RESEARCH ARTICLE

Fault Diagnosis Method for Imbalanced Data of Rotating Machinery Based on Time Domain Signal Prediction and SC-ResNeSt

HAITAO WANG, YIFAN GUO, XIANG LIU, JIE YANG, XIHENG ZHANG, AND LICHEN SHI

School of Mechanical and Electrical Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China

Corresponding author: Lichen Shi (bestslc@xauat.edu.cn)

ABSTRACT In an actual engineering environment, some rotating machines are usually in normal operation, but their time in a fault state is very short, which leads to a serious imbalance in the fault diagnosis datasets for rotating machinery, and gives the traditional network model the shortcomings of poor stability and low accuracy in practical engineering applications. To solve this problem, we propose a fault diagnosis method based on the combination of a new Dual-stage Attention-based Recurrent Neural Network (DA-RNN) and depth residual dispersion self-calibration convolution network (SC-ResNeSt). Firstly, a novel DA-RNN network with a gated cycle unit (GRU) as a coding-decoding unit was designed, and the network was used to predict and expand the scarce fault signals. Secondly, to make full use of the time domain information of vibration signals, a new image coding method, namely, Gram Angle Product Field (GAPF), was proposed. Then, because the traditional convolution layer lacks a dynamic receptive field to extract more representative features, self-calibrated convolution modules were introduced on the basis of the distraction network (ResNeSt), and a new network model, SC-ResNeSt, was established. Finally, the expanded vibration signal is converted into GAPF, which is used as the input for the SC-ResNeSt network to classify the fault types. To check the performance of the model, the Case Western Reserve University rolling bearing dataset and planetary gearbox dataset were used for testing. Ultimately, good results were obtained in a prediction experiment for bearing fault samples and a subsequent fault diagnosis experiment for bearings and gears, which verified the feasibility and practicability of the model.

INDEX TERMS Fault diagnosis, Gram angle field, rotary machine, self-calibration convolution, split-attention.

I. INTRODUCTION

With the continuous development of heavy industry, precision mechanical devices have been widely used in aerospace, construction, transportation, and other fields. Some rotating machinery such as bearings or gears as the key components of the mechanical transmission system, and their health plays an important role in the stable operation of mechanical equipment [1], [2]. Any damage to bearings and gears can lead to the sudden failure of the whole mechanical apparatus, resulting in large property losses and even more serious

The associate editor coordinating the review of this manuscript and approving it for publication was Ton Duc Do.

casualties [3]. According to statistics, more than half of mechanical equipment failures are related to bearings or gears [4]. Therefore, building a fault diagnosis model with good practicability and stability is important for avoiding casualties and economic losses, which is also a key topic explored by domestic and foreign experts [5], [6].

Fault diagnosis is important in pursuing the relationship between the monitoring data and the machine's health status [7]. Traditionally, this relationship is caught by abundant experience and expert knowledge of engineers. Some experienced engineers can determine machine malfunctions based on abnormal sounds or locate the bearing faults using advanced signal processing methods that analyze the

vibration signals [8]. However, in the face of multiple conditions, multiple fault information coupling, unknown and variable modes of vibration signals, covering all fault information is simply impossible [9]. In recent years, many Intelligent Fault Diagnosis (IFD) methods have emerged with the help of some machine learning theories. IFD is now widely used in the field of fault diagnosis of rotating machinery since it does not rely on the a priori knowledge of experts Ye et al. [10] developed the Particle Swarm Optimization algorithm based on the Support Vector Machine. They used the multidimensional feature vector of the vibration signal as an input to diagnose bearing faults. He et al. [11] constructed a Modified Deep Auto-Encoder (MDAE) driven by multi-source parameters, which successfully solved the feasibility of aero-engine fault diagnosis in a cross-domain. With the continuous growth in the scope of fault diagnosis objects, the number of sensors installed on the mechanical apparatus to be diagnosed, and the increasing sampling frequency, the obtained data samples have also increased exponentially. Consequently, the fault diagnosis field has joined the big-data era [12]. When the traditional fault diagnosis method faces huge amounts of data, it usually experiences problems such as insufficient model generalization, poor robustness, the difficulty of human operation, and largescale work content. Therefore, in the context of big data, more researchers seek to resolve the problem of fault diagnosis rotating machinery by using Deep Learning (DL) models [13].

Zhu et al. [14] imported Deep Belief Network (DBN) into the multi-sensor information fusion model to identify some unfamiliar fault modes in rotating machinery in an engineering environment and achieved good diagnostic results. Xu et al. [15] developed a hybrid DL model based on the Convolutional Neural Network (CNN) and multi-Grained Cascade Forest (GC-Forest). The inherent fault features of the signals were extracted by CNN and input into a GC-forest classifier to realize the fault mode identification of rolling bearings. Sun et al. [16] proposed an intelligent bearing fault diagnosis model based on multi-domain information fusion and improved residual dense network, and introduced the convolution attention mechanism which can discriminate the importance of features and further improve the feature extraction capability and efficiency of the diagnosis network. However, the vibration signal obtained from the engineering site usually exhibits non-stationary and modulated characteristics due to noise inclusion. From the perspective of deep learning, compared with one-dimensional convolution, the two-dimensional convolution mode can make the network learn more information [17]. Liu et al. [18] converted the original vibration signal into a two-dimensional time-frequency image (TF-image) by a kind of differential continuous wavelet transform and used this image as the input of the network to realize the fault diagnosis of gears. Tong and Zhang et al. [19], [20] converted the signal to Gram Angle Field (GAF) to perform fault diagnosis of rolling bearings to utilize the time domain information of the vibration signal fully Although the DL model has been widely used in the field of fault diagnosis of rotating machinery, it is still characterized by two challenges in engineering applications:

1) In the practical engineering environment, the mechanical apparatus is usually in normal operation, but the machine's time at fault is very short. Most notably, the number of catastrophic or unexpected fault data is low, leading to a serious imbalance in the fault datasets of rotating machinery. Therefore, in practical application, the effect of fault diagnosis based on the DL model may be very poor. The accuracy of fault type identification with more samples is usually very high, while the accuracy of fault identification with fewer samples may be very low [21].

2) Due to the complex structure of rotating machinery in the actual engineering environment, the vibration signals collected by the sensors usually contain strong noise when operating in harsh environments, masking the fault information in the signals. Hence, it is difficult for traditional deep-learning models to extract valuable fault features from these signals [22], [23].

Three mainstream methods exist to solve the sample imbalance problem of rotating machinery in a practical engineering environment One is to expand the number of fault samples based on oversampling [24], [25], [26]. The second method is the Cost-Sensitive (CS) learning method, which can solve the problem of data imbalance in classification duties and has been successfully used for various traditional DL engineering tasks [27], [28], [29], [30]. The third method is solving the problem of poor identification accuracy regarding equipment status caused by the scarcity of available data in the actual engineering environment via Transfer Learning (TL) [31], [32], [33]. However, based on oversampling data processing methods balance the datasets just by increasing the number of fault samples, the sample information is not increased, which may lead to over-fitting of the model [27]. Learning algorithms based on CS usually have difficulty in determining the size of cost-sensitive values and finding effective methods to measure the performance of cost-sensitive classifiers [34]. The fault diagnosis model that employs TL to solve imbalanced data needs to build a source domain data sample that is highly similar to the target domain Sometimes the signal patterns between the target domain data collected by the sensor and the source domain data are different, and the input data of diverse patterns will inevitably lead to difficulty in alignment between the two domains. However, it is very difficult to build a method that can completely align the two domains, and the problem of negative migration can easily occur [35].

In summary, a key issue for improving the effect of imbalanced fault diagnosis is how to exploit the data distribution of minority fault samples and adaptively generate good synthetic samples. Qin et al. [36]. constructed a Dual-stage Attention-based Recurrent Neural Network (DA-RNN),

which introduced input and time attention modules based on the traditional RNN. The authors included several LSTM modules in the coding and decoding unit of the network. The proposed model can adaptively assign different weights to the concomitant time series according to the relationship between the concomitant time series and the target time series achieving good prediction results. Compared to the traditional time series prediction models, the samples generated by DA-RNN can retain inherent nonlinear dynamic characteristics of the original data and the diagnosis results are more convincing [37], [38] Although DA-RNN is characterized by great advantages in extracting the hidden information in a time series and obtaining the long-term dependence of the sequence, a structure with LSTM as the coding and decoding unit often easily deteriorates the model's performance and creates difficulties in training the network due to the complex structure of LSTM [39].

In view of the above challenges, we designed a new DA-RNN network with GRU as the encoding-decoding unit, and used the network to predict and expand sparse fault samples to solve the imbalance problem of rotating machinery data samples in the actual engineering environment; In addition, we combined the split-attention network and self-calibrating convolutional module to construct a new network model, namely, SC-ResNeSt, which can adaptively adjust the size of the convolutional kernel to extract richer fault features according to the actual needs. The main contributions of this paper are as follows:

(1) A Dual-stage Attention-based Recurrent Neural Network with GRU as the encoding and decoding unit is designed. This configuration effectively solves the problems of the complex structure and difficult training in traditional DA-RNN networks. This network is used to predict and expand the fault samples according to the ratio between the normal data samples and the fault data samples of rotating machinery.

(2) To make the most of the time domain information of vibration signals, a new image coding method called Gram Angle Product Field (GAPF) is proposed in this paper. This method offers a unique mapping relationship with the time series and can reflect all the information of the vibration signals.

(3) Because the traditional convolution layer lacks a dynamic receptive field to extract more representative features, we introduce Self-Calibrated Convolution modules on the basis of the ResNeSt, establish a new network model (SC-ResNeSt), and use this model to diagnose the fault types of rotating machinery.

The rest of this paper is arranged as follows: Section II introduces the coding method of the Gram angle field. Section III introduces the DA-RNN and GRU in detail. Section IV introduces the fault diagnosis model proposed in this paper. Sections V and VI outline related experiments and analyses of the results. Finally, Section VII contains the conclusions.

II. GAF

Time domain analysis is the most fundamental signal analysis method, as its theory is simple and easy to reproduce, and provides the truest and most comprehensive information, and does not lose some information due to time-frequency transformations like spectrum analysis methods (CWT, STFT) is one of the most efficient and practical fault diagnosis methods [40], [41] Therefore, when processing the vibration signals of rotating machinery, time domain analysis has been widely used. However, the traditional time domain analysis method cannot achieve high accuracy in the fault diagnosis of rolling bearings. For the above reasons, we present a new method to transform vibration signals into two-dimensional images, namely, the Gram angle field (GAF), and apply this method to the diagnosis of rotary machine fault types. The GAF and time series are corresponding mapping relations that can fully reflect all the information of vibration signals, thus ensuring the dependence of the signal on time. The main process of constructing the Gram angle field is as follows:

1) Firstly, the time series $X = \{x_1, x_2, x_3 \cdots x_n\}$ with length n is normalized to 0 to 1 by (1).

$$\tilde{x}^i = \frac{(x_i - min\ X) + (x_i - max\ X)}{max\ X - min\ X} \tag{1}$$

2) The normalized time series is mapped onto a polar coordinate system by (2). Where t_i is the time stamp, N is a constant factor for adjusting the radial span of the polar coordinates, and $\tilde{X}$ is the normalized time series.

$$\begin{cases} \varphi_i = \arccos\left(\tilde{x}^i\right), & -1 \le \tilde{x}^i \le 1, \tilde{x}^i \in \tilde{X} \\ r_i = \dfrac{t_i}{N}, & t_i \in N \end{cases} \tag{2}$$

This is a new method able to represent the time series in a polar coordinate system. As time changes, the time series $\tilde{X} = \left\{\tilde{x}^1, \tilde{x}^2, \tilde{x}^3 \cdots \tilde{x}^n\right\}$ will be mapped to different angles and radii in polar coordinates creating twists between different corner points on the spanning circle, as shown in Fig. 1. By observing (2), because $-1 \le \tilde{x}^i \le 1, 0 \le \varphi_i \le \pi$, and the cosine function is monotonic in the interval $[0, \pi]$, therefore the coding of the time series in polar coordinates is a unique mapping relationship.

3) The difference/sum operation based on the sine or cosine function gives the Gram angle difference matrix or Gram angle sum matrix, as in the following equation. Equation (3) represents the Gram angle difference matrix, and (4) represents Gram angle sum matrix:

$$\begin{aligned} G_D &= \begin{pmatrix} \sin\left(\varphi_1 - \varphi_1\right) & \cdots & \sin\left(\varphi_1 - \varphi_n\right) \\ \vdots & \ddots & \vdots \\ \sin\left(\varphi_n - \varphi_1\right) & \cdots & \sin\left(\varphi_n - \varphi_n\right) \end{pmatrix} \\ &= \sqrt{I - \tilde{X}'^2} \cdot \tilde{X} - \tilde{X}' \cdot \sqrt{I - \tilde{X}^2} \end{aligned} \tag{3}$$

$$G_S = \begin{pmatrix} \cos\left(\varphi_1 + \varphi_1\right) & \cdots & \cos\left(\varphi_1 + \varphi_n\right) \\ \vdots & \ddots & \vdots \\ \cos\left(\varphi_n + \varphi_1\right) & \cdots & \cos\left(\varphi_n + \varphi_n\right) \end{pmatrix}$$
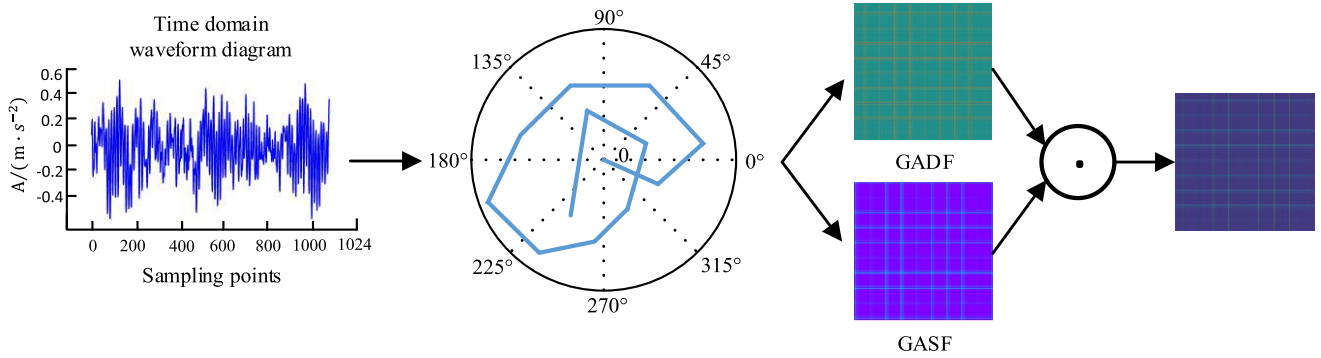
FIGURE 1. Graph angle field coding diagram.

$$= \tilde{X}' \cdot \tilde{X} - \sqrt{I - \tilde{X}'^2} \cdot \sqrt{I - \tilde{X}^2} \tag{4}$$

where I is the unit row vector. After the above formula calculation, the original time series can be transformed into a characteristic matrix distributed symmetrically along the diagonal. This matrix can reflect the information related to the vibration signal which will be called Gram matrix.

4): Finally, by scaling each element in the Gram matrix to between 0 to 255 by (5), thereby converting the two-dimensional matrix into a Gram angle field, where $I(jk)$ represents the pixel value at point (jk) in the image. Here, $int(\cdot)$ is a rounding function, and $G(jk)$ represents the element values corresponding to the j row and the k column in the Gram matrix.

$$I(j,k) = int\left(127.5\left(G\left(j,k\right) + 1\right)\right) \tag{5}$$

When choosing between GADF and GASF, Tong and Zhang were all selected by subjectively contrasting the color difference, rather than strictly using an experiment, which lacks rigor. Therefore, to combine the respective characteristics of GADF and GASF, the point product of the Gram angle difference matrix and Gram angle sum matrix, namely, GADF·GASF, is produced according to (6), as shown at the bottom of the page. From this product, we obtain a new matrix that combines the common characteristics of GADF and GASF: the Gram Angle Product Field (GAPF).

III. DA-RNN

DA-RNN is a prediction network with multiple time series as inputs. In DA-RNN, a dual-module attention structure of the input attention module and time attention module is adopted, and several LSTM basic units are integrated into the Encoder and Decoder. DA-RNN adaptively gives different weights to the concomitant time series according to the relationship between the concomitant time series and the target time series, which solves the problem of the traditional RNN difficulty in capturing the long-term dependence of the time series [36] As a variant of LSTM, Gated Recurrent Unit (GRU) integrates "forget gate" and "input gate" in LSTM into a single "reset gate." Compared to LSTM, GRU lacks a gate in its structure, so the parameters processed by the model are relatively few, and the difficulty of network training is simple. Therefore, we change the coding-decoding unit in the traditional DA-RNN from LSTM to GRU. In the following section, we describe the main principles of the DA-RNN model, Input Attention Mechanism, GRU model and Time Attention Mechanism in turn.

A. INPUT ATTENTION MECHANISM

The Input Attention Mechanism is shown in Fig. 2, assuming n time series $X = \left(x^1, x^2, \cdots, x^n\right)^T$ as an input to this module, which represents n concomitant time series, where T is the length of each time series; $x^k = (x_1^k x_2^k, \cdots, x_T^k)$ represents the k-th time series; And $x_t = (x_t^1 x_t^2, \cdots, x_t^n)$ represents the series information of the above n time series at t moment. In the input attention mechanism, the weight e_t^k of the kth concomitant time series at moment t is first calculated by (7) based on the correlation magnitude of the concomitant time series X to the target time series and combined with the hidden layer state h_{t-1} at moment $t - 1$ Then, the weight e_t^k is normalized by the Softmax layer, as shown in (8). Finally, we obtain the weighted time series $\tilde{x}_t = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \cdots, \alpha_t^n x_t^n)$ as input for the encoder. The hidden layer state h_t is updated by the GRU unit in the encoder to obtain the hidden layer state h_t at moment t, as in (9).

$$e_t^k = V_e tanh\left(W_e\left[h_{t-1}\right]\right) + U_e x^k + b_e \tag{7}$$

$$GAPF = \begin{pmatrix} \sin\left(\varphi_1 - \varphi_1\right) * \cos\left(\varphi_1 + \varphi_1\right) & \cdots & \sin\left(\varphi_1 - \varphi_n\right) * \cos\left(\varphi_1 + \varphi_n\right) \\ \vdots & \ddots & \vdots \\ \sin\left(\varphi_n - \varphi_1\right) * \cos\left(\varphi_n + \varphi_1\right) & \cdots & \sin\left(\varphi_n - \varphi_n\right) * \cos\left(\varphi_n + \varphi_n\right) \end{pmatrix} \tag{6}$$
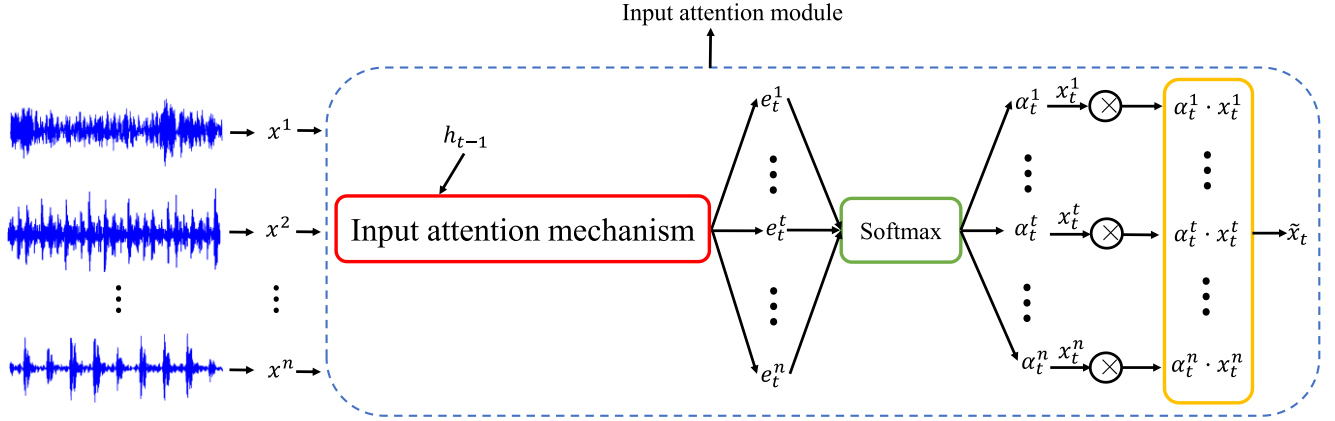
FIGURE 2. Schematic diagram of input attention mechanism.

$$\alpha_t^k = \frac{\exp\left(e_t^k\right)}{\sum_{i=1}^n \exp\left(e_t^i\right)} \tag{8}$$

$$h_t = f_{GRU}\left(h_{t-1}, \tilde{x}_t\right) \tag{9}$$

where e_t^k represents the weight parameter of the k-th concomitant time series at the t moment, and α_t^k represents the weight parameter normalized by the Softmax function; h_{t-1} represents the state of the hidden layer at moment $t-1$, and h_t represents the state of the hidden layer at moment t; x^k represents the k-th concomitant time series; $V_e W_e$, U_e and b_e are pre-training parameters; and $\tilde{x}_t$ represents a new concomitant time series after weighting.

B. GATED RECURRENT UNIT (GRU)

GRU integrates the "forget gate" and "input gate" in LSTM into a single "reset gate" that controls the update and output of data on the hidden layer through the two modules of the "reset gate" and "update gate." The basic unit of GRU is shown in Fig. 3 below. The "reset gate" is mainly used to describe the degree to which the data of the past moment were neglected. The smaller the threshold of this gate, the more data were neglected [42]. In this module, the hidden layer state h_{t-1} at the past moment and the input series x_t at the current moment are taken as inputs, and the output R_t is obtained by nonlinear mapping through the Sigmoid function, as shown in (10). The Sigmoid function can map the information in R_t to an interval of [0, 1]. When $R_t = 0$ it means to completely discard the information passed by the GRU unit in the previous moment, only the input information of the current moment is used as the input. Such a "gate" structure can make the current neuron release some interference data transmitted from the previous neuron, which is helpful in capturing the long-term dependence of the time series.

$$R_t = \sigma\left(x_t \cdot W_{xr} + h_{t-1}W_{hr} + b_r\right) \tag{10}$$

Then, the reset hidden layer data h_{t-1} is written to the candidate hidden layer state $\tilde{h}_t$ by the *tanh* function, as shown in (11):

$$\tilde{h}_t = tanh\left(x_t \cdot W_{xh} + (R_t \cdot h_{t-1})W_{hh} + b_h\right) \tag{11}$$
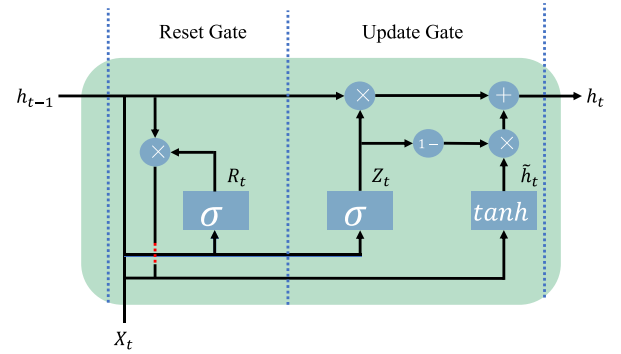

FIGURE 3. Schematic diagram of GRU model.

The main objective of the "update gate" is to control the data update in the hidden layer state h_t at moment t. The module takes the hidden layer output h_{t-1} of the past moment and the input x_t of the current moment as inputs. Then, the output data Z_t of the "update gate" are used to describe the degree to which the hidden layer state information of the past moment is brought into the hidden layer state of the current moment, which can be calculated by (12). The gate threshold is determined by the Sigmoid function. The larger the threshold, the more data the hidden layer state h_{t-1} has to pass to the next hidden layer state h_t

$$Z_t = \sigma\left(x_t \cdot W_{xz} + h_{t-1}W_{hz} + b_z\right) \tag{12}$$

Lastly, the hidden layer data h_{t-1} at the moment $t-1$ and candidate hidden layer data $\tilde{h}_t$ at moment t are combined using (13) to obtain the hidden layer state output h_t at moment t

$$h_t = Z_t \cdot h_{t-1} + (1 - Z_t) \cdot \tilde{h}_t \tag{13}$$

C. TIME ATTENTION MECHANISM

In the process of time series prediction, the features extracted from the concomitant time series are important factors that affect the prediction of a time series. It is not sufficient to
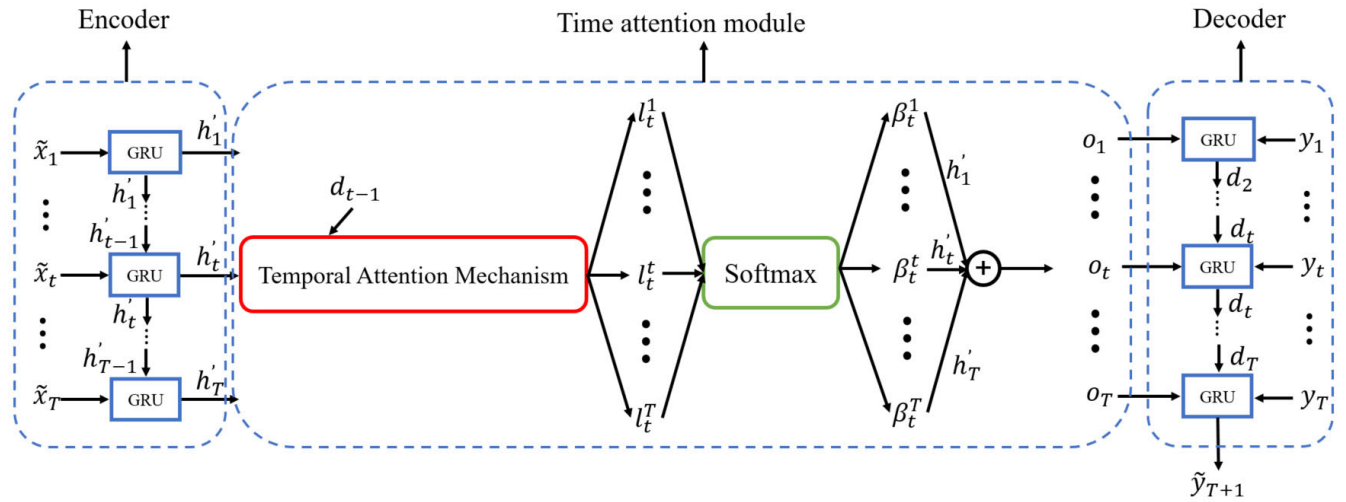
FIGURE 4. Schematic diagram of time attention mechanism.

accurately predict time series only via the spatial correlation between concomitant time series and target time series. As the time series becomes longer, the network will face the problem of gradient explosion and disappearance, so the purpose of introducing the time attention module is to select the hidden layer state that has the greatest correlation with the target series. The time attention mechanism is shown in Fig. 4. This module takes as input the output of the encoder, that is, the state $h'_i = f_{GRU}(h_{i-1}\tilde{x}_i)i \in [1, T]$ of the hidden layer at moment i The time attention weight is calculated as follows:

$$l_t^i = V_d \tanh\left(W_d[d_{t-1}] + U_d h'_i\right) \quad (14)$$

$$\beta_t^i = \frac{\exp(l_t^i)}{\sum_{j=1}^{T} \exp(l_t^i)} \quad (15)$$

where d_{t-1} is the hidden layer state at moment $t-1$; l_t^i is the weight parameter output by the i-th encoder at moment t; β_t^i represents the number of weight parameters normalized by the Softmax function; And V_d, W_d, U_d are the pre-training parameters. The calculation equation for the output O_t for the time attention module is shown in (16):

$$O_t = \sum_{i=1}^{T} \beta_t^i h'_t \quad (16)$$

where $O_t = (O_1 O_2, \cdots, O_T)$ output from the time attention mechanism and the historical target series $Y = (y_1 y_2, \cdots, y_T)$ are spliced together as the input of the decoder. The newly calculated time series $\tilde{Y}_t = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_T)$ is obtained by (17), and the hidden layer state in the decoder is updated by (18). Finally, the GRUs connected in series are used to predict the time series at $T+1$, as shown in (19):

$$\tilde{y}_t = W[y_t, o_t] + \tilde{b} \quad (17)$$

$$d_{t+1} = f_{GRU}(d_t, \tilde{y}_t) \quad (18)$$

$$\tilde{y}_{T+1} = V_y\left(W_y[d_T, o_T] + b_y\right) + b \quad (19)$$

where o_t is the output of the time attention mechanism at moment t; $\tilde{y}_t$ is the linear transformation of y_t; And d_t is the

hidden layer state of the decoder at moment t; $V_y W_y$ and b_y are the pre-training parameters; And $\tilde{y}_{T+1}$ is the predicted value of the target series at moment $T+1$.

To summarize, the main process for time series predictions using the DA-RNN model with GRUs as coding and decoding units are as follows. Firstly, the input of the given model is $X = (x^1 x^2, \cdots, x^n)$, which represents n concomitant time series. The model measures the correlation between n concomitant time series and the target series via the input attention mechanism and gives different weights to the concomitant time series according to the correlation between the concomitant time series and the target time series, thus generating a new concomitant time series $\tilde{X}$. Then, after $\tilde{X}$ is encoded by the encoder, it enters the time attention mechanism. According to the degree of influence on the target series, the attention module adaptively gives a relevant weight to the hidden layer state h_t and generates a new hidden layer state o_t according to the weight. Finally, o_t is decoded by the decoder, and the future target series is predicted by combining the historical target series. Therefore, according to DA-RNN model, we predict the scarce fault signals to expand the vibration signals and solve the serious imbalance between the normal state samples and the fault state samples of rotating machinery in an actual engineering environment.

IV. SC-ResNeSt
A. ResNeSt
With an increase in network depth, the explosion and disappearance of gradients often occur in the network. Therefore, the accuracy of sample identification will not increase with an increase in network layers. The deep residual convolution neural network (ResNet) proposed by He et al. [43] in 2015 not only solves the problem of network degradation well but also has strong feature extraction capabilities. Additionally, the residual block is the core part of ResNet. ResNeSt is a new variant of ResNet that introduces a split-attention
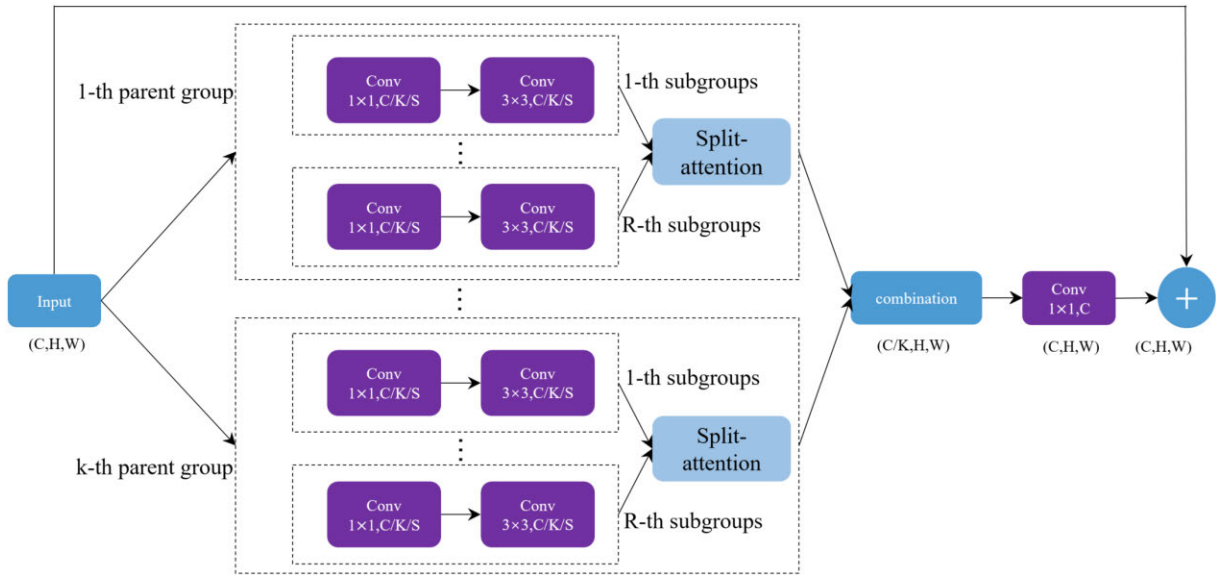
FIGURE 5. ResNeSt basic unit.

mechanism. Compared to ResNet, this new variant has stronger feature extraction capabilities. Fig. 5 shows the basic unit of ResNeSt, whose main principle is as follows. When an image is input into the network, firstly, the image is divided into K parent groups based on the segmentation structure. Then the image continues to divide into R subgroups within each parent group Thus, the image can be divided into $G = R \times K$ subgroups. That is, the whole network model is divided into groups and split. Then, each subgroup is convolved by 1×1 and 3×3 respectively, and it is summarized and input into the split-attention mechanism. Each subgroup feature is given a different weight by the split-attention mechanism. Finally, the characteristic map output by the aggregation is linearly combined with the characteristic map output by the residual module as the final output [44], [45].

The structure of the Split-Attention module is shown in Fig. 6. In this module, the idea of the SK attention mechanism is mainly imitated [46]. The aggregation feature E^k in the k-th parent group is formed by fusing the features of r subgroups:

$$E^k = \sum_{j=R(k-1)+1}^{Rk} E_j \tag{20}$$

where $j \in 1, 2, \cdots, RK; k \epsilon 1, 2, \cdots, K$. Here, E^k represents the aggregation feature in the k-th parent group, and E_j represents the output characteristics of the j-th subgroup, corresponding to the output characteristics based on the traditional 3×3 convolution layer in Fig. 5. R represents the number of subgroups, and k represents the k-th parent group, and K represents the number of parent groups.

The aggregated feature E^k is subjected to global average pooling processing to obtain the global context information of the channel:

$$s_c^k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} E_c^k(i,j) \tag{21}$$

where s_c^k represents the c-th channel value of the $1 \times 1 \times C/K$ feature map of E^k following the global pooling. $E_c^k(ij)$ represents the value at pixel point (i, j) in the c-th channel of the aggregation feature E^k.

Then, each subgroup is classified by the full connection layer, and the weight of each subgroup is adaptively calculated by the Softmax function, as shown in (22):

$$a_i^k(c) = \begin{cases} \dfrac{e^{G_i^c(s^k)}}{\sum_{j=0}^{R} e^{G_i^C(s^k)}}, & R > 1 \\[3mm] \dfrac{1}{1 + e^{-G_i^c(s^k)}}, & R = 1 \end{cases} \tag{22}$$

where $a_i^k(c)$ represents the characteristic weight of the i-th subgroup. G_i^c represents the attention weight function composed of a dense connection layer, full connection layer, and activation function ReLU.

Finally, the weighted features of each subgroup are combined to generate the aggregated output feature V^k of a single parent group. Where the aggregated features of the c-th channel are calculated by (23):

$$V_c^k = \sum_{i=1}^{R} a_i^k(c) E_{R(k-1)+i} \tag{23}$$

where V_c^k represents the weighted aggregation feature of the c-th channel in the k-th parent group, and $E_{R(k-1)+i}$ represents the output feature of the $R(k-1) + i$th subgroup.

B. SELF-CALIBRATION CONVOLUTION

The size of the traditional convolution kernel is fixed, which makes the traditional convolution model does not differentially learn the features of each channel and may cause the network to learn similar information. Such a structure also causes to a lack of dynamically changing sensory fields in the convolutional layer to extract more representative features.
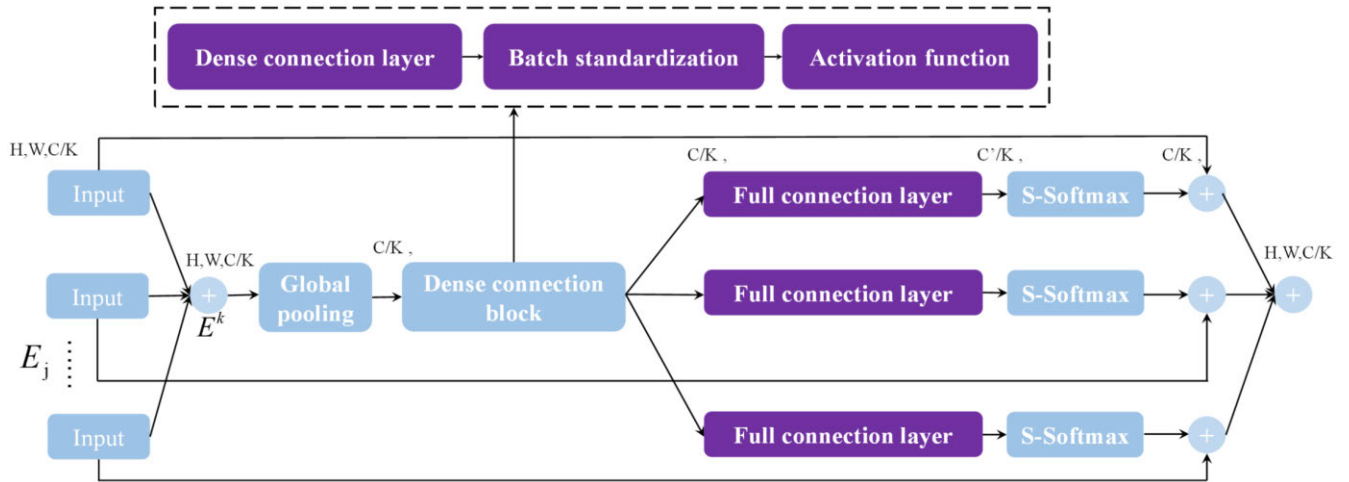
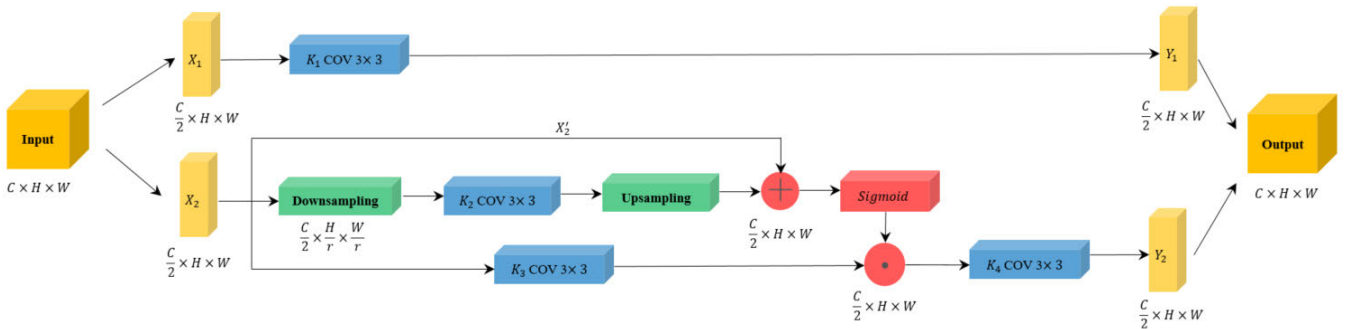**FIGURE 6.** Split-Attention mechanism.

FIGURE 7. 3 × 3 self-calibration convolution module.

In this paper, self-calibration convolution is introduced on the basis of ResNeSt [47], instead of the traditional 3×3 convolution layer in the ResNeSt. The structure of the self-calibration convolution module is shown in Fig. 7. When a $C \times H \times W$ feature map enters the 3×3 self-calibration convolution module, the map is divided into X_1 and X_2 groups, and the size of each group is $C/2 \times H \times W$. In the X_1 group, the traditional 3×3 convolution operation is performed through the convolution channel K_1 to obtain the feature Y_1, whose purpose is to preserve the original spatial features, as shown in (24). In group X_2, the input features are self-calibrated. Firstly, down-sampling processing is carried out for X_2 through the pool layer pair, and the potential space with a smaller resolution is obtained to enlarge the receptive field, and through the convolution channel K_2 and up-sampling processing, the features are extracted and the original resolution is restored. After activation by the activation function Sigmoid, it is used to calibrate the features extracted by the convolution channel K_3 Then, the characteristic Y_2 is output through the convolution channel K_4, as shown in (25). Finally, the outputs of Y_1 and Y_2 groups are aggregated as the final output, as shown in (26). In the self-calibration convolution module, the tasks performed by the four convolution channels

K_1, K_2, K_3, and K_4 are different, which is quite different from the traditional convolution mode:

$$Y_1 = W_{K_1} * W_1 \tag{24}$$
$$Y_2 = \sigma \left(W_{K_2} * X_2 + X_2' \right) * W_{K_3} * W_{K_4} \tag{25}$$
$$Y = (Y_1 + Y_2) \tag{26}$$

where W_{K_1}, W_{K_2}, W_{K_3} and W_{K_4} are the weight parameters of four 3×3 convolution kernels in Fig. 7, respectively, and σ is the activation function. Compared to the traditional convolution mode, the self-calibration convolution mode can adaptively change the size of the convolution kernel receptive field. Only the information of a specific area is considered, thereby avoiding the influence of interference information to a certain extent. Additionally, when the self-calibration convolution module is introduced, no additional network parameters are attached, nor is the complexity of the model changed.

V. FAULT DIAGNOSIS MODEL
In this section, the flow chart in Fig. 8 shows the fault diagnosis model combining DA-RNN with SC-ResNeSt. Firstly, the DA-RNN model with GRU as the coding-decoding unit
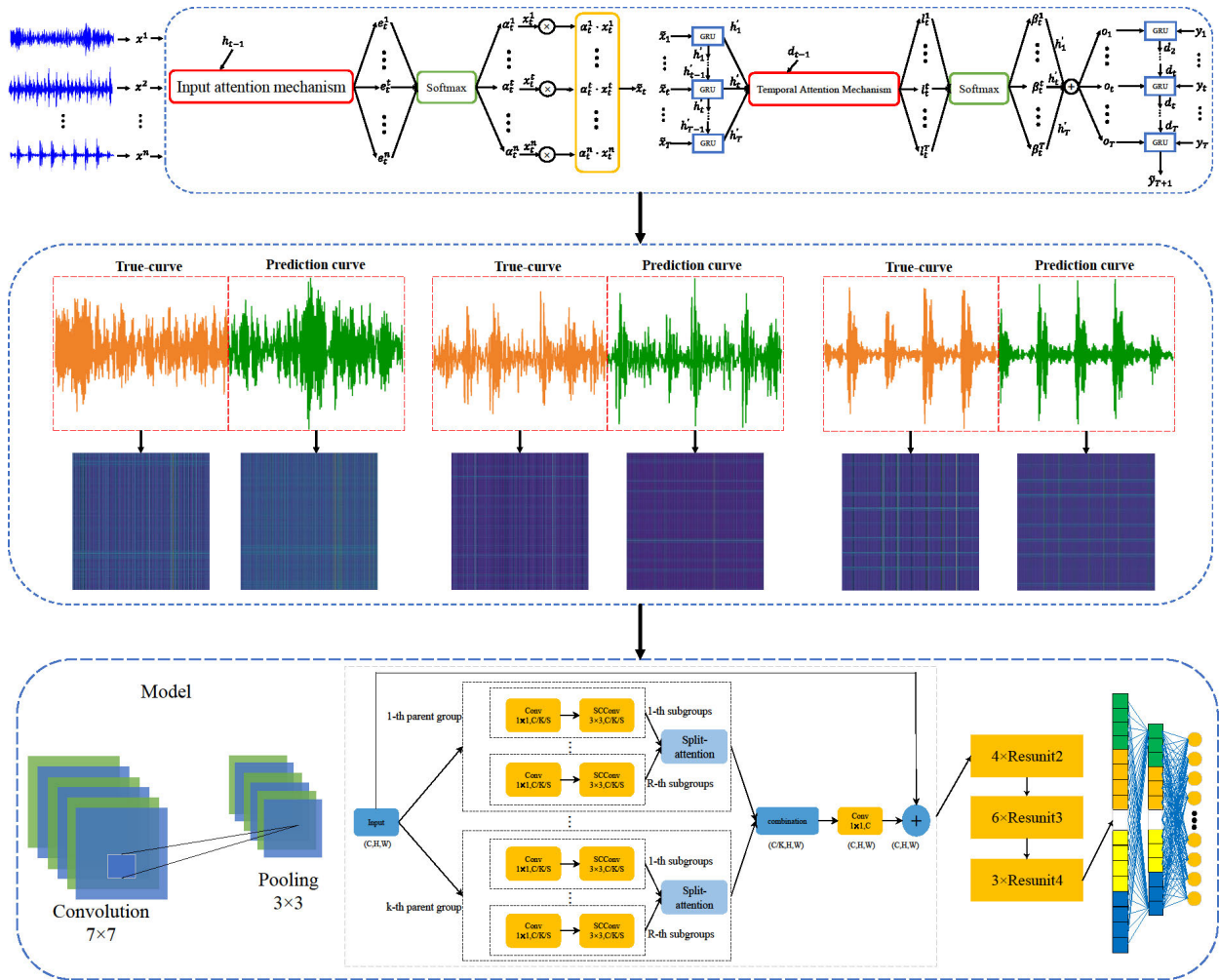
FIGURE 8. Network model.

predicts the target series according to the correlation degree of the concomitant time series to the target time series, so as to realize the expansion of rare fault signals and provide more training samples to the network model. Then, the expanded vibration signal is converted to GAPF as the input of SC-ResNeSt to realize the fault diagnosis of rotating machinery In the SC-ResNeSt model, the input image firstly extracts image features through the convolution layer, maximum pool layer, and four residual mechanisms with self-calibration convolution modules and then classifies fault types through the global pool layer, full connection layer, and Softmax layer. In the residual module, the Batch Normalization (BN) layer and ReLU nonlinear activation function are used to hasten convergence and reduce the calculation amount of the network. The Dropout layer is used at the end of the model to drop some neurons and solve the overfitting problem of the network.

VI. EXPERIMENTAL RESULTS AND ANALYSIS
### A. DATA SOURCE OF ROLLING BEARING
This paper adopts the rolling bearing dataset of Case Western Reserve University in the United States, and Fig. 9 shows the fault data acquisition device. The dataset consists of the vibration signal of the driver end, the vibration signal of the fan end, and the vibration signal of the base. The bearing type is a deep groove ball bearing SKF6205, with a sampling frequency of 12Khz. This dataset contains the vibration signals of rolling bearings under four rotating speeds. The fault types are inner ring fault (IR), outer ring fault (OR) and rolling element fault (RE). The damage diameter of each fault type is 0.18, 0.36, and 0.54 mm, respectively. Including the vibration signals of the normal condition (NC), there are 10 states in total.

B. DATA PREPROCESSING
1) PREDICTIVE EXPANSION OF VIBRATION SIGNAL
The time window T, encoder hidden layer size m, and decoder hidden layer size p are the three parameters in the DA-RNN model that need to be determined in advance. A time window that is too large or too small will affect the performance of the model, and the larger the hidden layer state of the encoder and decoder, the better the prediction effect of the network in the time series. Here, $T = 10$ and $m = 128, p = 128$ are selected, the time window T is used to segment the vibration signal, and the segmented vibration signal is then used as
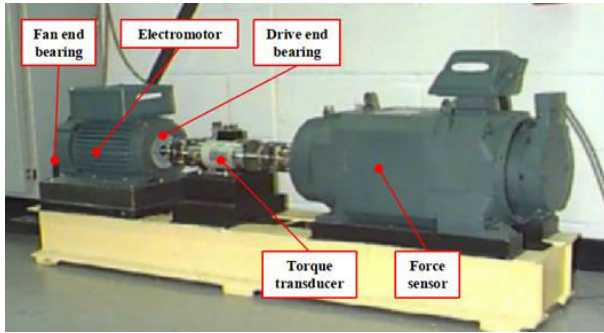
the input of the network model. In this section, the vibration signal of the CWRU dataset under the condition of a rotating speed of 1762 r/min is selected, with the fan vibration signal, base vibration signal, and temperature signal as the concomitant time series, and the vibration signal of the driving end as the target time series. We created six imbalanced datasets based on the ratio between healthy and faulty data samples, e.g., 2:1 for dataset A, 4:1 for dataset B, 5:1 for dataset C, 10:1 for dataset D, 50:1 for dataset E, and 100:1 for dataset F. And we define datasets A and B as mildly unbalanced datasets; Datasets C and D as moderately unbalanced datasets; and datasets E and F as heavily unbalanced datasets, as shown in Table 1. The iteration number of the model is set to 200 epochs, the learning rate is set to 0.001, the learning rate decreases by 10% every 10 iterations, and the number of samples in each batch is set to 32.

Table 2 shows the evaluation of the error between the predicted and true values of the network model after 200 training sessions in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), from which it can be observed that the values of MAE and RMSE gradually increase as the percentage of imbalance decreases. Fig. 10 shows the prediction results from DA-RNN for three failure types of the bearing under six unbalanced ratio conditions. From the figure, it can be observed that its predicted vibration curve is generally consistent with the actual vibration curve of the bearing, and combined with the MAE and RMSE values between the actual curve and the predicted curve in Table 2, it shows that the DA-RNN model has high accuracy.

2) VIBRATION SIGNAL → GAPF
In this section, the bearing vibration signals obtained from the DA-RNN model prediction are encoded as a Gram angle product field (GAPF), and the resolution size of each image is set to 256×256. Taking the imbalance ratio of 2:1 as an example, 1024 points are selected to display the four running states of the bearing, as shown in Fig. 11. In an actual engineering environment, the rolling bearing is in a healthy state most of the time, so there is no need to predict and expand the bearing vibration signal in a healthy state. From the figure, it can be clearly observed that the GAPF codes of different fault

types are quite different, which proves that it is theoretically feasible to use GAPF for fault diagnosis. The experimental results in the following section demonstrate this possibility. In the same fault type, the GAPF coding diagram generated by the actual vibration curve and the predicted vibration curve of the bearing is very similar. Although slightly different in some places, the features in the actual sample have the same distribution pattern as those in the predicted sample. These differences also prove that the DA-RNN model can predict the future target series according to the characteristics of the historical target series. This sample expansion can also increase the generalization and robustness of the whole fault diagnosis model.

t-distribution Stochastic Neighbor Embedding is an unsupervised and nonlinear manifold learning dimensionality reduction algorithm and is also a good feature visualization method [48]. Therefore, t-SNE technology is used to measure the difference between the actual samples and predicted samples of different bearing failures. For brevity, this section only makes the t-SNE visualization of bearing data with an imbalance ratio of 2:1, as shown in Fig. 12. From the figure, it can be found that the image samples generated by the DA-RNN model prediction can be aggregated into a cluster with the actual samples Although there are individual points that are not aggregated together, this further indicates that the DA-RNN model does predict the future target series based on the features of the historical target series.

C. EXPERIMENTAL RESULTS
The hardware and software information used in this experiment were as follows: the 64-bit Windows 10 operating system was used, the GPU was an NVIDIA GTX3070, and the CPU was an Intel i5. This program uses the Adam optimizer as the optimization function to update the training parameters of the network. The activation function is ReLu, and the number of iterations is 50 epochs. The learning rate is set to 0.001, and for every 10 iterations, the learning rate is reduced by 20%. The number of samples per batch is set to 32. To prevent the model from overfitting, we set the dropout value of the Dropout layer to 0.5. At the end of the model, the Softmax classification function is used to classify the fault types.

1) MODEL FEASIBILITY VERIFICATION
Pleiss et al. [49] constructed a Dense Convolutional Network (Densenet) that adopted dense connection blocks based on the traditional ResNet, strengthened the transmission of features from layer to layer, and ensured that the network could make full use of image features. Woo et al. [50] proposed a Convolutional Block Attention Module (CBAM), which is an attention mechanism based on the two dimensions of space and time and can be integrated into any CNN architecture. CBAM is a plug-and-play module and does not add any other redundant training parameters. Xie et al. [51] proposed a simple and highly modular image classification model that introduced

TABLE 1. Case western reserve university Rolling Bearing Failure Test DataSet.

Fault size		×		0.18mm			0.36mm			0.54mm		Imbalance ratio
Fault location		NC	RE	IR	OR	RE	IR	OR	RE	IR	OR	
Fault lable		0	1	2	3	4	5	6	7	8	9	
Train	A	120000	60000	60000	60000	60000	60000	60000	60000	60000	60000	2:1
	B	120000	30000	30000	30000	30000	30000	30000	30000	30000	30000	4:1
	C	120000	24000	24000	24000	24000	24000	24000	24000	24000	24000	5:1
	D	120000	12000	12000	12000	12000	12000	12000	12000	12000	12000	10:1
	E	120000	2400	2400	2400	2400	2400	2400	2400	2400	2400	50:1
	F	120000	1200	1200	1200	1200	1200	1200	1200	1200	1200	100:1
Test	A/B/C/D /E/F	12000	12000	12000	12000	12000	12000	12000	12000	12000	12000	1:1

TABLE 2. Prediction Results Of Six Unbalanced Proportions By da-rnn.

Fault size	Fault location	A		B		C		D		E		F	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
0.18mm	RE	0.049	0.068	0.053	0.07	0.055	0.073	0.057	0.071	0.065	0.082	0.067	0.088
	IR	0.063	0.066	0.066	0.087	0.069	0.093	0.07	0.09	0.135	0.202	0.133	0.191
	OR	0.057	0.06	0.06	0.077	0.063	0.081	0.077	0.096	0.141	0.19	0.149	0.241
0.36mm	RE	0.12	0.16	0.13	0.17	0.129	0.171	0.13	0.173	0.138	0.186	0.145	0.192
	IR	0.071	0.091	0.07	0.093	0.071	0.094	0.073	0.095	0.152	0.251	0.179	0.308
	OR	0.12	0.16	0.14	0.18	0.143	0.184	0.147	0.188	0.165	0.288	0.176	0.295
0.54mm	RE	0.093	0.13	0.11	0.14	0.096	0.14	0.111	0.154	0.169	0.243	0.18	0.253
	IR	0.051	0.064	0.053	0.067	0.056	0.07	0.059	0.77	0.091	0.89	0.175	0.102
	OR	0.121	0.174	0.13	0.18	0.133	0.19	0.139	0.21	0.162	0.813	0.221	0.95

the idea of Group Convolution into the ResNet network, deepened the network in the horizontal space, and improved the classification accuracy of ResNet without increasing the complexity of the network. In this section, the control experiments of ResNet, ResNeXt, CBAM-Net, DenseNet and SC-ResNeSt are added We first analyzed how the imbalance rate of the samples affects the fault diagnosis model. The specific experimental setup was as follows: the six imbalance datasets were not expanded in any way, and after converting them to GAPF, the image samples were divided into training and test sets according to the ratio of 7:3 to serve as the input of ResNet ResNeXt, CBAM-Net, DenseNet, and SC-ResNeSt. The experimental results are shown in Fig. 13. From the figure, we can see that the recognition accuracy of the models was not greatly affected by the various network models when they were faced with mildly imbalance data, and all of them could achieve fault recognition accuracy of more than 98%; When the models were faced with moderately imbalance data, the recognition accuracy of each model decreased by [2%-5%]; When the models face heavily imbalance data, the recognition accuracy of various models decreased to between [50%-60%], and the models have lost the ability to recognize the fault-type.

To verify the feasibility of the fault diagnosis model presented in this paper, the bearing vibration signal of the CWRU bearing dataset at 1762 r/min is selected. Firstly, the DA-RNN model is used to predict and expand the vibration signal according to six imbalanced ratios. Then the signal is converted into GAPF, and the training samples and test samples are split according to the ratio of 7:3. Lastly, it is input into SC-ResNeSt to realize the classification of bearing fault types. At the beginning of training, with an increase in iteration times, the model keeps learning and optimizing until it converges. Fig. 14 shows the fault diagnosis results under six imbalanced proportions. As shown in this Fig. 14, the network is able to achieve more than 90% fault identification accuracy at the beginning of training. After about five iterations, the fault identification accuracy stabilizes around 100%, and the loss function curve steadily declines, indicating that the network is not over-fitted. Ultimately, average fault identification accuracy rates of 99.74%, 99.64%, 99.57%, 99.4%, 99.25% and 9907% were achieved in the experiments under the six imbalanced proportion conditions, which verified the feasibility of this method. To further analyze the detailed classification results of the model for each failure type of the bearing, confusion matrices were drawn for each of the six unbalanced scale conditions, as shown in Fig. 15. In the confusion matrix, the column labels express the real labels of fault types, and the row labels express the prediction labels of fault types. It can be observed from Fig. 15 that all fault types of bearings can achieve very high recognition accuracy under the condition of six proportions,
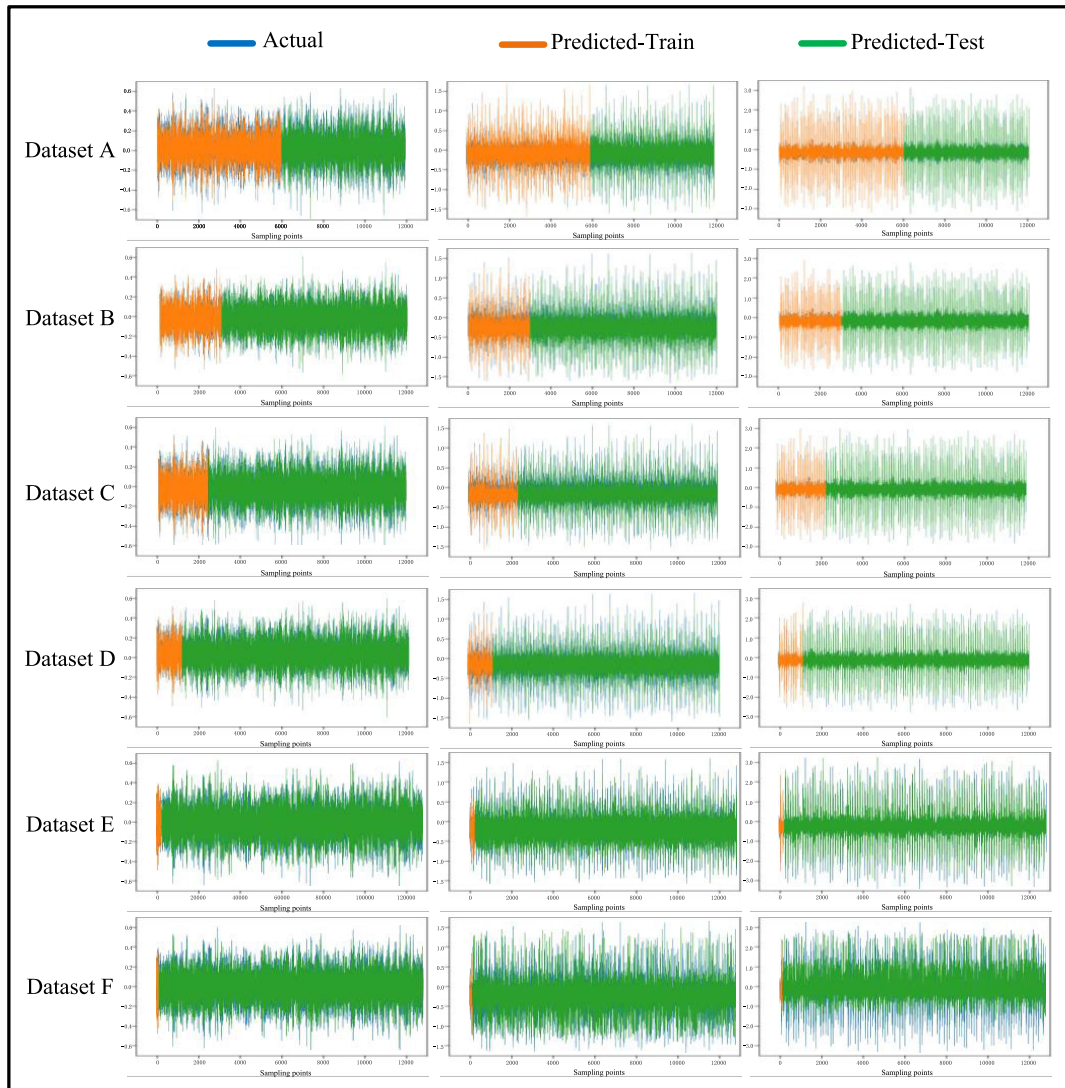
FIGURE 10. DA-RNN's schematic diagram of prediction results of six unbalanced proportions.

which further demonstrates the feasibility of our fault diagnosis model.

To verify the superiority of the SC-ResNeSt fault diagnosis model, the data samples obtained by DA-RNN expansion are used as the input of each network model, and the experimental results are shown in Fig 16 It can be seen in Fig 16 that the SC-ResNeSt fault diagnosis model proposed in this paper can achieve more than 99% fault identification accuracy under six imbalanced proportions when the same signal prediction expansion method is adopted. Moreover, no matter what kind of deep learning model is adopted, with an increase in the imbalance ratio, the accuracy of fault identification will change in a small range, which further shows that the vibration signal predicted by the DA-RNN model with GRU as a coding and decoding unit can be comparable to the original vibration signal, proving that the DA-RNN has high prediction accuracy. In summary, the fault diagnosis model

based on the combination of DA-RNN and SC-ResNeSt proposed in this paper can well solve the severe imbalance problem of bearing samples in the actual engineering environment.

At present, deep learning models have made numerous achievements in visual recognition, natural language processing, reinforcement learning, and other tasks. However, due to the over-parameterized black-box nature of the model, its interpretability in feature extraction still cannot be solved. It is usually difficult to explain the classification results of the model theoretically [52]. Therefore, to verify the effect of SC-ResNeSt on GAPF feature extraction layer by layer, the t-SNE algorithm is used to visualize the features of the output results of the Input Layer, 1-st Residual Block, 3-rd Residual Block and Output Layer of the network. In this section, the fault dataset with an imbalanced ratio of 2:1 is used as the input, and the visualization results of t-SNE characteristics
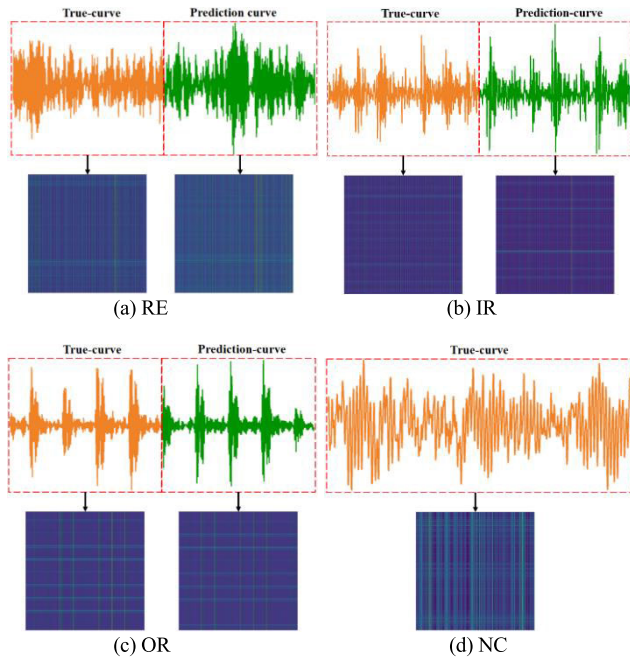
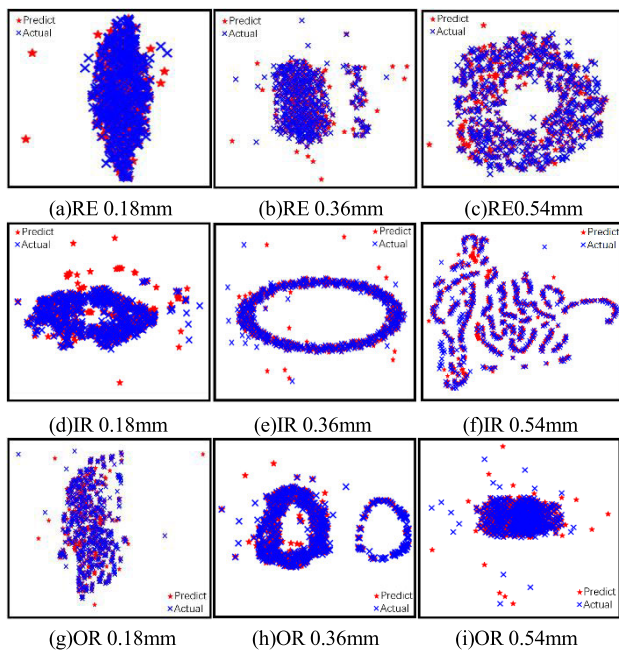**FIGURE 11.** Vibration Signal →GAPF.

FIGURE 12. Schematic diagram of difference visualization.

are shown in Fig. 17. As shown in Fig. 17, when the image first enters the network, because feature is not yet extracted, the distribution of the 10 fault types remains disordered, and various fault types cannot be clustered. When the image reaches the residual layer of the network, the network learns certain fault characteristics, and various fault types can be roughly clustered, but there is still confusion at the fault boundary. When the image reaches the final output layer, the network has fully learned the fault features, and the effect of fault classification is also significantly improved.
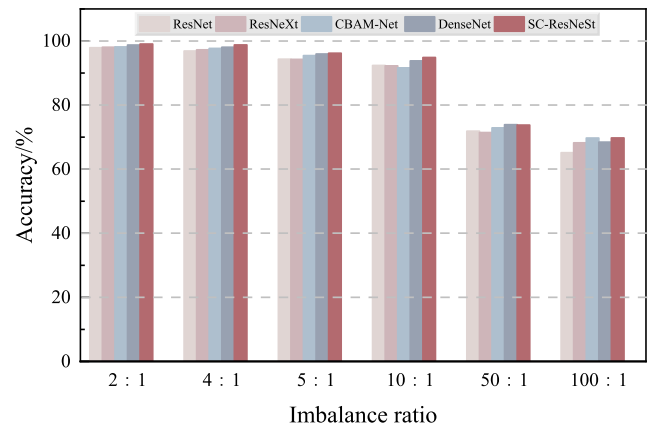
FIGURE 13. Effect of imbalance rate of samples on different fault diagnosis models.

2) MODEL NOISE IMMUNITY VERIFICATION

In an actual engineering environment, the fault diagnosis model for rotating machinery not only faces the challenge of an imbalance between the normal state samples and the fault state samples, but also the fact that mutual vibration and friction between various parts of the machine will inevitably produce noise, resulting in the vibration signal collected by the sensor easily becoming polluted by noise, which obscures the fault information in the vibration signal. Therefore, the interference of strong noise on vibration signal acquisition and the existence of imbalance in rotating machinery data samples always coexist. To verify the noise immunity of the fault diagnosis model in this paper, in this section, the bearing vibration signals under the conditions of unbalance ratio of 4:1, 10:1 and 100:1 are selected, and the pre-processing for faulty samples is the same as above.

The model anti-noise test experiment was specifically designed as follows: the samples used for model training remained unchanged, and white Gaussian noise with signal-to-noise ratios (SNR) of -9, -6, -3, and 0 dB was added to the test samples to simulate the noise in an actual engineering environment. Additionally, a comparison with other common deep learning models (such as ResNet, ResNeXt, CBAM-Net, and DenseNet) was added. Fig. 18 shows the experimental results. It can be seen that with the continuous improvement of SNR, the fault identification accuracy of various network models is gradually improved. Under an imbalanced ratio of 4:1, the fault identification accurate rate intervals of the five models in the four SNR are, respectively, [9707%, 97.87%], [97.21%, 9812%] [97.74%, 98.63%] [9807%, 98.73%] and [98.82%, 99.41%]. Under an imbalanced ratio of 10:1, the fault identification accurate rate intervals of the five models in the four SNR are, respectively, [96.87%, 97.32%] [97.04%, 97.56%] [97.18%, 97.86%] [97.47%, 98.23%] and [98.21%, 99.06%]. Under an imbalanced ratio of 100:1, the fault identification accurate rate intervals of the five models in the four SNR are, respectively, [9407%, 9537%], [94.05%, 95.86%] [9474%, 96.18%] [9588%, 9707%] and [97.64%, 9887%]. This result
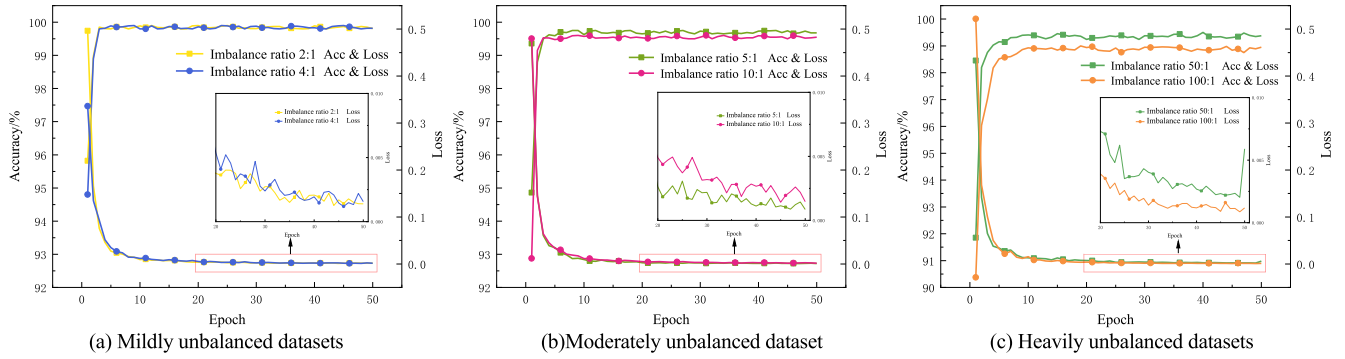
FIGURE 14. Six kinds of unbalanced proportion fault identification accuracy curves.
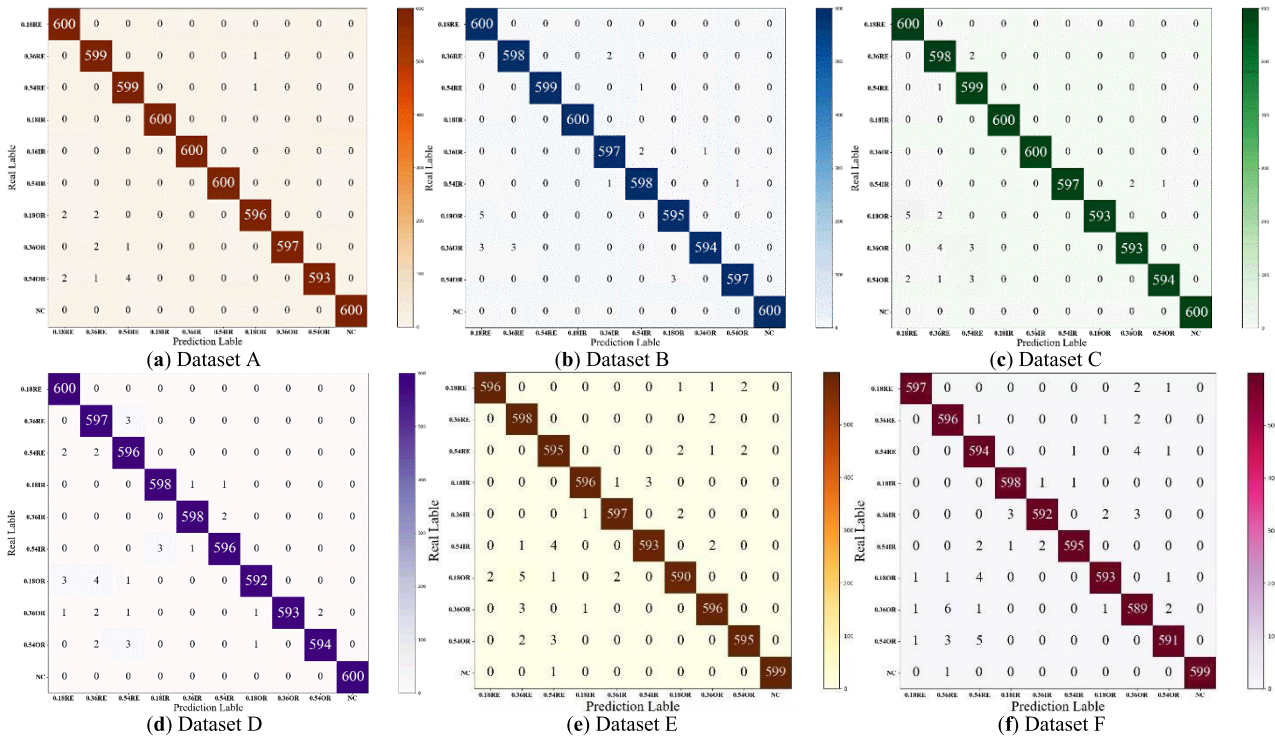

FIGURE 15. Confusion matrix of six unbalanced proportions.

occurs because, compared to the other three models, ResNet and ResNeXt have simpler structures and lack a corresponding attention mechanism, which makes it difficult for the network to extract the corresponding fault features in a strong noise environment. DenseNet builds dense connection layer that enhances the transfer of features from layer to layer CBAM-Net uses a space-channel attention mechanism based on ResNet, but unlike SC-ResNeSt, both lack a self-calibration convolution module, which makes the network unable to adaptively extract fault features according to actual needs during training.

3) COMPARED WITH OTHER IMAGE CODING METHODS

Francis et al. [53] proposed using STFT to transform the vibration signal of bearings into time-frequency images (TF image) and use the TF image as the input of the network

model to identify the faults. Xiao et al. [54] directly converted the original vibration signal of bearings into a gray image, that was used as the input of the network for fault diagnosis. To further demonstrate the superiority of this method, in this section, the following image coding methods are compared with our proposed Gram angle product field (GAPF): 1) using Short-Time Fourier Transform to convert vibration signals into a time-frequency diagram as the input of SC-ResNeSt; 2) using Xiao's method to directly convert the vibration signals into a gray image as the input of SC-ResNeSt; 3): using GADF coding to convert vibration signals into the Gram angle difference field as the input of SC-ResNeSt; 4): using GASF coding to convert the vibration signals into the Gram angle sum field as the input of SC-ResNeSt. In this section, the bearing vibration signals under the conditions of unbalance ratio of 4:1, 10:1 and 100:1 are selected, and the

TABLE 3. Average fault ident Average fault identification accuracy of different input modes.

Input Mode	GAPF	GADF	GASF	Time-frequency image	Gray scale image
Dataset B	99.64%	99.24%	99.31%	97.87%	96.93%
Dataset D	99.4%	98.25%	97.89%	96.7%	95.88%
Dataset F	99.07%	97.14%	97.11%	95.78%	94.23%
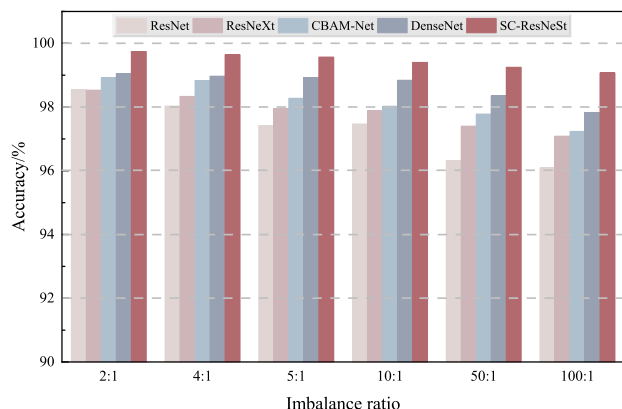
FIGURE 16. Accuracy of fault identification of different network models under six unbalanced proportions.

(a) Input Layer   (b) 1-st Residual Block

(c) 3-rd Residual Block   (d) Output Layer
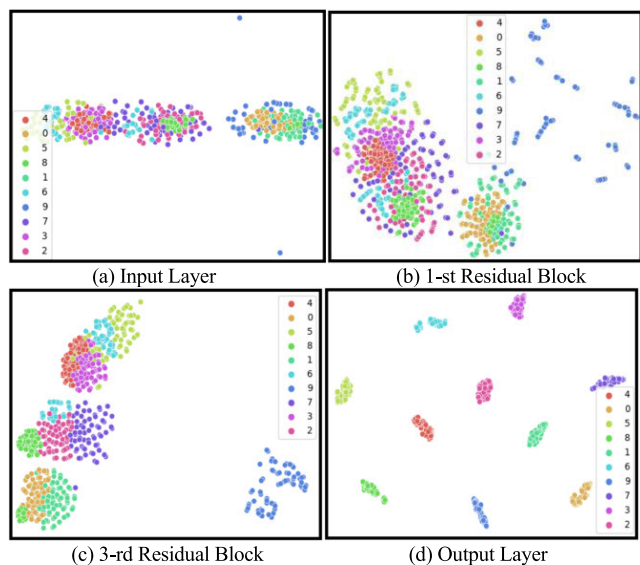
FIGURE 17. SC-ResNeSt model feature extraction visualization diagram.

predicted expansion of the fault samples as well as the pre-processing method is consistent with the above. According to a ratio of 7:3, the data are divided into a training set and test set, and the resolution of the generated image size is set to 256×256. One sample of each of the above four image encoding methods is taken for display, as shown in Fig 19 The above five images are input into the network model, in which hyperparameters such as the BatchSize and Epoch of the model remain unchanged. The experimental results are shown

in Table 4. The accuracy of fault identification in the table is the average of 50 iterations. It can be seen from the table that when the imbalance ratio is 4:1, the fault identification accuracy of the five input modes is 99.64%, 9924%, 9931%, 9787%, and 96.93%. When the imbalance ratio is 10:1, the fault identification accuracy of the five input modes is 99.4%, 98.25%, 97.89%, 96.7%, and 95.88%. When the imbalance ratio is 10:1, the fault identification accuracy of the five input modes is 9907%, 9714%, 97.11%, 95.78%, and 9423%. This occurs because the Gram angle field and vibration signal of the bearing are mutually mapped, which can reflect all the information of the time series, and will not lose some information due to Fourier transform like in spectrum analysis. When GADF and GASF are used as network inputs, their fault classification effect is equivalent because GAPF combines the characteristics of GADF and GASF, and its fault identification accuracy rate is improved by [0.33%-0.4%] [1.15%-1.51%], and [1.93%-1.96%], respectively, under the three imbalanced proportion conditions.

4) VERIFICATION EXPERIMENT OF PLANETARY GEARBOX DATASET

To further verify the practicality of the fault diagnosis model for rotating machinery in an actual engineering environment, the network model of this paper is further verified via the planetary gearbox dataset, and the fault data acquisition test bench is shown in Fig. 20. The gearbox is a two-stage transmission. The first stage is a fixed-axis gearbox, and the number of teeth is 77 and 55, respectively. The second stage is a planetary gearbox, and the number of teeth of the gear ring, planetary gear, and sun gear is 72, 27, and 18, respectively. In this dataset, the vibration signals of gears at five rotational speeds are collected at seven measuring points. Under each rotational speed, the gears have three states: normal condition (NC), broken condition (BK), and cracked condition (CK). The sampling frequency is 51.2Khz. In this section, the vibration signal of the gearbox at the measuring point 1 and the rotating speed of 1260 r/min is selected, and the fault data samples are still predicted and expanded by the DA-RNN model according to the imbalance ratios of 4:1 10:1, and 100:1. MAE and RMSE are still used as indicators to measure the prediction accuracy of DA-RNN model for the above three unbalanced datasets, as shown in Table 4.

Table 5 presents the experimental dataset, and Fig. 21 shows the fault diagnosis results under the three imbalanced ratios. It can be observed from the figure that, in the three
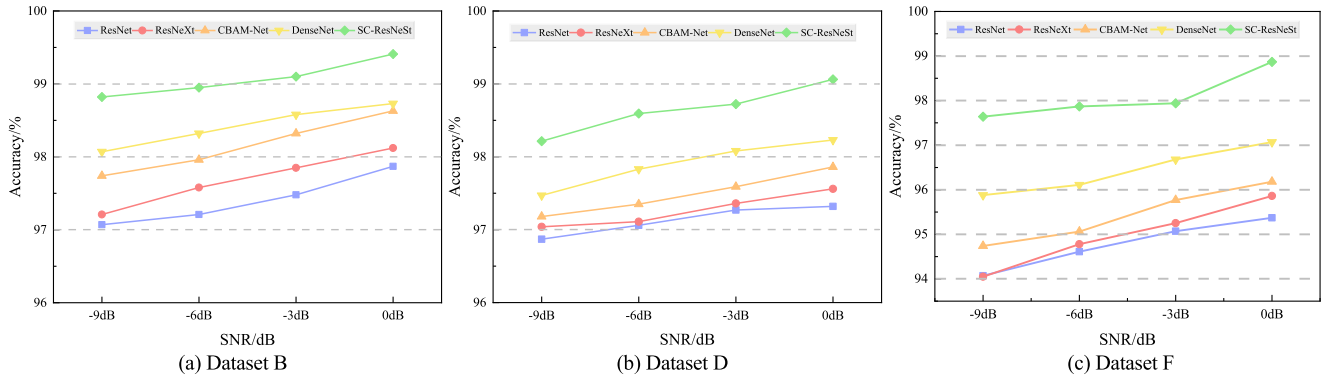
(a) Dataset B (b) Dataset D (c) Dataset F

FIGURE 18. Fault identification accuracy of different network models under different signal-to-noise ratios.

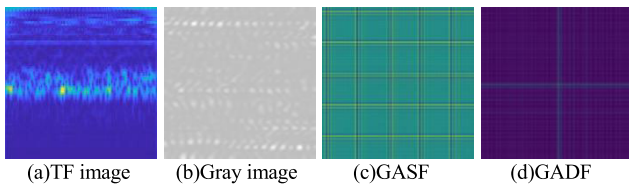(a)TF image (b)Gray image (c)GASF (d)GADF

FIGURE 19. Show pictures of different image coding methods.

TABLE 4. Prediction Results Of Three Unbalanced Proportions By DA-RNN.

Fault type	Imbalance ratio 4:1		Imbalance ratio 10:1		Imbalance ratio 100:1	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
BK	0.095	0.194	0.122	0.238	0.217	0.405
CK	0.084	0.129	0.104	0.149	0.136	0.184

FIGURE 20. Planetary gearbox fault data test bed.

FIGURE 21. Accuracy curve and loss curve of fault identification under three scales.
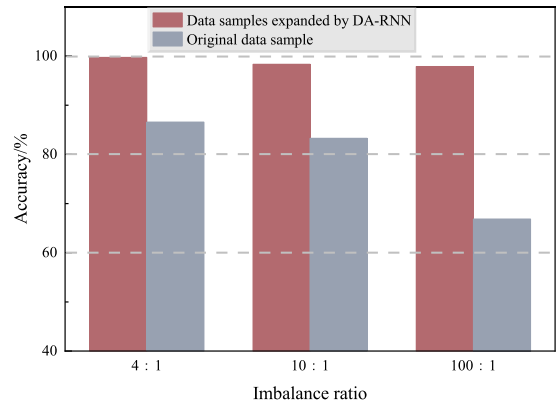

FIGURE 22. Accuracy curve and loss curve of fault identification under three scales.

scales, the network can achieve a high fault identification accuracy at the beginning of training, and the loss function curve drops steadily, indicating that there is no over-fitting occurring in the network. Finally, the average fault identification accuracy of 30 iterations in three unbalanced ratios conditions is 99.72% 98.31%, and 9787% respectively, which demonstrates the feasibility of the fault diagnosis model in an actual engineering environment.

Fig 22 shows the effect of the original data samples and the data samples expanded by DA-RNN on the classification accuracy of SC-ResNeSt. From the figure, it can be observed

that when the original data samples are used as the input of SC-ResNeSt, the recognition accuracy of the model is between [60%-85%] and the model has almost lost the ability to recognize the fault types. Next, we continue to use the confusion matrix to evaluate the detailed classification results of the model for various gear fault modes, as shown in Fig. 23. Here, when the imbalance ratio is 4:1, only a small number of CK faults will be mistaken by the model as BK faults. In the
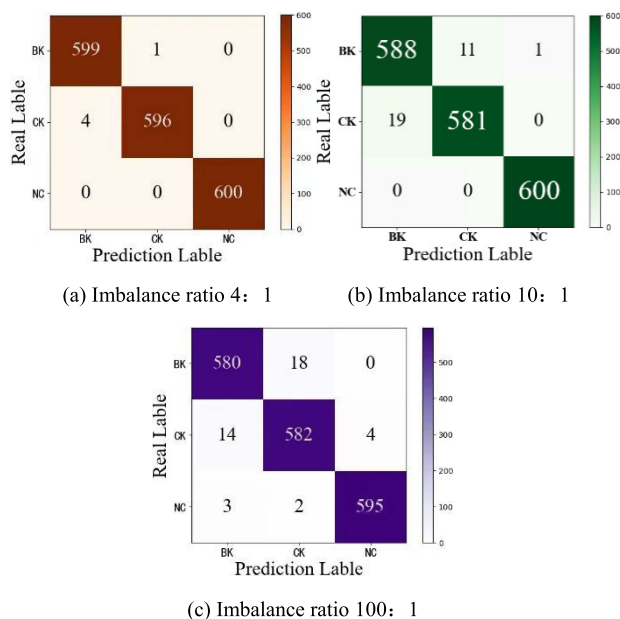
(a) Imbalance ratio 4：1   (b) Imbalance ratio 10：1

(c) Imbalance ratio 100：1

FIGURE 23. Three kinds of confusion matrices with unbalanced proportions.

TABLE 5. Experimental dataset of planetary gearbox.

Fault type		BK	CK	NC
Imbalance ratio 4	Training Samples	1400	1400	1400
：1	Testing samples	600	600	600
Imbalance ratio 10	Training Samples	1400	1400	1400
：1	Testing samples	600	600	600
Imbalance ratio	Training Samples	1400	1400	1400
100：1	Testing samples	600	600	600

case of a 10:1 imbalance ratio, BK will be slightly confused with CK. In the case of a 100:1 imbalance ratio, the confusion between CK faults and BK faults appears to be more serious.

VII. CONCLUSION

In this paper, DA-RNN and SC-ResNeSt were combined to solve the problem of the serious data imbalance faced by rotating machinery in an actual engineering environment using time-domain signal predictions. Ultimately, good diagnostic results were obtained in the experiments for the CWRU bearing dataset and planetary gearbox dataset, and the following conclusions were drawn:

(1) In this paper, MAE and RMSE were used as the evaluation indexes to determine the prediction accuracy of the new DA-RNN model, and good prediction results were obtained with the CWRU bearing dataset. In the experiments of visualizing the discrepancy between real image samples and predicted image samples using t-SNE technique, the experimental results further demonstrate that DA-RNN has high prediction accuracy.

(2) The fault diagnosis model of SC-ResNeSt used in this paper achieved good fault diagnosis results with the CWRU

bearing dataset and planetary gearbox dataset because of the split-attention mechanism and self-calibration convolution module. Compared to the current mainstream deep learning model, SC-ResNeSt was found to offer better robustness and anti-noise properties.

(3) Compared to other image coding methods such as time-frequency images and gray images, the accuracy of fault identification can be increased by 1%-5% by using the Gram Angle Product Field as the input of the network, which proves the superiority of the GAPF image coding method.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

REFERENCES

[1] I. El-Thalji and E. Jantunen, "A summary of fault modelling and predictive health monitoring of rolling element bearings," *Mech. Syst. Signal Process.*, vols. 60–61, pp. 252–272, Aug. 2015, doi: 10.1016/j.ymssp.2015.02.008.

[2] W. Cui, G. Meng, A. Wang, X. Zhang, and J. Ding, "Application of rotating machinery fault diagnosis based on deep learning," *Shock Vib.*, vol. 2021, Dec. 2021, Art. no. 3083190, doi: 10.1155/2021/3083190.

[3] W. Huang, S. Li, L. Mao, Y. Ma, J. Wang, C. Shen, and Z. Zhu, "Research on multi-source sparse optimization method and its application in compound fault detection of gearbox," *J. Mech. Eng.*, vol. 57, no. 7, pp. 87–99, Apr. 2021, doi: 10.3901/JME.2021.07.087.

[4] H. Wang and L. Pu, "Bearing fault diagnosis of split attention network based on deep subdomain adaptation," *Appl. Sci.*, vol. 12, no. 24, Dec. 2022, Art. no. 12762, doi: 10.3390/app122412762.

[5] X. Zhang, B. Zhao, and Y. Lin, "Machine learning based bearing fault diagnosis using the case western reserve university data: A review," *IEEE Access*, vol. 9, pp. 155598–155608, 2021, doi: 10.1109/ACCESS.2021.3128669.

[6] S. Yang, X. Kong, Q. Wang, Z. Li, H. Cheng, and K. Xu, "Deep multiple auto-encoder with attention mechanism network: A dynamic domain adaptation method for rotary machine fault diagnosis under different working conditions," *Knowl.-Based Syst.*, vol. 249, Aug. 2022, Art. no. 108639, doi: 10.1016/j.knosys.2022.108639.

[7] X. Dai and Z. Gao, "From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis," *IEEE Trans. Ind. Informat.*, vol. 9, no. 4, pp. 2226–2238, Nov. 2013, doi: 10.1109/TII.2013.2243743.

[8] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587, doi: 10.1016/j.ymssp.2019.106587.

[9] Y. Ge, F. Zhang, and Y. Ren, "Adaptive fault diagnosis method for rotating machinery with unknown faults under multiple working conditions," *J. Manuf. Syst.*, vol. 63, pp. 177–184, Apr. 2022, doi: 10.1016/j.jmsy.2022.03.009.

[10] M. Ye, X. Yan, and M. Jia, "Rolling bearing fault diagnosis based on VMD-MPE and PSO-SVM," *Entropy*, vol. 23, no. 6, p. 762, Jun. 2021, doi: 10.3390/e23060762.

[11] Z. He, Z. Shao, Z. Ding, H. Jiang, and J. Cheng, "Modified deep autoencoder driven by multisource parameters for fault transfer prognosis of aeroengine," *IEEE Trans. Ind. Electron.*, vol. 69, no. 1, pp. 845–855, Jan. 2022, doi: 10.1109/TIE.2021.3050382.

[12] C. Qian, J. Zhu, Y. Shen, Q. Jiang, and Q. Zhang, "Deep transfer learning in mechanical intelligent fault diagnosis: Application and challenge," *Neural Process. Lett.*, vol. 54, no. 3, pp. 2509–2531, Jan. 2022, doi: 10.1007/s11063-021-10719-z.

[13] Y. Lei, F. Jia, D. Kong, J. Lin, and S. Xing, "Opportunities and challenges of machinery intelligent fault diagnosis in big data era," *J. Mech. Eng.*, vol. 54, no. 5, pp. 94–104, Mar. 2018, doi: 10.3901/JME.2018.05.094.

[14] D. Zhu, X. Cheng, L. Yang, Y. Chen, and S. X. Yang, "Information fusion fault diagnosis method for deep-sea human occupied vehicle thruster based on deep belief network," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9414–9427, Sep. 2022, doi: 10.1109/TCYB.2021.3055770.

[15] Y. Xu, Z. Li, S. Wang, W. Li, T. Sarkodie-Gyan, and S. Feng, "A hybrid deep-learning model for fault diagnosis of rolling bearings," *Measurement*, vol. 169, Feb. 2021, Art. no. 108502, doi: 10.1016/j.measurement.2020.108502.

[16] J. Sun, J. Wen, C. Yuan, Z. Liu, and Q. Xiao, "Bearing fault diagnosis based on multiple transformation domain fusion and improved residual dense networks," *IEEE Sensors J.*, vol. 22, no. 2, pp. 1541–1551, Jan. 2022, doi: 10.1109/JSEN.2021.3131722.

[17] C. Ma, Z. Sheng, Y. Xu, and K. Zhang, "Fault diagnosis of rolling bearing based on adaptive frequency slice wavelet transform," *Trans. Chin. Soc. Agricult. Eng.*, vol. 35, no. 10, pp. 34–41, May 2019, doi: 10.11975/j.issn.1002-6819.2019.10.005.

[18] Y. Liu, W. Wen, Y. Bai, and Q. Meng, "Self-supervised feature extraction via time–frequency contrast for intelligent fault diagnosis of rotating machinery," *Measurement*, vol. 210, Mar. 2023, Art. no. 112551, doi: 10.1016/j.measurement.2023.112551.

[19] Y. Tong, X. Pang, and Z. Wei, "Fault diagnosis method of rolling bearing based on GADF-CNN," *J. Vib. Shock*, vol. 40, no. 5, pp. 247–253, Mar. 2021, doi: 10.13465/j.cnki.jvs.2021.05.032.

[20] X. Zhang, M. Zhang, and X. Li, "Rolling bearing fault mode recognition based on 2D image and CNN-BiGRU," *J. Vib. Shock*, vol. 40, no. 23, pp. 194–201, Dec. 2021, doi: 10.13465/j.cnki.jvs.2021.23.026.

[21] F. Zhou, S. Yang, H. Fujita, D. Chen, and C. Wen, "Deep learning fault diagnosis method based on global optimization GAN for unbalanced data," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104837, doi: 10.1016/j.knosys.2019.07.008.

[22] H. Sun, C. Wang, and X. Cao, "An adaptive anti-noise gear fault diagnosis method based on attention residual prototypical network under limited samples," *Appl. Soft Comput.*, vol. 125, Aug. 2022, Art. no. 109120, doi: 10.1016/j.asoc.2022.109120.

[23] X. Yu, B. Tang, and L. Deng, "Fault diagnosis of rotating machinery based on graph weighted reinforcement networks under small samples and strong noise," *Mech. Syst. Signal Process.*, vol. 186, Mar. 2023, Art. no. 109848, doi: 10.1016/j.ymssp.2022.109848.

[24] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New imbalanced bearing fault diagnosis method based on sample-characteristic oversampling technique (SCOTE) and multi-class LS-SVM," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107043, doi: 10.1016/j.asoc.2020.107043.

[25] C. Shen, H. Zhang, S. Meng, and C. Li, "Augmented data driven self-attention deep learning method for imbalanced fault diagnosis of the HVAC chiller," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105540, doi: 10.1016/j.engappai.2022.105540.

[26] Y. Zhang, X. Li, L. Gao, L. Wang, and L. Wen, "Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning," *J. Manuf. Syst.*, vol. 48, pp. 34–50, Jul. 2018, doi: 10.1016/j.jmsy.2018.04.005.

[27] Y. Wu, J. Wang, L. Yang, and M. Yu, "Survey on cost-sensitive deep learning methods," *Comput. Sci.*, vol. 46, no. 5, pp. 1–12, May 2019, doi: 10.11896/j.issn.1002-137X.2019.05.001.

[28] P. Peng, W. Zhang, Y. Zhang, H. Wang, and H. Zhang, "Imbalanced fault diagnosis based on particle swarm optimization and sparse auto-encoder," in *Proc. IEEE 24th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, Dalian, China, May 2021, pp. 210–213, doi: 10.1109/CSCWD49262.2021.9437742.

[29] J. He, L. Yin, J. Liu, C. Zhang, and H. Yang, "A fault diagnosis method for unbalanced data based on a deep cost sensitive convolutional neural network," *IFAC-PapersOnLine*, vol. 55, no. 3, pp. 43–48, 2022, doi: 10.1016/j.ifacol.2022.05.008.

[30] Z. Wu, H. Zhang, J. Guo, Y. Ji, and M. Pecht, "Imbalanced bearing fault diagnosis under variant working conditions using cost-sensitive deep domain adaptation network," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116459, doi: 10.1016/j.eswa.2021.116459.

[31] Y. Lei, B. Yang, Z. Du, and N. Lv, "Deep transfer diagnosis method for machinery in big data era," *J. Mech. Eng.*, vol. 55, no. 7, pp. 1–8, May 2020, doi: 10.3901/JME.2019.07.001.

[32] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2019, doi: 10.1109/TIE.2018.2877090.

[33] Y. Ding, M. Jia, J. Zhuang, Y. Cao, X. Zhao, and C.-G. Lee, "Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions," *Rel. Eng. Syst. Saf.*, vol. 230, Feb. 2023, Art. no. 108890, doi: 10.1016/j.ress.2022.108890.

[34] X. Tao, S. Hao, D. Zhang, and P. Xu, "Overview of classification algorithms for unbalanced data," *J. Chongqing Univ. Posts Telecommun.*, vol. 25, no. 1, pp. 101–110+121, Feb. 2013, doi: 10.3979/j.issn.1673-825X.2013.01.017.

[35] M. Hakim, A. A. B. Omran, A. N. Ahmed, M. Al-Waily, and A. Abdellatif, "A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations," *Ain Shams Eng. J.*, vol. 14, no. 4, Apr. 2023, Art. no. 101945, doi: 10.1016/j.asej.2022.101945.

[36] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," 2017, *arXiv:1704.02971*.

[37] J. Li, Y. Liu, and Q. Li, "Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method," *Measurement*, vol. 189, Feb. 2022, Art. no. 110500, doi: 10.1016/j.measurement.2021.110500.

[38] Y. Liu, C. Gong, L. Yang, and Y. Chen, "DSTP-RNN: A dual-stage two-phase attention-based recurrent neural network for long-term and multi-variate time series prediction," *Expert Syst. Appl.*, vol. 143, Apr. 2020, Art. no. 113082, doi: 10.1016/j.eswa.2019.113082.

[39] Y. Cheng and Y. Morimoto, "Triple-stage attention-based multiple parallel connection hybrid neural network model for conditional time series forecasting," *IEEE Access*, vol. 9, pp. 29165–29179, 2021, doi: 10.1109/ACCESS.2021.3059861.

[40] H. Y. Pan, Y. Yang, X. Li, J. D. Zheng, and J. S. Cheng, "Symplectic geometry mode decomposition and its application to rotating machinery compound fault diagnosis," *Mech. Syst. Signal Process.*, vol. 114, pp. 189–211, Jan. 2019, doi: 10.1016/j.ymssp.2018.05.019.

[41] Y. Lv, W. Zhao, Z. Zhao, W. Li, and K. K. H. Ng, "Vibration signal-based early fault prognosis: Status quo and applications," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101609, doi: 10.1016/j.aei.2022.101609.

[42] W. Li, Q. Shi, K. Wang, and X. Cheng, "Runoff prediction based on variational mode decomposition and deep gated network," *J. Hydroelectr. Eng.*, vol. 39, no. 3, pp. 34–44, Mar. 2020, doi: 10.11660/slfdxb.20200304.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[44] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[45] G. Yuan, J. Liu, H. Liu, R. Qu, and H. Zhou, "Classification of cigarette appearance defects based on ResNeSt," *J. Yunnan Univ., Natural Sci. Ed.*, vol. 44, no. 3, pp. 464–470, Dec. 2022, doi: 10.7540/j.ynu.20210257.

[46] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[47] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10093–10102, doi: 10.1109/CVPR42600.2020.01011.

[48] Z. Jiang, J. Zheng, H. Pan, and Z. Pan, "Rolling bearing fault diagnosis method based on multiscale time irreversibility and t-SNE manifold learning," *J. Vib. Shock*, vol. 36, no. 17, pp. 61–68 and 84, Dec. 2016, doi: 10.13465/j.cnki.jvs.2017.17.010.

[49] G. Pleiss, D. Chen, G. Huang, T. Li, L. van der Maaten, and K. Q. Weinberger, "Memory-efficient implementation of DenseNets," 2017, *arXiv:1707.06990*.

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

[51] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016, *arXiv:1611.05431*.

[52] H. Yang, X. Li, and W. Zhang, "Interpretability of deep convolutional neural networks on rolling bearing fault diagnosis," *Meas. Sci. Technol.*, vol. 33, no. 5, Feb. 2022, Art. no. 055005, doi: 10.1088/1361-6501/ac41a5.

[53] R. F. R. Junior, I. A. D. S. Areias, M. M. Campos, C. E. Teixeira, L. E. B. da Silva, and G. F. Gomes, "Fault detection and diagnosis in electric motors using convolution neural network and short-time Fourier transform," *J. Vib. Eng. Technol.*, vol. 10, no. 7, pp. 2531–2542, Oct. 2022, doi: 10.1007/s42417-022-00501-3.

[54] X. Xiao, J. Wang, Y. Zhang, Q. Guo, and S. Zong, "Two-dimensional convolution neural network optimization method for bearing fault diagnosis," *Proc. CSEE*, vol. 39, no. 15, pp. 4558–4568, Aug. 2019, doi: 10.13334/j.0258-8013.pcsee.182037.

JIE YANG received the bachelor's degree in mechanical engineering from Hunan Agricultural University, in 2021. He is currently pursuing the master's degree in mechanical and electrical engineering with the Xi'an University of Architecture and Technology. His current research interests include deep learning, fault diagnosis, and life prediction.

HAITAO WANG received the master's degree in electrical theory and new technology and the Ph.D. degree in electronic science and technology from Northwest Polytechnical University, China, in 2009 and 2013, respectively. He is currently an Associate Professor with the Xi'an University of Architecture and Technology, China, where he is also the Director of the Electromechanical System Detection and Control Research Institute. His current research interests include deep learning, fault diagnosis and prediction, and reliability analysis.

XIHENG ZHANG received the bachelor's degree in mechanical engineering from the Xi'an University of Architecture and Technology, in 2021, where he is currently pursuing the master's degree in mechanical and electrical engineering. His current research interests include information fusion, the fault diagnosis of rotating machinery, intelligent equipment operation, and flexible sensor preparation.

YIFAN GUO received the bachelor's degree in mechanical engineering from the Xi'an University of Architecture and Technology, in 2021, where he is currently pursuing the master's degree in mechanical and electrical engineering. His current research interests include information fusion, the fault diagnosis of rotating machinery, and intelligent equipment operation and maintenance based on big data driving.

XIANG LIU received the bachelor's degree in mechanical engineering from the Baoji University of Arts and Sciences, in 2020. She is currently pursuing the master's degree in mechanical and electrical engineering with the Xi'an University of Architecture and Technology. Her current research interests include information fusion, the fault diagnosis of rotating machinery, intelligent equipment operation, and nonlinear system identification.

LICHEN SHI received the master's degree in mechanical engineering and the Ph.D. degree in electronic science and technology from the Xi'an University of Architecture and Technology, China, in 1998 and 2009, respectively. She is a currently a Professor with the Xi'an University of Architecture and Technology. Her current research interests include deep learning, mechanical system dynamics, equipment condition monitoring, and fault diagnosis. She is the Executive Director of the Shaanxi Vibration Engineering Society.

• • •