data were utilized as inputs into the improved swin transformer network for the purposes of fault feature extraction and classification. While preserving the benefits of the hierarchical architecture, the network sequentially integrates the MSCF-EFFN, DSW-MSA, and DWCAM. This integration enables the model to achieve a stronger cross-scale representation capability at shallow layers, adaptive modeling of non-stationary impact features within each block of the backbone network, and efficient feature screening and enhancement at the output of each stage.

Prior to integration into the backbone network, the input time-frequency spectrograms undergo an initial process of patch partitioning and linear embedding. In this phase, the two-dimensional images are divided into fixed-size patches and subsequently mapped to a high-dimensional feature space. Following this, patch merging operations are conducted across different stages to progressively reduce the feature resolution while simultaneously increasing the channel dimensions, thereby achieving a hierarchical feature representation. This approach ensures that the network is capable of capturing increasingly abstract time-frequency features in a layer-by-layer manner, all while maintaining computational efficiency.

To demonstrate a more intuitive understanding of the network design, this study presents a summary of the system parameter configurations in Table 1. The table delineates the network branches, main modules, and corresponding output dimensions for each stage. During the hierarchical iterative process, the feature resolution is gradually reduced, whereas the channel dimensions are progressively expanded, ultimately

enabling the intelligent classification of *N* fault types. In essence, the proposed structural enhancements are designed to address the challenges of weak transient impacts and cross-scale feature coupling under complex operating conditions. By improving the discriminative capability of the learned representations, the backbone network provides more separable high-level features for the subsequent classification stage.

Table 1. Multi-scale depth-wise separable convolution swin transformer network parameters.

Network branch	Main Modules	Output Size
Input layer	/	224×224×3
MSCF-EFFN	Multi-scale convolution block Spatial attention block Depth-wise separable convolution block	112×112×12
Stage 1	Patch partition Linear embedding Swin transformer block ×2	56×56×48 56×56×96 56×56×96
DWCAM-1	Spatial attention block Channel attention block Depth-wise separable convolution block	56×56×96
Stage 2	Patch merging Swin transformer block ×2	28×28×192
DWCAM-2	DWCAM	28×28×192
Stage 3	Patch merging Swin transformer block ×6	14×14×384
DWCAM-3	DWCAM	14×14×384
Stage 4	Patch merging Swin transformer block ×2	7×7×768
DWCAM-4	DWCAM	7×7×768
Classification layer	Fully connected layer + Softmax	*N classes*

To achieve this, the network parameters were determined as follows. First, the hierarchical architecture (stage depths, channel dimensions, and attention heads) follows the Swin-T baseline configuration, which has been widely validated in prior studies to balance computational complexity and representation capacity through its window-partitioned attention mechanism. This provides a principled starting point for model design. Second, the kernel sizes in the proposed modules were computed based on the characteristic scales of vibration components in the time-frequency domain: large kernels (7×7) correspond to low-frequency trends spanning longer temporal durations, medium kernels (3×3) match mid-frequency resonances, and small kernels (1×1) capture high-frequency transient impacts. These scale mappings are derived from the time-frequency resolution of the input spectrograms. Third, the final configuration was validated through systematic ablation studies to ensure that the chosen parameters achieve an optimal trade-off between diagnostic accuracy and computational efficiency.

Specifically, the proposed modules are theoretically motivated by the fundamental goal of classification: learning a feature representation with high inter-class separability and low intra-class variance. MSCF-EFFN enhances linear separability at the input stage by explicitly encoding multi-scale components. DSW-MSA improves intra-class compactness through deformable attention that adaptively clusters fault-related patterns. DWCAM preserves discriminative information across stages by recalibrating spatial and channel features. Together, they construct a progressively structured feature space that facilitates the final Softmax classification.

3.3. Fault classification

Building upon this progressively structured feature space, the final feature tensor extracted by the proposed architecture serves as the discriminative representation for fault classification and possesses dimensions of 7×7×768, containing 768 channels of high-dimensional semantic features, with each channel corresponding to distinct fault pattern information. To utilize this tensor for fault category prediction, global average pooling (GAP) is first applied to compute the mean value of each channel's 7×7 feature map, thereby compressing it into a one-dimensional vector of length 768. This process retains the

global semantic information of the channels while effectively reducing redundant features and the parameter scale. This 768-dimensional vector is then introduced into a fully connected (FC) layer, which translates the high-dimensional features to the classification space by learning weight parameters, producing a response vector of length N, where N denotes the number of fault categories. To aid in interpretation and decision-making, the softmax function is further applied to normalize this vector into a probability distribution, ensuring that each component P_N signifies the probability of the sample belonging to the N-th fault category, with the total of all category probabilities equal to 1. Finally, the category with the highest probability is selected as the prediction result, achieving an end-to-end intelligent diagnosis from raw signals to fault categories. For example, in a four-class task (normal, outer race fault, inner race fault, rolling element fault), if the softmax output is [0.05, 0.10, 0.80, 0.05], the sample is identified as an inner-race fault.

4. Experimental analysis and verification

To further validate the effectiveness of the proposed MSCF-EFFN-DSW-DWCAM framework derived in Section 3, comprehensive experimental evaluations are conducted under dynamic disturbance conditions. The theoretical modules introduced in the previous section are integrated into a unified training architecture to construct the complete multi-scale depth-wise separable convolution swin transformer network, which is subsequently applied to the in-wheel motor fault diagnosis task. The experimental data are collected from a self-built in-wheel motor fault test bench operating under dynamic disturbance conditions. These dynamic disturbance conditions involve rotational speed fluctuations, load variations, and multi-source noise coupling, which lead to non-stationary and attenuated fault characteristics. Compared with conventional steady-state conditions, such dynamically disturbed operating conditions more closely resemble real-world engineering environments and thus provide a more rigorous framework for evaluating model robustness and generalization capability. Accordingly, the proposed model is systematically evaluated under these dynamic disturbance conditions to comprehensively assess its diagnostic performance in practical engineering applications. The following subsections detail the experimental configuration, dataset construction, and validation methodology.

4.1. Experiments and data

Fig. 6 shows the experimental platform devised to evaluate the in-wheel motor under dynamic disturbances, with the objective of closely replicating real-world vehicle operating conditions. Specifically, the stator of the in-wheel motor was affixed to the suspension using the same installation method as employed in actual vehicles. The motor was embedded inside the wheel hub, with its rotor securely connected to the hub. The system was powered by an electric vehicle and controlled by a dedicated motor controller. A hydraulic platform was used to apply different vertical loads, with the initial height adjusted based on the readings from the pressure sensors situated between the suspension and the fixture, thereby achieving four load conditions: 0, 0.5, 1.0, and 1.5 T. To ensure the reproducibility of the experimental conditions, an STM microcontroller replaced the throttle pedal for precise control at different rotational speeds. Additionally, a roller support structure was installed on the experimental platform, with its upper end contacting the outer surface of the tire and its lower end fixed to the hydraulic platform, thereby further improving the realism of the operating conditions. During the experiments, motor samples with different health conditions were sequentially mounted and tested, including normal operation and bearing fault conditions involving the inner race, outer race, and rolling elements at different severity levels.

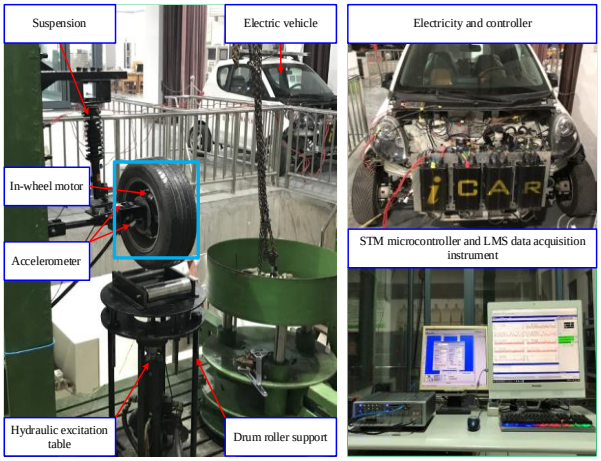

Figure 6. In-wheel motor fault test bench under dynamic disturbances.

In this study, nine bearing fault states were examined and categorized as follows: S1 denotes the normal state; S2–S4

correspond to slight, moderate, and severe inner race faults, respectively; S5–S7 represent slight, moderate, and severe outer race faults, respectively; and S8–S9 indicate moderate and severe rolling element faults.

The experimental conditions encompassed a range of rotational speeds from 100 to 700 r/min, in conjunction with the previously mentioned four vertical loads for comparative analysis. Specifically, seven speed levels and four vertical loads were paired in all possible combinations, resulting in 28 operating conditions that provided comprehensive and diverse data support for subsequent model training and validation.

To facilitate signal acquisition, two accelerometers were installed in the horizontal and vertical orientations of the in-wheel motor stator shaft to capture the vibration signals during operation. The sampling frequency was set to 100 kHz with a sampling duration of 20 s. To ensure that each sample covered a complete vibration cycle, the sample length was set to 10240 points, and a sliding segmentation strategy with a 50% overlap rate was adopted. Based on the total amount of data collected, 300 samples were obtained for each category and subsequently divided into training, validation, and test sets in a 7:1.5:1.5 ratio, thereby providing data support for the construction and validation of the fault diagnosis model.

For model training, a fault diagnosis model was constructed based on the proposed multi-scale depth-wise separable convolution swin transformer network. The model hyperparameters were configured as follows: a batch size of 8, 100 training epochs, the stochastic gradient descent (SGD) optimizer, an initial learning rate of 5×10^{-4}, and a weight decay coefficient of 1×10^{-5}. The learning rate was kept constant throughout the training process to ensure stable convergence. The network was trained in an end-to-end manner using the cross-entropy loss function for multi-class fault classification. The operating environment comprised a Windows 11 system with hardware specifications including an AMD R7-8845H CPU, an NVIDIA RTX 4060 GPU, and 16 GB of RAM. The programming language employed was Python 3.11.7, and the deep learning framework used was PyTorch 2.5.1. Within this environment, training samples were input into the proposed network to iteratively extract and enhance fault features, ultimately producing the classification prediction results.

4.2. Experimental Validation

To validate the effectiveness, adaptability, and superiority of the multi-scale depth-wise separable convolution swin transformer fault diagnosis model, this study provides in-wheel motor fault data subjected to dynamic disturbances, and employs the hold-out method, cross-validation, and controlled variable method for comparative validation to analyze the model's performance in the field of in-wheel motor fault diagnosis.

1) Model Validation Under Identical Operating Conditions. In this experiment, seven different rotational speed conditions were selected under a 0.5 T vertical load, with each rotational speed considered as an identical operating condition. The hold-out method was used to partition the experimental dataset under the same operating conditions into training, testing, and validation subsets, which were used for model training, testing, and validation, respectively. To mitigate experimental randomness, different training samples were selected in the same proportion from the experimental data under identical conditions to train the multi-scale depth-wise separable convolution swin transformer model, while the remaining data were allocated for testing and validation. This procedure was repeated 10 times, and the average value was calculated to determine the state recognition accuracy of the model under the specified operating condition, as illustrated in Fig. 7.

The multi-scale depth-wise separable convolution swin transformer model proposed in this study achieved a diagnostic accuracy exceeding 97.6% for all nine states of in-wheel motor, with an average diagnostic accuracy of 98.7%, thereby demonstrating exceptional generalization performance. Notably, the model exhibited superior identification efficacy for rolling element faults, with its diagnostic rates consistently ranking among the highest. This indicates the model's strong capability in capturing the impact vibration characteristics induced by such faults. Furthermore, at higher rotational speeds ranging from 500 to 700 r/min, the diagnostic accuracy for all fault types experienced a slight decline. This phenomenon can be attributed to the increase in motor operational noise at elevated speeds and the potential proximity of fault characteristic frequencies to the natural frequencies of the transmission system, leading to more complex vibration signal components and certain interference with feature extraction. Nevertheless, the model's minimum recognition rate under these

conditions remained above 97.6%, with fluctuations confined within a reasonable range, fully meeting the accuracy requirements for practical engineering applications and demonstrating the model's satisfactory robustness and practicality.

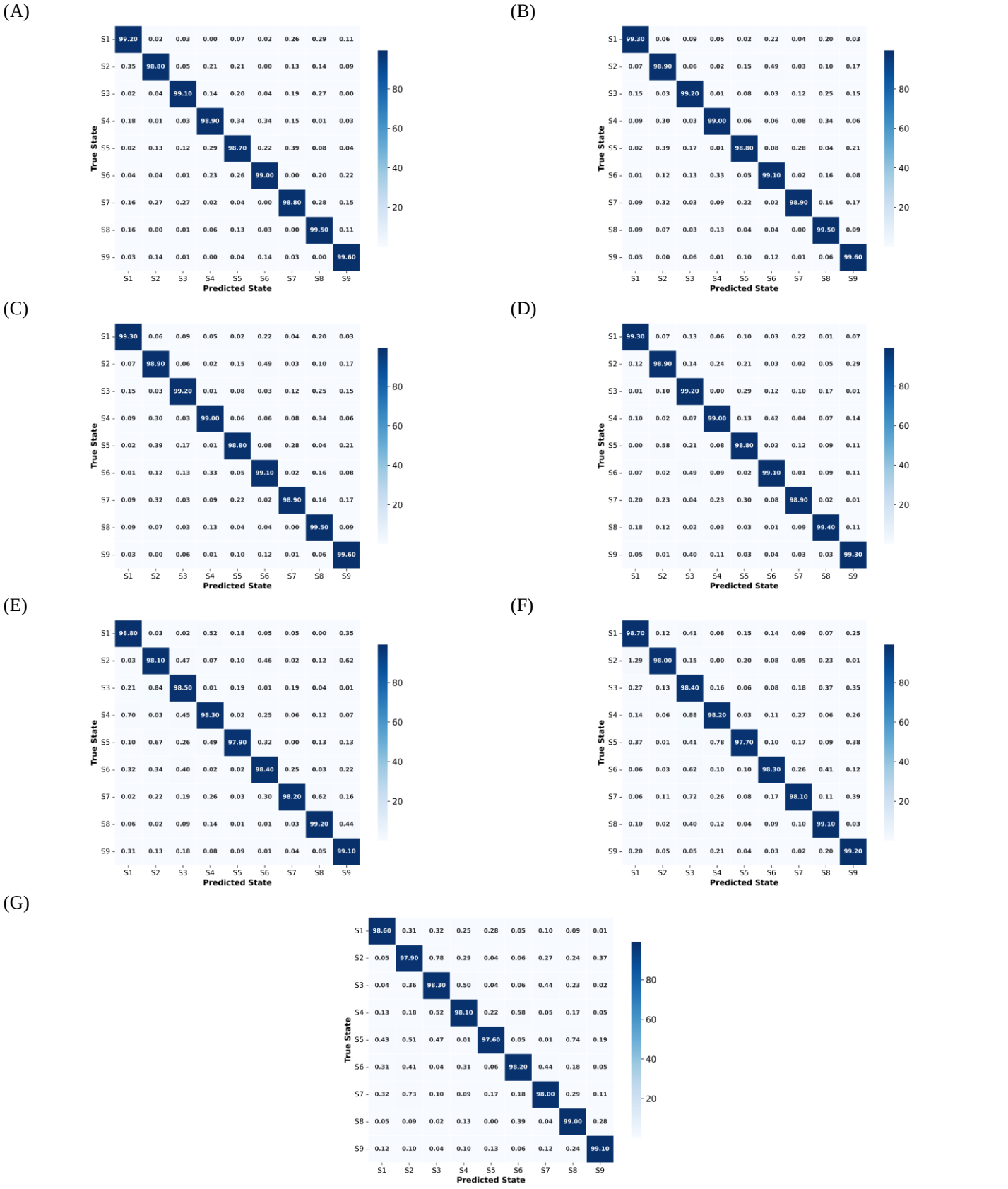

Figure 7. Model validation results under identical operating conditions (A) 100 r/min, (B) 200 r/min, (C) 300 r/min, (D) 400 r/min, (E) 500 r/min, (F) 600 r/min, and (G) 700 r/min.

2) Cross-Condition Model Validation. In practical applications, each in-wheel motor operates under varying conditions. Consequently, cross-validation techniques were adopted to assess the stability and generalization capability of the proposed fault diagnosis model under similar operating conditions. In this context, "similar condition" refers to scenarios characterized by comparable rotational speeds or similar load levels.

Two cross-validation schemes were designed in this study. In the first scheme, the vertical load was held constant, and experimental data from two non-consecutive rotational speeds were selected as training samples, while data from the intermediate speed served as testing samples. For example, under a 0 T load, the data at 100 and 300 r/min were used as training data (labeled as 100 and 300), whereas the data at 200 r/min were used as testing data. In the second scheme, the rotational speed remained constant, and data from two non-consecutive vertical loads were selected as training samples, whereas data from the intermediate load were used as testing samples. For instance, at a rotational speed of 100 r/min, data under 0 and 1 T loads were used as training data (labeled as 0 and 1), whereas data under a 0.5 T load were used as testing data. Throughout the implementation of both cross-validation schemes, the hyperparameters, optimization algorithm, learning rate, and model architecture were unchanged. Table 2 presents the cross-validation schemes and corresponding results for the diagnostic model.

Table 2. Cross-condition model validation.

Condition of Training Data		**Condition of Testing Data**		**Recognition Accuracy of Each State**									**Average Accuracy**
Load	**Speed**	**Load**	**Speed**	S1	S2	S3	S4	S5	S6	S7	S8	S9	
0	100 & 300	0	200	99.3%	98.1%	97.8%	96.9%	99.4%	99.2%	99.3%	99.5%	99.7%	99.1%
0	200 & 400	0	300	97.9%	96.5%	96.8%	95.9%	98.0%	97.7%	97.9%	98.2%	98.4%	97.8%
0.5	200 & 400	0.5	300	98.5%	97.1%	97.3%	96.4%	98.8%	98.6%	98.9%	99.0%	99.2%	98.6%
0.5	300 & 500	0.5	400	96.9%	95.8%	95.5%	94.7%	97.5%	97.1%	97.4%	97.8%	98.0%	97.3%
1.0	300 & 500	1.0	400	97.0%	96.2%	96.3%	95.8%	97.8%	97.6%	98.0%	98.2%	98.6%	97.4%
1.0	400 & 600	1.0	500	94.8%	94.0%	94.3%	93.6%	95.1%	95.3%	95.5%	95.9%	96.1%	95.2%
1.5	400 & 600	1.5	500	95.8%	94.7%	94.9%	94.3%	96.2%	96.0%	96.3%	96.8%	97.2%	96.1%
1.5	500 & 700	1.5	600	98.0%	96.8%	97.0%	96.4%	98.4%	98.2%	98.5%	98.6%	98.9%	98.0%
0 & 1.0	100	0.5	100	99.1%	98.2%	98.4%	97.5%	99.3%	99.0%	99.2%	99.4%	99.6%	99.0%
0 & 1.0	200	0.5	200	97.1%	95.9%	96.3%	95.2%	97.4%	97.0%	97.5%	97.9%	98.2%	97.4%
0 & 1.0	300	0.5	300	96.4%	95.6%	95.8%	95.0%	97.0%	97.1%	97.3%	97.5%	98.1%	96.9%
0.5&1.5	400	1.0	400	98.1%	97.0%	97.2%	96.5%	98.5%	98.3%	98.4%	98.7%	99.0%	98.0%
0.5&1.5	500	1.0	500	97.9%	97.0%	97.1%	96.3%	98.2%	98.0%	98.1%	98.4%	98.8%	97.8%
0.5&1.5	600	1.0	600	97.2%	96.3%	96.5%	95.8%	97.8%	97.6%	97.9%	98.1%	98.6%	97.5%
0.5&1.5	700	1.0	700	97.7%	96.8%	97.1%	96.2%	98.3%	98.1%	98.4%	98.6%	98.9%	97.9%

Overall, the diagnostic results of both cross-validation schemes were satisfactory. Under similar operating conditions, the proposed diagnostic model achieved a recognition accuracy exceeding 93.6% for all operational states of the in-wheel motor, with a normal state identification accuracy consistently above 97.8%. Although the cross-validation accuracy exhibits a slight decrease compared to the results obtained through hold-out validation under identical conditions, it still fulfills the requirements for engineering application. The average recognition accuracy across similar operating conditions is approximately 97.8%, with only one specific condition (using data from 400 r/min and 600 r/min under 1 T vertical load as training data) resulting in an average accuracy below 96%. The comprehensive analysis indicated that the proximity effect of the resonance frequency at 500 r/min significantly influenced the experimental data. Furthermore, the recognition accuracy differed between the two cross-validation schemes. In the first scheme, approximately 50% of the motor states achieved a recognition accuracy of over 97.8% under similar conditions, with the highest reaching 99.1%, although some cases fell below 96%. In the second scheme, the maximum recognition accuracy reached 99.0%, with the average accuracy across all models maintained at a high level. This demonstrates that the multi-scale depth-wise separable convolution swin transformer diagnostic model exhibits greater sensitivity to rotational speed variations than to vertical load changes.

3) Comparative Experiments with Different Models. To further validate the diagnostic adaptability of the proposed method at different rotational speeds, comparative experiments were conducted at four rotational speeds (100, 300, 500, and 700 r/min) under a 0.5 T vertical load. Each experiment was repeated 10 times, and the average values were calculated. The compared models included ViT [45], 2DCNN [46], and the proposed multi-scale depth-wise separable convolution-based swin transformer. All three models were trained and validated using the same data partitioning and parameter configuration. As shown in Fig. 8, the proposed method exhibited superior state recognition accuracy across all rotational speeds compared to the other models, with its performance advantage being particularly notable under conditions significantly influenced by resonance at 300 r/min and 500 r/min. Overall, the proposed method maintained the highest average recognition rate and the smallest fluctuation range across different rotational speeds, thereby demonstrating its stronger robustness and fault feature extraction capability under multi-condition scenarios.

4) Ablation Validation. To further validate the efficacy of the multi-scale depth-wise separable convolution swin transformer model, an ablation study was designed using the controlled variable method. Under the operating condition of a 0.5 T vertical load and a rotational speed of 500 r/min, six comparative schemes were developed by progressively introducing various improved structures: (1) SW-T: The baseline model, using the original swin transformer architecture without any improvement modules. (2) SW-T-MSCAM: Incorporating the MSCAM after each stage to augment multi-scale feature selection in both spatial and channel dimensions. (3) DSW-T-MSCAM: Introducing a deformable attention mechanism based on (2), which allows the attention window to dynamically adjust with the input, thereby improving the adaptive capture capability for non-stationary impact features. (4) EFFN-DSW-T-MSCAM: Adding the MSCF-EFFN at the model input based on (3) to enhance shallow cross-scale representation and accelerate model convergence. (5) DSW-T-DWCAM: Replacing the MSCAM in scheme (3) with the improved DWCAM to evaluate the enhancement effect of DWCAM on feature selection and refinement while maintaining computational efficiency. (6) EFFN-DSW-T-DWCAM: The comprehensive proposed framework, integrating the three improvement modules—MSCF-EFFN, DSW-MSA, and DWCAM—as the final multi-scale depth-wise convolution fault diagnosis model presented in this study. To ensure equitable comparison, all the above schemes were implemented under identical operating conditions and a unified training configuration.

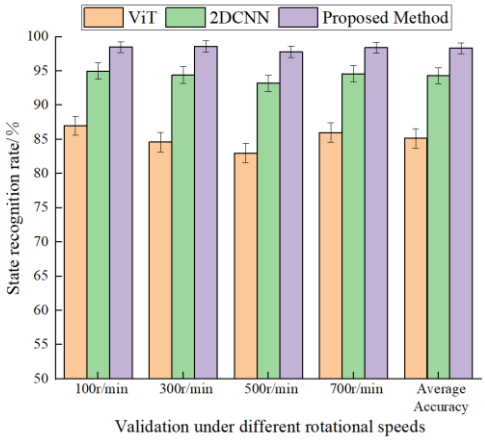

Figure 8. Performance comparison of different models.

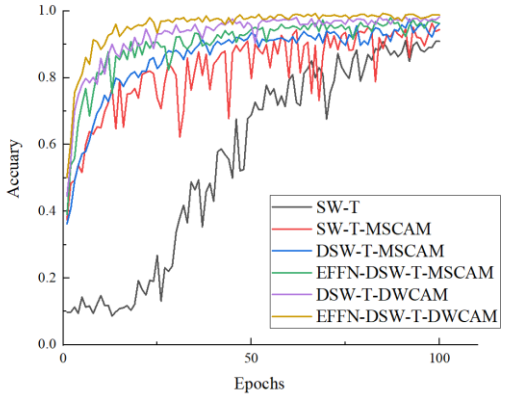

Figure 9. Ablation study accuracy validation.

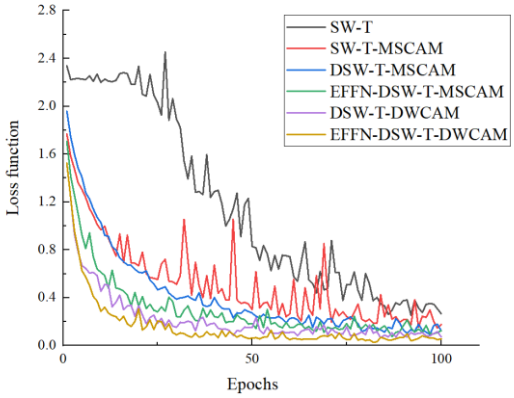

Figure 10. Ablation study loss function validation.

Figs 9 and 10 depict the accuracy and loss convergence curves of the above schemes. The baseline SW-T model demonstrates the slowest convergence behavior, with the accuracy stabilizing only after approximately 80 epochs (around

80% of the total training process) and remaining below 90%. In addition, its loss curve exhibits a gradual decline with pronounced oscillations, indicating limited feature extraction capability and insufficient training stability.

The introduction of MSCAM leads to a marked enhancement in convergence speed, with accuracy stabilizing at approximately 50–60 epochs (50–60% of total training) and reaching about 95%. However, the strong nonlinear interactions induced by MSCAM result in noticeable fluctuations during training. The integration of DSW-MSA further improves convergence stability, achieving stable accuracy at around 45 epochs (approximately 45%) and increasing the final accuracy to approximately 96.5%, thereby demonstrating its efficacy in modeling non-stationary impact features.

The incorporation of MSCF-EFFN at the input stage enables the model to achieve an accuracy over 90% within about 30 epochs (30% of total training), alongside rapid and stable convergence of loss. This finding substantiates the critical role of MSCF-EFFN in enhancing shallow cross-scale features and expediting the optimization process.

A direct comparison between DSW-T-MSCAM and DSW-T-DWCAM indicates that substituting MSCAM with the proposed DWCAM results in a notable enhancement in performance. Under identical network depth and training configurations, DSW-T-DWCAM achieves a higher accuracy of approximately 97.8% with diminished loss fluctuations. This outcome demonstrates that DWCAM facilitates more effective feature selection and refinement without the necessity for shallow feature enhancement modules.

The comprehensive model, EFFN-DSW-T-DWCAM demonstrates exceptional convergence characteristics. It achieves an accuracy exceeding 95% within the initial 20 epochs (20% of total training), with the loss curve converging rapidly and smoothly. Furthermore, the model attains a peak validation accuracy of 99.12%.

The ablation results collectively demonstrate that each proposed module contributes positively to model optimization from different perspectives. Specifically, MSCF-EFFN primarily improves convergence efficiency, DSW-MSA enhances the modeling capability for non-stationary features, and DWCAM further refines feature selection and representation. The synergistic integration of these modules results in significant improvements in convergence behavior and diagnostic accuracy. At the same training length of 100 epochs, the proposed full model converges approximately 40–60% earlier than the baseline SW-T, demonstrating significant improvement in optimization efficiency

5. Conclusion

This study addresses the challenge of weak transient impact feature extraction and cross-scale feature coupling in in-wheel motor fault diagnosis under complex operating conditions. By proposing a multi-scale depth-wise separable convolution swin transformer network, the framework systematically enhances the feature representation capability through three synergistic modules: the MSCF-EFFN enriches shallow multi-scale representations, the DSW-MSA adaptively captures irregular transient impacts, and the DWCAM refines deep feature selection. Together, these components construct a progressively structured feature space with high inter-class separability, which fundamentally improves the classification performance. This behavior also reflects an implicit clustering effect in the learned feature space, where samples from the same fault category are compactly grouped while different categories are well separated

Experimental results on a dynamic disturbance test bench, covering nine health states and 28 operating conditions, demonstrate that the proposed method achieves superior convergence stability and recognition accuracy compared to existing approaches, with a peak accuracy of 99.12%. The ablation studies further validate the individual and collective contributions of the proposed modules, confirming their effectiveness in addressing the inherent limitations of conventional swin transformer architectures for fault diagnosis tasks. Theoretically, this work provides a novel perspective on integrating multi-scale convolution, deformable attention, and channel refinement mechanisms within a hierarchical Transformer framework, offering insights into architectural design for non-stationary signal analysis. Practically, the proposed method exhibits strong robustness under varying operating conditions, highlighting its potential for real-world deployment in electric vehicle intelligent maintenance systems.

Future research will focus on expanding experimental scenarios with real-world road operation data to further enhance generalization capability. Efforts will also be directed toward

lightweight model design and multi-modal sensor fusion to enable real-time monitoring and on-board deployment in vehicular environments.

References

1. Zhao Z, Taghavifar H, Du H, Qin Y, Dong M and Gu L. In-Wheel Motor Vibration Control for Distributed-Driven Electric Vehicles: A Review. IEEE Transactions on Transportation Electrification 2021; 7(4): 2864–2880. https://doi.org/10.1109/TTE.2021.3074970.
2. Yan K, Hu Z, Hu J, Li J, Zhang B, Song J, Li J, Chen L, Li H and Xu L. A critical review of radial field in-wheel motors: technical progress and future trends. eTransportation 2024; 22: 100353. https://doi.org/10.1016/j.etran.2024.100353.
3. Wei J, Zhao Z, Lu E, Liu S, Hu X, Zhou Q, Xu C. Adaptive backstepping tracking control for differential drive vehicles under longitudinal slipping conditions. Biosystems Engineering 2026; 261: 104339. https://doi.org/10.1016/j.biosystemseng.2025.104339.
4. Xue H, Wu M, Zhang Z, Wang H. Intelligent diagnosis of mechanical faults of in-wheel motor based on improved artificial hydrocarbon networks. ISA Transactions 2022; 120: 360–371. https://doi.org/10.1016/j.isatra.2021.03.015.
5. Hussain F. Model predictive control system based on direct yaw moment control for 4WID self-steering agriculture vehicle. International Journal of Agricultural and Biological Engineering 2021; 14(2): 175–181. https://doi.org/10.25165/j.ijabe.20211402.5283.
6. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi A K. Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing 2020; 138: 106587. https://doi.org/10.1016/j.ymssp.2019.106587.
7. Yu Z, Li Y, Du X, Liu Y. Threshing cylinder unbalance detection using a signal extraction method based on parameter-adaptive variational mode decomposition. Biosystems Engineering 2024; 244: 26–41. https://doi.org/10.1016/j.biosystemseng.2024.05.010.
8. Pang J, Li Y, Ji J, Xu L. Vibration excitation identification and control of the cutter of a combine harvester using triaxial accelerometers and partial coherence sorting. Biosystems Engineering 2019; 185: 25–34. https://doi.org/10.1016/j.biosystemseng.2019.02.013.
9. Castilla-Gutierrez J, Fortes Garrido J C, Davila Martin J M and Grande Gil J A. Evaluation procedure for blowing machine monitoring and predicting bearing SKFNU6322 failure by power spectral density. Eksploatacja i Niezawodnosc–Maintenance and Reliability 2021; 23(3): 522-529. https://doi.org/10.17531/ein.2021.3.13.
10. Tao Y, Ge C, Feng H, Xue H, Yao M, Tang H, Liao Z, Chen P. A novel approach for adaptively separating and extracting compound fault features of the in-wheel motor bearing. ISA Transactions 2025; 159: 337–351. https://doi.org/10.1016/j.isatra.2025.01.042.
11. Ma C, Zhang W, Meng L, Yang M, Zhang K, Xu Y. A dual-objective optimized reweighted overlapping group sparse framework integrating frequency slice function for robust bearing fault diagnosis. Mechanical Systems and Signal Processing 2026; 242: 113678. https://doi.org/10.1016/j.ymssp.2025.113678.
12. Guo J, He Q, Zhen D, Gu F. Morphological convolution undecimated wavelet: A novel frequency demodulation analysis method for bearing fault diagnosis. IEEE Transactions on Instrumentation and Measurement 2025; 74: 3522008. https://doi.org/10.1109/TIM.2025.3551821.
13. Li H, Wang T, Zhang F, Chu F. AutoVMDPgram: An effective method for fault diagnosis of rolling bearing. IEEE Transactions on Neural Networks and Learning Systems 2024; 36(8): 15233–15243. https://doi.org/10.1109/TNLS.2024.3518079.
14. Wang B, Xiong Y, Tan L. A high-precision aeroengine bearing fault diagnosis based on spatial enhancement convolution and vision transformer. IEEE Transactions on Instrumentation and Measurement 2025; 74: 1–15. https://doi.org/10.1109/TIM.2024.3502884.
15. Wang J, Zheng J, Pan H, Tong J, Liu Q. Refined composite multiscale slope entropy and its application in rolling bearing fault diagnosis. ISA Transactions 2024; 152: 371–384. https://doi.org/10.1016/j.isatra.2024.07.008.
16. Li J, Luo W, Bai M, Song M. Fault diagnosis of high-speed rolling bearing in the whole life cycle based on improved grey wolf optimizer–least squares support vector machines. Digital Signal Processing 2024; 145: 104345. https://doi.org/10.1016/j.dsp.2023.104345.
17. Chiang H S, Shih D H, Lin B, Shih M H. An APN model for arrhythmic beat classification. Bioinformatics 2014; 30(12): 1785–1786. https://doi.org/10.1093/bioinformatics/btu120.
18. Borlea I D, Precup R E, Dragan F, Borlea A B. Centroid update approach to K-means clustering. Advances in Electrical and Computer Engineering 2017; 17(4): 3–10. https://doi.org/10.4316/AECE.2017.04001.
19. Jing N. Neural network-based pattern recognition in the framework of edge computing. Romanian Journal of Information Science and Technology 2024; 27(1): 106–119. https://doi.org/10.59277/ROMJIST.2024.1.08.

20. Andonovski G, Leite D, Precup R E, Gomide F, Pratama M, Škrjanc I. Advancements in data-driven evolving fuzzy and neuro-fuzzy control: A comprehensive survey. Applied Soft Computing 2025; 186: 114058. https://doi.org/10.1016/j.asoc.2025.114058.
21. Zhang Q, Deng L. An intelligent fault diagnosis method of rolling bearings based on short-time Fourier transform and convolutional neural network. Journal of Failure Analysis and Prevention 2023; 23(2): 795–811. https://doi.org/10.1007/s11668-023-01616-9.
22. Yan R, Shang Z, Xu H, Wen J, Zhao Z, Chen X, Gao R X. Wavelet transform for rotary machine fault diagnosis: 10 years revisited. Mechanical Systems and Signal Processing 2023; 200: 110545. https://doi.org/10.1016/j.ymssp.2023.110545.
23. Li J, Liu Y, Wu X, Kong X, Cai B. Fault diagnosis in open circuit of inverters on electrical discharge milling machines using adaptive Gaussian wavelet convolutional network. Measurement 2025; 248: 116856. https://doi.org/10.1016/j.measurement.2025.116856.
24. Jiang G, Wang J, Wang L, Xie P, Li Y, Li X. An interpretable convolutional neural network with multi-wavelet kernel fusion for intelligent fault diagnosis. Journal of Manufacturing Systems 2023; 70: 18–30. https://doi.org/10.1016/j.jmsy.2023.06.015.
25. Sethi M R, Subba A B, Faisal M, Sahoo S, Raju D K. Fault diagnosis of wind turbine blades with continuous wavelet transform based deep learning model using vibration signal. Engineering Applications of Artificial Intelligence 2024; 138: 109372. https://doi.org/10.1016/j.engappai.2024.109372.
26. Ruan D, Wang J, Yan J, Gühmann C. CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis. Advanced Engineering Informatics 2023; 55: 101877. https://doi.org/10.1016/j.aei.2023.101877.
27. Sun J, Yang F, Cheng J, Wang S, Fu L. Nondestructive identification of soybean protein in minced chicken meat based on hyperspectral imaging and VGG16-SVM. Journal of Food Composition and Analysis 2024; 125: 105713. https://doi.org/10.1016/j.jfca.2023.105713.
28. An Y, Zhang K, Liu Q, Chai Y, Huang X. Rolling bearing fault diagnosis method based on periodic sparse attention and LSTM. IEEE Sensors Journal 2022; 22(12): 12044–12053. https://doi.org/10.1109/JSEN.2022.3173446.
29. Gao D, Zhu Y, Ren Z, Yan K, Kang W. A novel weak fault diagnosis method for rolling bearings based on LSTM considering quasi-periodicity. Knowledge-Based Systems 2021; 231: 107413. https://doi.org/10.1016/j.knosys.2021.107413.
30. Hou Y, Wang J, Chen Z, Ma J, Li T. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer. Engineering Applications of Artificial Intelligence 2023; 124: 106507. https://doi.org/10.1016/j.engappai.2023.106507.
31. Yang Z, Cen J, Liu X, Xiong J, Chen H. Research on bearing fault diagnosis method based on transformer neural network. Measurement Science and Technology 2022; 33(8): 085111. https://doi.org/10.1088/1361-6501/ac66c4.
32. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence 2022; 45(1): 87–110. https://doi.org/10.1109/TPAMI.2022.3152247.
33. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing 2021; 452: 48–62. https://doi.org/10.1016/j.neucom.2021.03.091.
34. Jin X, Xie Y, Wei X, Zhao B, Chen Z, Tan X. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recognition 2022; 121: 108159. https://doi.org/10.1016/j.patcog.2021.108159.
35. Woo S, Park J, Lee J Y, Kweon I S. CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), Munich, GER, 2018: 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.
36. Zhang J, Zhang M, Wang D, Yang M, Liang C. Multi-scale convolutional sparse attention transformer: A lightweight fault diagnosis model for rotating machinery. Neurocomputing 2025; 650: 130934. https://doi.org/10.1016/j.neucom.2025.130934.
37. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2021: 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.
38. Zeng F, Ren X, Wu Q. A fault diagnosis method for motor vibration signals incorporating swin transformer with locally sensitive hash attention. Measurement Science and Technology 2024; 35(4): 046121. https://doi.org/10.1088/1361-6501/ad1cc4.
39. Sun X, Ding H, Li N, Dong X, Sun J, Zheng G. Intelligent fault diagnosis method for shearer rocker gear based on swin transformer and multiscale convolution parallel integration. IEEE Transactions on Instrumentation and Measurement 2025; 74: 1–16. https://doi.org/10.1109/TIM.2025.3551002.
40. Zhou T, Yao D, Yang J, Meng C, Li A, Li X. DRSwin-ST: An intelligent fault diagnosis framework based on dynamic threshold noise reduction and sparse transformer with shifted windows. Reliability Engineering & System Safety 2024; 250: 110327.

https://doi.org/10.1016/j.ress.2024.110327.

41. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020: 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155.
42. Rahman M M, Munir M, Marculescu R. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, 2024: 11769–11779.
43. Adekunle A A, Fofana Issouf, Picher P, Rodriguez-Celis E M, Arroyo-Fernandez O H and Zemouri R. Optimizing deep learning predictive models: A comprehensive review of RNN and its variant architectures. Applied Soft Computing 2025; 185: 114015. https://doi.org/10.1016/j.asoc.2025.114015.
44. Tang H, Zu X, Guo Y, Jiang X, Wang J, Lin R, Xue H, Wang H. A novel incremental method with dynamic learnable pruning mechanism for low-speed machinery fault diagnosis. Engineering Applications of Artificial Intelligence 2026; 166: 113562. https://doi.org/10.1016/j.engappai.2025.113562.
45. Xiang L, Bing H, Li X, Hu A. A frequency channel-attention based vision transformer method for bearing fault identification across different working conditions. Expert Systems with Applications 2025; 262: 125686. https://doi.org/10.1016/j.eswa.2024.125686.
46. Ma Y, Wen G, Cheng S, He X and Mei S. Multimodal convolutional neural network model with information fusion for intelligent fault diagnosis in rotating machinery. Measurement Science and Technology 2022; 33(12): 125109. https://doi.org/10.1088/1361-6501/ac7eb0.