Multi-Branch Joint Time-Frequency Transformer for Domain Generalization Fault Diagnosis of Rotating Machinery

Qitong Chen[1], Liang Chen[2*], Hong Zhuang[3], Qi Li[4] and Wenjing Zhou[5]

[1,2,5] *School of Mechanical and Electric Engineering, Soochow University, Suzhou, Jiangsu, 215000, China*
qtchen0730@163.com
chenl@suda.edu.cn
wjzhou1@stu.suda.edu.cn

[3] *School of Management and Engineering, Nanjing University, Nanjing, Jiangsu, 210000, China*
hzhuang@suda.edu.cn
[4] *Department of Mechanical Engineering, Tsinghua University, Beijing 100000, China*
liq22@tsinghua.org.cn

ABSTRACT

Conventional time-frequency Transformers primarily focus on the global features of signals in the time-frequency domain while neglecting the local features in both the time domain and frequency domain. This limitation hinders the ability of the model to effectively capture the shared features among time, frequency, and time-frequency representations. To address this issue, a Multi-Branch Joint Time-Frequency Transformer (MBJTF-Transformer) is proposed for domain generalization (DG) fault diagnosis of rotating machinery. Specifically, a time-branch Transformer is designed to extract temporal features, while a frequency-branch Transformer captures frequency-domain information. In addition, a time-frequency Transformer is employed to learn the shared representations across time, frequency, and time-frequency domains. Finally, a multi-decision fusion strategy of MBJTF-Transformer is adopted to enhance the generalization capability of the model. Experimental results on both the SCARA (Selective Compliance Assembly Robot Arm, SCARA) dataset and the PU (Paderborn University) bearing dataset demonstrate that the proposed MBJTF-Transformer achieves superior DG performance compared to multiple state-of-the-art sequential models.

1. INTRODUCTION

Rotating machinery has become essential in modern industry, operating under complex and variable working conditions. These variations often lead to distribution shifts between training and real-world data, making fault diagnosis not only crucial but also challenging. As a result, improving domain generalization (DG) performance has become increasingly important to ensure accurate and reliable fault detection across unseen operating conditions, thereby enhancing safety and reducing maintenance costs (Y. Chen, Zhang, Yan, & Xie, 2025).

In recent years, DG-based fault diagnosis (DGFD) has attracted widespread attention from researchers. Ma et al. proposed a 3D dynamic convolutional network for multi-source DG diagnosis of rotating equipment, which leverages feature activation and non-causal suppression to adaptively learn domain-invariant features (Ma, Wei, Zhang, Kong, & Du, 2024). Wang et al. proposed a Multi-Scale Style Generative and Adversarial Contrastive Networks (MSG-ACN) model for machinery fault diagnosis, which generates diverse auxiliary samples and learns domain-invariant features through adversarial contrastive learning to achieve robust performance under unseen working conditions (Wang, Ren, Shen, Huang, & Zhu, 2024). Zhao et al. proposed a semantic-discriminative augmentation-driven network to tackle the challenge of class imbalance in DG. By introducing a semantic regularization-based mixup strategy to enrich minority classes and applying triplet loss to enhance feature discrimination, their method improves generalization to unknown domains (Zhao & Shen, 2024). Chen et al. explored domain-generalizable diagnostic signals from the perspective of signal analysis and designed a lightweight convolutional neural network for DG fault diagnosis of industrial robots, demonstrating certain practical engineering value (Q. Chen, Li, Wu, Chen, & Shen, 2024).

Although these CNN-based methods can achieve DG fault diagnosis with relatively low model complexity, convolutional neural networks often overlook the sequential nature of signals, which may negatively impact diagnostic performance (Xiao, Shao, Wang, Yan, & Liu, 2024). To address this limitation, some researchers have developed sequential models for fault diagnosis of rotating machinery. Xiao et al. proposed a Bayesian variational Transformer (Bayesformer) for training an ensemble of networks, where all attention weights are modeled as random variables to mitigate the adverse effects of operating condition variations and noise interference (Xiao et al., 2024). Liu et al. developed an Attention Contrastive Calibration Transformer (ACCT), which segments time series into blocks to extract domain-invariant features and employs a region confusion-based data augmentation strategy to enhance generalization (Liu, Chen, He, Shi, & Zhou, 2023). Huang et al. introduced a causal Transformer to address out-of-distribution generalization, which builds a causal layer to extract causal relations across domains and incorporates a domain discriminator combined with a gradient reversal layer to learn domain-invariant representations (Huang, Wang, Zhou, Ning, & Song, 2023).

While time-series models effectively consider sequential dependencies and achieve promising diagnostic performance, they often neglect the frequency-domain information of signals. Integrating frequency-domain features into sequential models can help models make more reliable decisions. For example, Ding et al. proposed a Time-Frequency Transformer (TFT) to extract time-frequency features from vibration signals, achieving superior diagnostic accuracy compared to baseline and state-of-the-art methods (Ding, Jia, Miao, & Cao, 2022). Wang et al. developed a Neural-Transformer based on multi-head spatiotemporal impulse self-attention, which extracts global time-frequency features from 2D time-frequency representations and achieves 93.14% diagnostic accuracy on three datasets (Wang et al., 2024). Li et al. proposed the Frequency-Time Modality Transformer (FTM-Transformer) for bearing fault diagnosis, which combines multivariate decomposition of time-frequency features and discrete wavelet transform to extract frequency-domain and time-domain features, outperforming vision Transformers and residual networks (Li, Wang, & Wu, 2023).

Although these time-frequency models capture global features in both time and frequency domains, they rarely consider the role of local features in the time and frequency domains for decision-making. To address this gap, this paper proposes a Multi-Branch Joint Time-Frequency Transformer (MBJTF-Transformer) for DG fault diagnosis of rotating machinery. The main contributions are summarized as follows:

1. The MBJTF-Transformer framework is proposed, in which time-domain, frequency-domain, and time-frequency features are integrated to address the DG problem in rotating machinery fault diagnosis.
2. A multi-branch feature extraction mechanism is designed, consisting of a time-branch Transformer and a frequency-branch Transformer to capture local features, as well as a time-frequency Transformer to extract shared cross-domain representations, thereby overcoming the limitation of conventional models that focus only on global features.
3. A multi-decision fusion strategy is introduced to enhance model generalization, and extensive experiments conducted on the SCARA and PU (Paderborn University) datasets demonstrate that the proposed method outperforms several state-of-the-art sequential models in terms of DG performance.

The remainder of this paper is organized as follows: Section 2 presents the proposed method, Section 3 provides experimental analysis, and Section 4 concludes the paper.

2. METHODS

The overall framework, as illustrated in Figure 1, consists of four main components: time-domain embedding, frequency-domain embedding, time-frequency domain embedding, and a multi-branch joint time-frequency Transformer.

Firstly, the raw input signals are separately segmented into fixed-length patches in the time domain and frequency domain. The time-domain and frequency-domain embedding modules transform these patch sequences into latent representations through a series of linear projections and nonlinear activations, effectively encoding local contextual information. Similarly, the time-frequency embedding module integrates the concatenated time and frequency features to learn joint representations that capture cross-domain correlations.

The extracted embeddings are then fed into three Transformer branches: time, frequency, and time-frequency. The time-domain and frequency-domain Transformers leverage multi-head self-attention to capture global dependencies within each domain, while the time-frequency Transformer employs multi-head cross-attention to model interactions between time and frequency representations.

Finally, the outputs from these three branches are fed into individual classification heads to predict fault classes, which are then fused to obtain the final diagnostic decision. This multi-branch design allows the model to fully utilize both local and global features in the time, frequency, and time-frequency domains, enhancing its generalization capability under unseen operating conditions.

2.1. Time-domain Embedding

The time-domain embedding module is responsible for transforming the segmented time-domain patches into latent representations suitable for sequential modeling. As shown in Figure 1, the raw time-domain input signal has a length of L and consists of C channels. This signal is first divided into

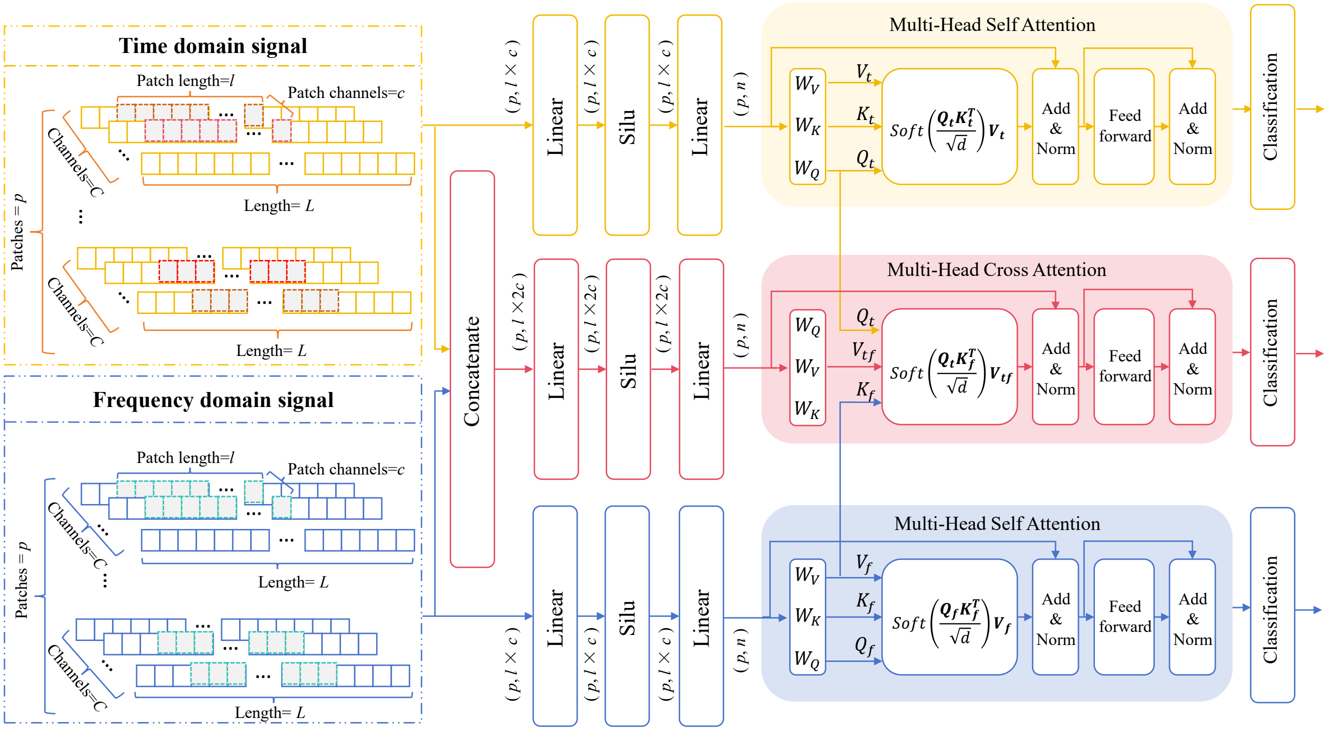

Figure 1. The overall structure of the MBJTF-Transformer.

p overlapping or non-overlapping patches, each with a patch length of l and patch channels c.

Each patch is then flattened to form a vector of dimension $l \times c$. These patch vectors, with an initial shape of $(p, l \times c)$, are projected into a lower-dimensional latent space by a series of linear transformations and a nonlinear activation function (SiLU). Specifically, the embedding process can be expressed as (1):

$$\mathbf{E}_t = \mathrm{Linear}_2(\mathrm{SiLU}(\mathrm{Linear}_1(\mathbf{P}_t))) \qquad (1)$$

where $\mathbf{P}_t \in \mathbb{R}^{p\times(l\times c)}$ denotes the sequence of patch vectors, and $\mathbf{E}_t \in \mathbb{R}^{p\times n}$ is the resulting time-domain embedding matrix, with n being the embedding dimension.

This transformation effectively encodes local contextual patterns within each patch while preserving the sequential structure of the time-series data. The resulting embeddings serve as the input to the subsequent time-domain Transformer branch, enabling the model to learn temporal dependencies for fault diagnosis.

2.2. Frequency-domain Embedding

The frequency-domain embedding module follows a similar procedure to process the frequency-domain representation of the original signal. Formally, the embedding process can be expressed as (2):

$$\mathbf{E}_f = \mathrm{Linear}_2(\mathrm{SiLU}(\mathrm{Linear}_1(\mathbf{P}_f))) \qquad (2)$$

where $\mathbf{P}_f \in \mathbb{R}^{p\times(l\times c)}$ denotes the patch sequence in the frequency domain, and $\mathbf{E}_f \in \mathbb{R}^{p\times n}$ is the resulting embedding matrix. This embedding module captures localized frequency-domain patterns, which complement the temporal information and are crucial for modeling the spectral characteristics of rotating machinery signals.

2.3. Time-frequency Embedding

The time-frequency embedding module aims to capture joint information by combining the features extracted from both time and frequency domains. Specifically, the outputs of the time-domain and frequency-domain patch embeddings (each with shape $p \times c$) are first concatenated along the feature dimension, resulting in a combined representation of shape $p \times 2c$.

This concatenated representation is then passed through an additional linear layer, a SiLU activation, and another linear transformation to project it into the final embedding space. The embedding process can be expressed as (3):

$$\mathbf{E}_{tf} = \mathrm{Linear}_2(\mathrm{SiLU}(\mathrm{Linear}_1(\mathrm{Concat}(\mathbf{P}_t, \mathbf{P}_f)))) \qquad (3)$$

yielding embeddings $\mathbf{E}_{tf} \in \mathbb{R}^{p\times n}$. By fusing time and frequency information early in the feature space, this module enables the subsequent time-frequency Transformer to learn cross-domain dependencies and shared representations, further enhancing the model's ability to generalize under unseen operating conditions.

2.4. Multi-branch Joint Time-Frequency Transformer

After the embedding process, the extracted features from the time domain, frequency domain, and time-frequency domain are separately fed into three Transformer branches to capture complementary information.

The time-domain Transformer branch receives the time-domain embeddings $\mathbf{E}_t \in \mathbb{R}^{p\times n}$ and applies a multi-head self-attention mechanism to learn temporal dependencies within the sequence. Specifically, query, key, and value matrices (Q_t, K_t, V_t) are computed by projecting $\mathbf{E}_t$ through learnable weight matrices W_Q, W_K, W_V. The attention output is calculated as (4):

$$\mathrm{Attention}(Q_t, K_t, V_t) = \mathrm{Softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d}}\right) V_t \qquad (4)$$

where d is the scaling factor related to the embedding dimension. This process enables the model to focus on relevant temporal patterns across patches.

The frequency-domain Transformer branch follows an analogous design. It processes the frequency-domain embeddings $\mathbf{E}_f$ through multi-head self-attention, using its own set of projections (Q_f, K_f, V_f), to model spectral dependencies.

The time-frequency Transformer branch differs in that it employs a multi-head cross-attention mechanism. Here, the concatenated embeddings from the time and frequency domains $\mathbf{E}_{tf}$ serve as the value and key (K_f, V_{tf}), while the time-domain embeddings $\mathbf{E}_t$ are used to compute the query matrix Q_t. The cross-attention output is calculated as (5):

$$\mathrm{Attention}(Q_t, K_f, V_{tf}) = \mathrm{Softmax}\left(\frac{Q_t K_f^\top}{\sqrt{d}}\right) V_{tf} \qquad (5)$$

which allows the model to align and integrate information between temporal and spectral representations.

Each Transformer branch includes standard feed-forward and normalization layers to refine the extracted features. Finally, the outputs from the three branches are passed into separate classification heads to predict fault categories, and their predictions are fused to produce the final diagnostic decision.

Overall, this multi-branch design addresses a key limitation of existing time-frequency models by capturing not only global but also local features in both the time and frequency domains. Through the combination of time, frequency, and time-frequency Transformers, MBJTF-Transformer learns richer and more robust representations that generalize well to unseen working conditions.

2.5. Multi-decision Fusion Strategy

Let the multi-source domain training dataset be $\{(\mathbf{X}_i^s, y_i^s)\}_{i=1}^N$, where $\mathbf{X}_i^s \in \mathbb{R}^{L\times C}$ is the input signal from source domain $s \in \{1, \ldots, S\}$, and $y_i^s \in \{1, \ldots, K\}$ is the corresponding fault label.

After patching and embedding, we obtain three feature sequences $\mathbf{E}_t, \mathbf{E}_f, \mathbf{E}_{tf} \in \mathbb{R}^{p\times n}$. The model employs three Transformer-based feature extractors: time-domain branch $T_t(\cdot)$, frequency-domain branch $T_f(\cdot)$, and time-frequency branch $T_{tf}(\cdot)$, which produce predicted class probabilities. Specifically, the predicted class probabilities for the time domain, frequency domain, and time-frequency domain are calculated by (6), (7), and (8), respectively:

$$\hat{\mathbf{y}}_t = T_t(\mathbf{E}_t) \qquad (6)$$

$$\hat{\mathbf{y}}_f = T_f(\mathbf{E}_f) \qquad (7)$$

$$\hat{\mathbf{y}}_{tf} = T_{tf}(\mathbf{E}_{tf}) \qquad (8)$$

The cross-entropy losses for each branch are calculated by (9), (10), and (11), respectively.

$$\mathcal{L}_t = \mathcal{L}_c(\hat{\mathbf{y}}_t, y^s) \qquad (9)$$

$$\mathcal{L}_f = \mathcal{L}_c(\hat{\mathbf{y}}_f, y^s) \qquad (10)$$

$$\mathcal{L}_{tf} = \mathcal{L}_c(\hat{\mathbf{y}}_{tf}, y^s) \qquad (11)$$

Finally, a multi-decision fusion strategy is applied by averaging the three losses:

$$\mathcal{L}_{fusion} = \frac{\mathcal{L}_t + \mathcal{L}_f + \mathcal{L}_{tf}}{3} \qquad (12)$$

This strategy encourages each branch to collaboratively learn complementary time, frequency, and time-frequency representations, improving generalization to unseen target domains.

2.6. Model Optimization

The model is trained to minimize the fused loss $\mathcal{L}_{fusion}$ defined in (12). The training process adopts the Adam optimizer with a learning rate η, and the model parameters θ are updated iteratively to minimize the objective:

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{fusion}^{(i)} \qquad (13)$$

where N is the total number of training samples across all source domains. The final trained model is expected to generalize well to unseen target domains by leveraging the fused decision from multiple specialized branches.

3. EXPERIMENTS

3.1. Experimental Platform Introduction

The SCARA robot ball screw dataset and the PU bearing dataset were used for fault diagnosis experiments. The SCARA test platform is shown in Figure 2 (Q. Chen, Chen, et al., 2024). Five types of electrical signals from the robot drive motor were collected: J3_Current feedback, J3_U-phase feedback current, J3_V-phase feedback current, J3_W-phase feedback current, and J3_d-axis feedback current. The experiments were conducted under four working conditions with different loads: 0 kg, 3 kg, 6 kg, and 9 kg. The health states of the ball screw include normal, spiral nut stuck, spline nut stuck, and ball shedding, with a sampling frequency of 1600 Hz. The PU test platform is illustrated in Figure 3 (Lessmeier, Kimotho, Zimmer, & Sextro, 2016). Both vibration signals from bearings and phase current signals from the motor were collected at a sampling frequency of 25,600 Hz. The health states of the bearings include normal, outer race fault, and inner race fault. The experiments were conducted under four working conditions, as detailed in Table 1.

Table 1. Four working condition settings of PU dataset.

No.	Rotational speed	Torque	Radial force	Name of Setting
0	1500 rpm	1500 rpm	1000 N	N15_M07_F10
1	900 rpm	900 rpm	1000 N	N09_M07_F10
2	1500 rpm	1500 rpm	1000 N	N15_M01_F10
3	1500 rpm	1500 rpm	400N	N15_M07_F04

To evaluate the diagnostic performance of the model under unseen operating conditions, the four working conditions of both datasets were divided into source domains and unknown target domains. The source domain data are labeled and accessible during training, while the target domain data are unlabeled and inaccessible. The transfer tasks are summarized in Table 2.

Table 2. Transfer tasks settings.

Datasets	Tasks	Source domains	Unknown domains
PU	T0	No.1, 2, 3	No.0
	T1	No.0, 2, 3	No.1
	T2	No.0, 1, 3	No.2
	T3	No.0, 1, 2	No.3
SCARA	T0	3, 6 and 9kg	0kg
	T3	0, 6 and 9kg	3kg
	T6	0, 3 and 9kg	6kg
	T9	0, 3 and 6kg	9kg

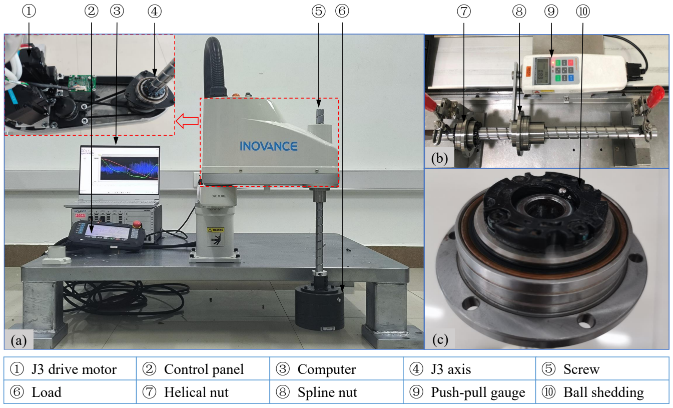

Figure 2. The SCARA robot platform used for experiments.

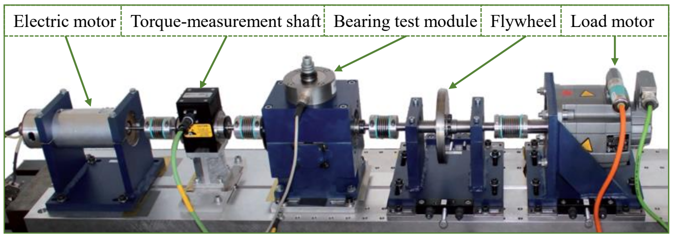

Figure 3. The PU bearing platform used for experiments.

3.2. Experimental setup

All experiments were conducted on a workstation equipped with an Intel i7-9700 CPU and an RTX2080 GPU. To reduce the influence of randomness, each transfer task was independently repeated 10 times. For the SCARA dataset, each class contains 50 samples, while for the PU dataset, each class contains 61 samples. The source domain data were divided into training and validation sets with a ratio of 7:3. The batch size was set to 32.

3.3. Hyperparameter Selection

To ensure optimal performance of the proposed MBJTF-Transformer, a grid search strategy was employed to determine the most suitable learning rate (Q. Chen, Li, et al., 2024). Specifically, the learning rate was varied within the range of [0.00005, 0.0001, 0.0005, 0.001, 0.005], and the average diagnostic accuracy across all DG tasks was used as the evaluation metric. As shown in Figure 4, the model achieved accuracies of 89.03%, 93.03%, 94.92%, 94.72%, and 62.37% under the corresponding learning rates. When the learning rate was too small (e.g., 0.00005), the convergence process was relatively slow, leading to suboptimal accuracy. Conversely, a larger learning rate (e.g., 0.005) caused unstable optimization and severe performance degradation. The learning rate of 0.0005 provided the best trade-off between convergence speed and stability, achieving the highest diagnostic accuracy of 94.92%. Therefore, 0.0005 was selected as the optimal learning rate for all subsequent experiments.

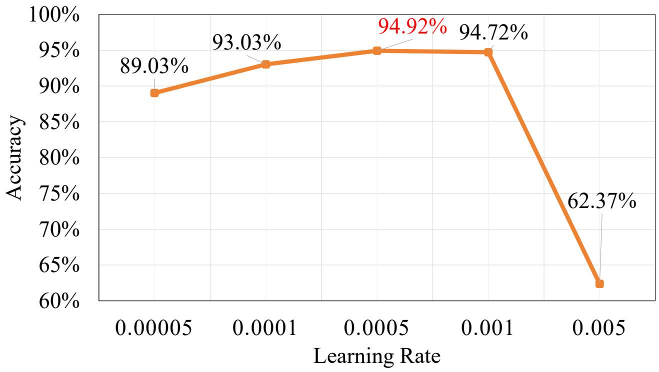

Figure 4. The diagnostic accuracy of the proposed model under different learning rates.

3.4. Comparative experimental analysis

Table 3 represents the DG results on the SCARA and PU datasets, including both ablation experiments and comparisons with several state-of-the-art sequential models: Autoformer (Wu, Xu, Wang, & Long, 2021), Transformer (Vaswani et al., 2017), DLinear (Zeng, Chen, Zhang, & Xu, 2023), FEDformer (T. Zhou et al., 2022), Informer (H. Zhou et al., 2021), LightTS (Zhang et al., 2022), Reformer (Kitaev, Kaiser, & Levskaya, 2020), and ETSformer (Woo, Liu, Sahoo, Kumar, & Hoi, 2022).

To evaluate the contribution of each branch in the proposed MBJTF-Transformer (denoted as MBJTF-T), three ablation variants were designed: MBJTF-T_A1 uses only the output of the time-frequency branch, MBJTF-T_A2 uses only the frequency-domain branch, and MBJTF-T_A3 uses only the time-domain branch for decision-making.

The MBJTF-T achieves the highest average accuracy of 94.92%, significantly outperforming its ablated variants. MBJTF-T_A1 and MBJTF-T_A2 achieve average accuracies of 91.86% and 88.80%, respectively, while MBJTF-T_A3 performs notably worse with 72.65%. These results demonstrate that each individual branch contributes differently to the overall performance, and the multi-decision fusion strategy in MBJTF-T effectively combines their complementary information, leading to better generalization to unseen operating conditions.

Among these benchmark models, Reformer achieved the highest average accuracy (93.36%) across all DG tasks, followed closely by LightTS with an average accuracy of 93.24%. The superior performance of Reformer and LightTS primarily stems from their architectural optimizations for temporal feature modeling and computational efficiency. Reformer incorporates a Locality-Sensitive Hashing (LSH) attention mechanism, which effectively reduces the computational complexity of long-sequence modeling while preserving global dependencies. LightTS, on the other hand, adopts a simple MLP-based structure combined with two refined downsampling strategies, enabling it to capture both local and global dynamic variations with fewer parameters, thereby enhancing model generalization. Other methods, such as Informer (91.62%) and DLinear (90.37%), also demonstrate competitive performance. However, the proposed MBJTF-T outperforms all of them, achieving an average accuracy of 94.92%. This superior performance highlights the effectiveness of integrating time-domain, frequency-domain, and time–frequency representations through a multi-branch architecture combined with a multi-decision fusion strategy. The consistent improvement over Reformer and LightTS further confirms that our model can better capture both local and global features, thereby achieving robust DG performance under unseen operating conditions.

3.5. Training Process Analysis

In the T3 transfer task, the training processes on the PU and SCARA datasets were monitored, as shown in Figure 5. It can be observed that the source-domain test accuracy rises rapidly and converges quickly in the early stages, indicating that the model effectively learns source-domain features. In contrast, the target-domain accuracy increases more slowly at first but gradually approaches the source-domain performance as training progresses, demonstrating strong cross-domain generalization capability. Meanwhile, both source and target domain loss values decrease significantly within the first 20 epochs and then stabilize, suggesting a smooth training process without overfitting. The proposed MBJTF-Transformer exhibits fast convergence and stable transfer characteristics during training, effectively extracting domain-invariant features and achieving reliable cross-domain fault diagnosis performance.

3.6. Feature Visualization Analysis

To further evaluate the DG capability of our model, we conducted feature visualization on the T6 transfer task of the SCARA dataset, as shown in Figure 6. S-Normal indicates source domain samples labeled as Normal, while T-Normal indicates target domain samples labeled as Normal. In the first row of Figure 6, we visualize the feature distributions produced by MBJTF-T and its three ablation variants. In MBJTF-T_A1, the overall clustering is still reasonable, but the class boundaries are less distinct compared to MBJTF-T. In MBJTF-T_A2, samples of the same fault type remain roughly grouped, yet the boundary between classes becomes blurrier, suggesting reduced discriminative power. In MBJTF-T_A3, the feature clusters are still formed, but the separation between different classes is less clear, and the decision boundary is not as sharp as in MBJTF-T. These observations indicate that although each single branch can capture useful fault features to some extent, integrating time, frequency, and time-frequency representations jointly con-

Table 3. Diagnostic results of ablation experiments and comparative experiments.

Methods	SCARA				PU				Average
	T0	T3	T6	T9	T0	T1	T2	T3	
MBJTF-T	74.50±4.76	92.30±6.84	**99.65±0.41**	**97.25±2.49**	99.67±0.53	**99.89±0.23**	96.17±1.91	**99.89±0.23**	**94.92%**
MBJTF-T_A1	66.30±9.79	80.15±9.87	98.15±1.84	92.00±4.89	99.45±0.36	**100.00±0.00**	**99.45±0.68**	99.34±0.72	91.86%
MBJTF-T_A2	56.50±5.73	71.35±3.74	96.40±1.85	90.35±3.28	99.56±0.43	**99.95±0.17**	**98.03±1.39**	98.25±1.63	88.80%
MBJTF-T_A3	74.70±0.54	**98.35±1.00**	**98.25±1.60**	**97.75±2.26**	60.38±5.93	49.13±7.37	34.43±0.81	68.25±6.41	72.65%
Autoformer	45.60±10.07	52.60±19.27	67.00±17.83	72.10±15.95	99.51±0.87	91.58±6.68	**98.74±0.89**	**100.00±0.00**	78.39%
Transformer	61.95±3.88	80.10±4.35	79.70±5.41	71.20±7.62	**99.95±0.17**	97.54±0.97	66.67±0.00	**99.95±0.17**	82.13%
DLinear	74.90±7.00	91.45±1.89	93.40±2.21	87.55±2.95	98.14±1.65	96.89±1.82	83.39±5.32	97.21±1.30	90.37%
FEDformer	58.05±6.40	77.00±5.29	87.80±4.24	81.95±4.45	**99.84±0.37**	97.05±1.91	69.95±3.62	**99.89±0.23**	83.94%
Informer	73.35±5.46	**93.40±2.87**	95.20±1.62	91.75±2.20	99.67±0.38	99.62±0.52	80.22±6.40	99.78±0.38	91.62%
LightTS	**82.35±11.44**	85.80±10.61	97.30±2.67	86.80±4.56	**99.73±0.39**	98.74±1.00	95.46±4.55	99.73±0.53	**93.24%**
Reformer	**85.25±4.81**	**97.25±1.89**	**99.45±0.60**	**94.85±3.29**	**99.95±0.17**	92.40±7.41	78.14±6.26	99.62±0.52	**93.36%**
ETSformer	**83.25±8.30**	92.40±4.50	82.65±11.46	76.75±6.73	99.13±1.37	78.63±13.83	72.68±26.99	96.67±7.40	85.27%

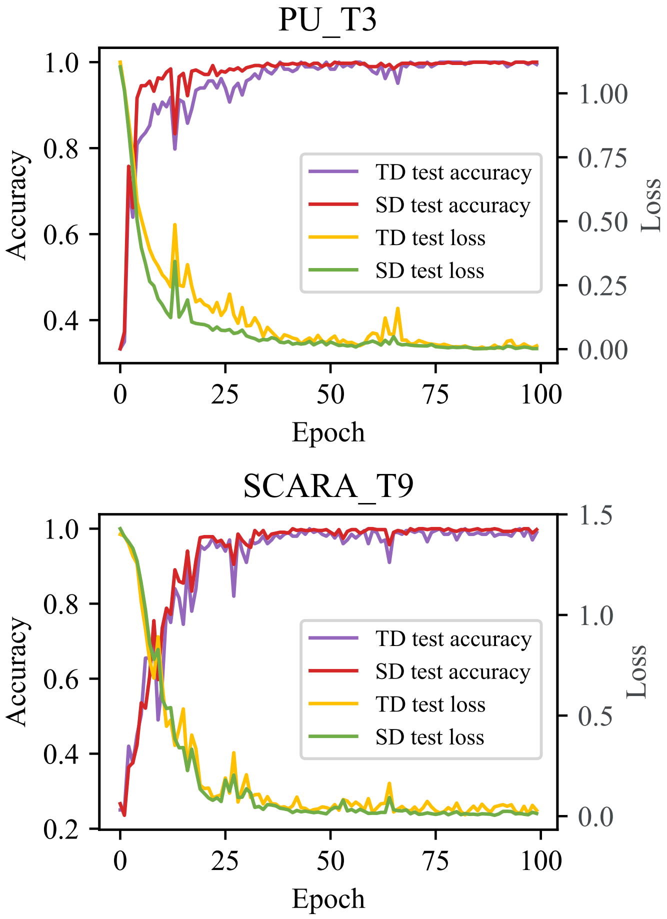

Figure 5. The training process of the MBJTF-Transformer on the T3 transfer task.

tributes to clearer decision boundaries and better domain alignment.

The lower part of Figure 6 presents the feature distributions of eight state-of-the-art sequential models. Among these, Reformer and LightTS exhibit relatively better DG, with source and target domain samples of the same class closer together and clearer boundaries between fault classes. However, the separation is still less distinct than MBJTF-T. Other models such as Autoformer, FEDformer, and Informer show significant overlap across different fault classes and domains, indicating weaker capability to extract domain-invariant features. Overall, the visualization results clearly demonstrate that MBJTF-T captures more discriminative and domain-aligned representations, leading to improved generalization on unseen target domains.

4. CONCLUSION

In this paper, we proposed a novel Multi-Branch Joint Time-Frequency Transformer (MBJTF-T) for DG fault diagnosis of rotating machinery. The model employs separate time, frequency, and time-frequency embedding modules, which are further processed by corresponding Transformer branches to extract complementary local and global features. A multi-decision fusion strategy is introduced to enhance the model's robustness across unseen target domains. Experimental results on the SCARA and PU datasets demonstrate that MBJTF-T significantly outperforms state-of-the-art sequential models in terms of DG performance. The feature visualization analysis further confirms that MBJTF-T can effectively align samples from source and target domains, leading to clearer decision boundaries and improved diagnostic accuracy.

For future work, we plan to explore how to extend the proposed framework to tackle the few-shot fault diagnosis problem. By integrating meric-learning or prototype-based approaches, we aim to enhance the adaptability and practicality of MBJTF-T in real-world industrial applications with scarce labeled data.

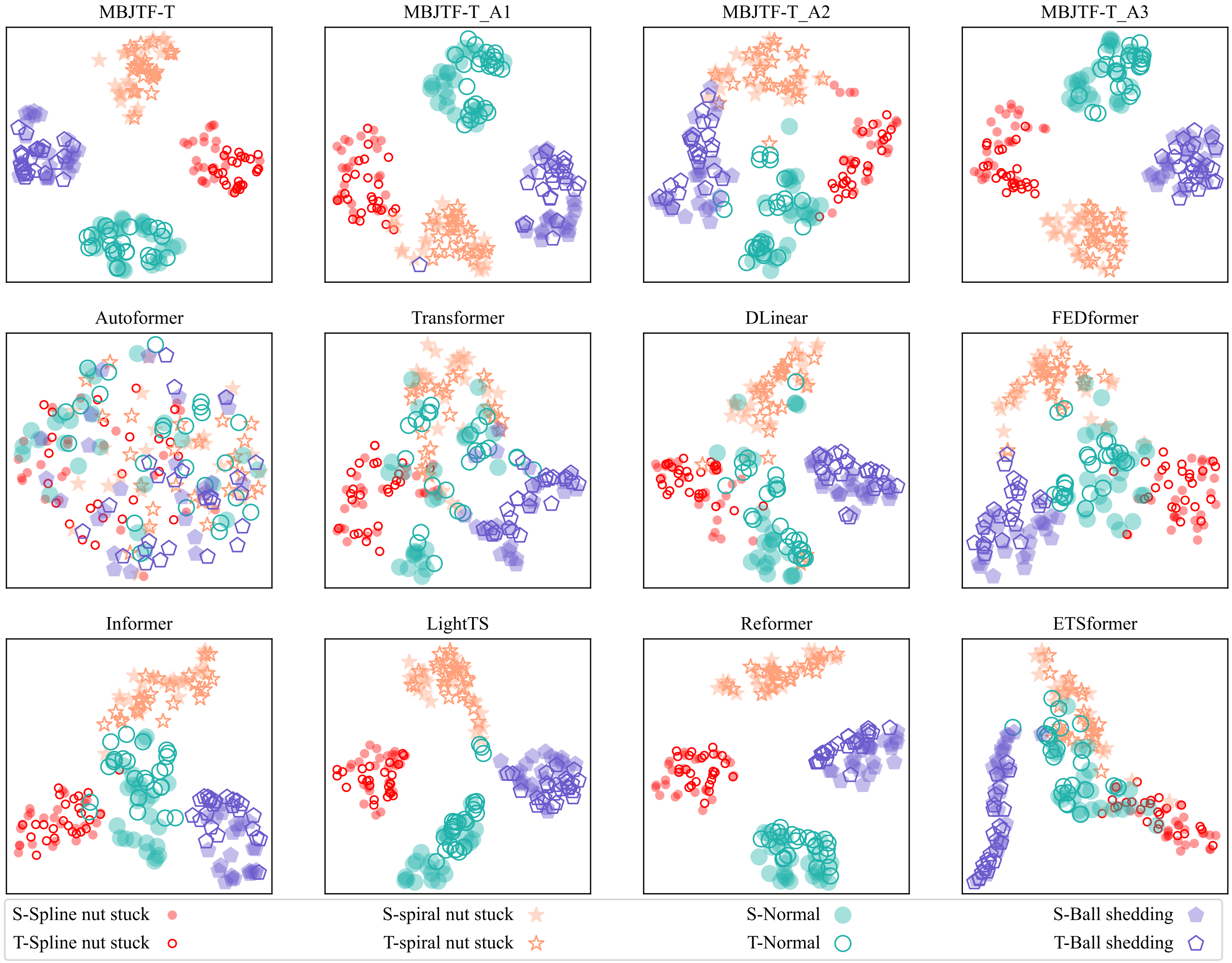

Figure 6. The t-SNE visualization of the feature representations learned by the MBJTF-Transformer.

Acknowledgment

This work was financially supported by the National Natural Science Foundation of China under Grants 52375114, 52272440, and 52175056, and by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX25_3465.

References

Chen, Q., Chen, L., Li, Q., Shi, J., Wang, D., & Shen, C. (2024). Metric learning-based few-shot adversarial domain adaptation: A cross-machine diagnosis method for ball screws of industrial robots. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1-10. doi: 10.1109/TIM.2024.3403183

Chen, Q., Li, Q., Wu, S., Chen, L., & Shen, C. (2024). Fault diagnosis for ball screws in industrial robots under variable and inaccessible working conditions with non-vibration signals. *Advanced Engineering Informatics*, *62*, 102617.

Chen, Y., Zhang, D., Yan, R., & Xie, M. (2025). Applications of domain generalization to machine fault diagnosis: A survey. *IEEE/CAA Journal of Automatica Sinica*.

Ding, Y., Jia, M., Miao, Q., & Cao, Y. (2022). A novel time–frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. *Mechanical Systems and Signal Processing*, *168*, 108616.

Huang, H., Wang, R., Zhou, K., Ning, L., & Song, K. (2023). Causalvit: Domain generalization for chemical engineering process fault detection and diagnosis. *Process Safety and Environmental Protection*, *176*, 155–165.

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *Phm society european conference* (Vol. 3).

Li, K., Wang, C., & Wu, H. (2023). Multimodal transformer for bearing fault diagnosis: A new method based on frequency-time feature decomposition.

Liu, S., Chen, J., He, S., Shi, Z., & Zhou, Z. (2023). Few-shot learning under domain shift: Attentional contrastive calibrated transformer of time series for fault diagnosis under sharp speed variation. *Mechanical systems and signal processing*, *189*, 110071.

Ma, H., Wei, J., Zhang, G., Kong, X., & Du, J. (2024). Causality-inspired multi-source domain generalization method for intelligent fault diagnosis under unknown operating conditions. *Reliability Engineering & System Safety*, *252*, 110439.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, J., Ren, H., Shen, C., Huang, W., & Zhu, Z. (2024). Multi-scale style generative and adversarial contrastive networks for single domain generalization fault diagnosis. *Reliability Engineering & System Safety*, *243*, 109879.

Woo, G., Liu, C., Sahoo, D., Kumar, A., & Hoi, S. (2022). Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, *34*, 22419–22430.

Xiao, Y., Shao, H., Wang, J., Yan, S., & Liu, B. (2024). Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis. *Mechanical Systems and Signal Processing*, *207*, 110936.

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 11121–11128).

Zhang, T., Zhang, Y., Cao, W., Bian, J., Yi, X., Zheng, S., & Li, J. (2022). *Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures.* Retrieved from `https://arxiv.org/abs/2207.01186`

Zhao, C., & Shen, W. (2024). Imbalanced domain generalization via semantic-discriminative augmentation for intelligent fault diagnosis. *Advanced Engineering Informatics*, *59*, 102262.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 11106–11115).

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning* (pp. 27268–27286).

BIOGRAPHIES

Qitong Chen received the B.Eng. degree in electrical engineering from Heilongjiang Bayi Agricultural University, Daqing, China, in 2021. He is currently pursuing the Ph.D. degree in Computer Science and Technology at Soochow University, Suzhou, China. His research interests include fault diagnosis of industrial robots, fault diagnosis of rotating machinery, domain generalization, adversarial learning, few-shot learning, and transfer learning.

Liang Chen received the Ph.D. degree in control engineering from a joint Ph.D. program with Zhejiang University, Hangzhou, China, and TU Berlin, Berlin, Germany, in 2009. He is currently a professor with the Department of Automation Engineering, the School of Mechanical and Electric Engineering, Soochow University, Suzhou, China. His research interests include intelligent control and deep learning-based intelligent sensoring and fault diagnosis.

Hong Zhuang received the B.Eng. degree in Electrical Engineering and Automation from Soochow University Wen Zheng College, Suzhou, China. She is currently pursuing further studies in Engineering Management at Nanjing University, Nanjing, China. Presently, she serves as a research assistant at the School of Mechanical and Electric Engineering, Soochow University. Her research interests include transfer learning and domain generalization.

Qi Li received the B.Eng. and M.Eng. degrees in electrical engineering and control theory & control engineering from Soochow University, Suzhou, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in mechanical engineering with Tsinghua University, Beijing, China. His research interests lie in the intersection of trustworthy AI and reliable prognostic and health management (PHM).

Wenjing Zhou received the B.Eng. degree in light chemical engineering from Soochow University, Suzhou, China, in 2023. She is currently pursuing the M.S. degree in control science and engineering at Soochow University, Suzhou, China. Her research interests include intelligent fault diagnosis of rotating machinery, domain adaptation, domain generalization, and transfer learning.