ELECTRIC LOCOMOTIVE BEARING FAULT DIAGNOSIS BASED ON THE SELF-ATTENTION-BASED DEEP NETWORK

Qian TAO[1,*], Junxian ZHANG[1], Yang ZHOU[1], Qian ZHANG[2]

Accurate fault diagnosis of electric locomotive bearings is critical for railway safety, yet traditional methods relying on signal decomposition and manual feature extraction struggle with noise and complex fault types. To overcome these challenges, this study proposes a self-attention deep network (SADN) framework that combines spatiotemporal feature learning with an incremental stochastic configuration classifier. The dual-branch architecture includes: (1) a temporal branch with a self-attention Long Short-Term Memory (LSTM) enhanced by timestamp attention gating (TSAG) and parametric soft-threshold shrinkage (PSTS) for dynamic noise suppression and global feature modeling; and (2) a spatial branch with a channel attention recalibration residual module (CARRM) and adaptive threshold truncation for improved local feature discrimination. A hierarchical fusion strategy merges temporal tensor unfolding and spatial multi-scale pooling to retain multi-granularity features and boost class separability. To address limited generalization in fixed-depth networks and Softmax classifiers, a dynamic depth extension (DDE) based on error entropy adjusts the network topology, while stochastic configuration networks (SCNs) incrementally generate hidden nodes for strong approximation under constraints. Experiments show SADN achieves 84.1% accuracy at SNR = 0 dB, 98.6% at 10 dB, and 99.7% at 20 dB, outperforming 1D-CNN and LSTM across SNRs.

Keywords: Self-attention deep networks, Electric locomotive bearings, Dynamic depth extension, Stochastic configuration networks

1. Introduction

Since the Industrial Revolution, mechanical equipment has been vital to industrial development. In modern railway systems, electric locomotives are core power units, whose stability directly affects operational efficiency and safety. Among key components, bearings play a crucial role in supporting rotating shafts, reducing friction, and transmitting loads. However, long-term high-speed operation under complex loads exposes bearings to fatigue, lubrication failure, temperature fluctuations, and external impacts, leading to wear, spalling, and cracking. These faults increase vibration and energy consumption, potentially resulting in severe

[1] CRRC NANJING PUZHEN CO., LTD., NanJing, China, 010700090506@crrcgc.cc
[2] School of Electrical and Automation Engineering, Hefei University of Technology, Hefei, China

equipment damage and safety risks. Therefore, accurate and timely bearing fault diagnosis is essential. Early detection prevents fault escalation, ensures reliable operation, reduces maintenance costs, extends service life, and enhances the safety and stability of railway networks—ultimately supporting broader economic and social development.

Conventional fault diagnosis methods for electric locomotive bearings often rely on non-stationary signal analysis techniques like Empirical Mode Decomposition (EMD) and Variational Mode Decomposition (VMD). In 2016, a VMD-based approach was proposed, which decomposed vibration signals into intrinsic mode functions and extracted features using multi-scale fractal dimensions and energy metrics. An optimized Support Vector Machine (SVM) classifier achieved a diagnostic accuracy of up to 99.75% [1]. However, the performance of VMD heavily depends on the selection of mode numbers and penalty factors, requiring parameter tuning. To address this, the Grey Wolf Optimization (GWO) algorithm was employed to adaptively optimize VMD parameters and the SVM's kernel function, significantly improving fault diagnosis accuracy [2]. Furthermore, an intelligent hybrid diagnosis approach combining Parameter Optimized Variational Mode Decomposition with Weighted Compound Kurtosis (POA-VMD-WCK) and Gramian Angular Difference Field combined with a Shifted Window Transformer (GADF-Swin Transformer) was introduced. This methodology effectively addresses intrinsic challenges associated with feature extraction and consequently enhances fault recognition accuracy in the context of bearing diagnosis [3]. A multi-stage fault diagnosis strategy for rolling bearings, utilizing Ensemble Empirical Mode Decomposition (EEMD) for feature extraction, employs a Belief-Rule-Base (BRB)-optimized Particle Swarm Optimization (PSO) to precisely adjust model parameters, achieving improved accuracy and computational efficiency [4]. Additionally, utilized a fuzzy expert system for motor bearing fault diagnosis, incorporating a similarity separation method to automatically derive fuzzy rules from numerical data [5]. To enhance noise robustness, an improved interval overlap method was employed to select input feature vectors based on a predefined validity metric. However, traditional methods still face key limitations. Non-stationary signal decomposition can be unstable when handling complex multi-fault signals, causing feature loss or overlap. Additionally, reliance on manually selected features introduces human bias and reduces noise robustness. As a result, researchers are increasingly turning to self-attention fused deep networks to overcome these challenges and improve fault diagnosis performance.

Deep learning has been widely applied to intelligent bearing fault diagnosis, addressing the limitations of traditional methods. By automatically extracting meaningful features from large volumes of raw signal data, these techniques significantly improve diagnostic accuracy and computational efficiency. For

example, an improved selective ensemble deep learning approach has demonstrated notable performance gains in rolling bearing fault identification [6]. In 2017, a Convolutional Deep Belief Network (CDBN) was applied to diagnose faults in electric locomotive bearings, showing strong performance even under challenging operational conditions [7]. Moreover, a Gaussian-guided adversarial adaptive transfer network was developed to diagnose rolling bearing faults, showing excellent performance in managing cross-condition situations [8]. Similarly, an Enhanced Deep Autoencoder (EDAE) was utilized for fault diagnosis on multiple devices, significantly boosting the model's generalization capability across different equipment [9]. Additionally, the integration of an optimized deep sparse autoencoder with a Gated Recurrent Unit (GRU) model has been demonstrated to enhance the precision of fault classification [10].

Modern modeling systems increasingly integrate Neural Networks with compatible parallel architectures to enhance processing capacity and predictive power. This approach effectively addresses the limitations of rigid Fault Tree analysis, enabling more comprehensive and adaptive risk and probability assessments [11]. Attention mechanisms enhance deep learning models by focusing on critical information, improving both efficiency and accuracy. Recently, they have gained traction in fault diagnosis. For example, a bidirectional LSTM with attention has been applied to rolling bearing fault diagnosis, effectively capturing key temporal features [12]. Similarly, the attention mechanism in deep learning models was enhanced to better emphasize critical features, leading to improved diagnostic accuracy [13]. Overall, deep learning has shown strong potential in bearing fault diagnosis, offering significant improvements in accuracy, adaptability to complex operating conditions, and cross-device transfer capability.

The attention mechanism emulates the human visual perception process, where limited attention capacity necessitates prioritization of key information while disregarding irrelevant details, thereby enhancing task efficiency. Attention mechanisms can be categorized into hard and soft types, depending on the way features are focused on during processing [14]. Hard attention assigns binary values (0 or 1) to features, focusing strictly on selected elements and ignoring others, resulting in a "black-and-white" effect. In contrast, soft attention assigns continuous weights between 0 and 1, enabling more flexible and comprehensive global information processing.

The residual attention network incorporates the soft attention mechanism to enhance the traditional residual network structure [15]. The network comprises stacked residual attention modules, each with two branches. The right branch extracts feature via conventional convolutions, while the left applies soft attention for feature reweighting. After aligning the attention scale with the right branch's output, the two are multiplied. To prevent excessive feature map reduction, a

shortcut connection—similar to Residual Network (ResNet)—is added between the branches. This design enhances performance and eases training in deep networks.

Self-attention, a form of soft attention, models relationships between different positions within input data without external guidance, enabling effective latent feature extraction. Its strong performance has made it a research focus in neural networks. In 2019, the Global Context Network (GCNet) was introduced to integrate self-attention mechanisms into convolutional networks [16]. GCNet captures global context from convolutional feature maps by computing weights and performing weighted averaging to produce the final output. Compared to traditional convolutional networks that focus mainly on local features, GCNet enhances global information modeling capabilities.

The self-attention fusion deep network is a multi-modal fault diagnosis framework tailored for rolling bearings under complex industrial conditions. It features a dual-branch heterogeneous architecture, combining an enhanced self-attention LSTM for global temporal feature modeling and a self-attention ResNet for local spatial feature extraction. To improve robustness under strong noise, a dynamic depth adjustment mechanism and an incremental classifier optimization strategy are incorporated. The main contributions of this study are as follows:

To mitigate non-stationary noise and feature degradation in fault signals, this study proposes a spatiotemporal dual-path self-attention architecture. The temporal branch integrates time-scale attention gating (TSAG) and parameterized soft-threshold shrinkage (PSTS) to dynamically suppress noise and extract fault impact features. The spatial branch employs a channel attention recalibration residual module (CARRM) with adaptive threshold truncation to enhance the SNR of local features.

A hierarchical multi-modal fusion strategy combines temporal tensor flattening and spatial multi-scale pooling to construct a joint spatiotemporal feature space, preserving temporal hierarchies while improving spatial class separability.

To enhance adaptability under complex conditions, a dynamic depth expansion (DDE) mechanism based on semantic confidence and error entropy is introduced. Combined with an incremental stochastic configuration classifier, the model adaptively adjusts its topology to maintain generalization under parameter constraints, addressing the limitations of fixed-structure networks.

2. The proposed method

Traditional LSTM and residual networks are sensitive to noise when extracting temporal and spatial features from vibration signals, limiting diagnostic accuracy. To improve noise robustness, we integrate Global Context Network (GCNet) and soft-thresholding, then fuse their outputs for better fault identification. The proposed self-attention deep fusion model (Figure 1) includes an adaptive

hierarchical self-attention LSTM, an adaptive hierarchical self-attention residual network, a fused feature vector, and a Stochastic Configuration Networks (SCNs) classifier.

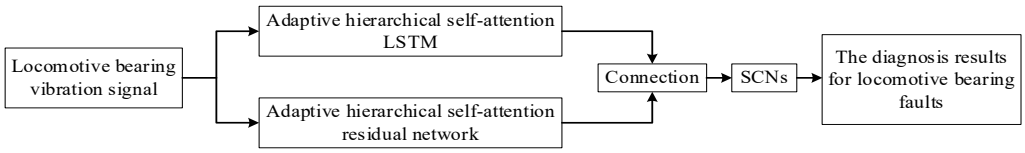


Fig. 1. Model structure diagram based on self attention fusion deep network

2.1 GCNet and Soft-Thresholding Strategy

The self-attention mechanism selectively captures important features from input data by learning internal dependencies, without relying on external information, while suppressing irrelevant data. GCNet is a typical self-attention network that extracts features by leveraging such internal information [17]. Its structure consists of three key steps: (1) Global Attention Pooling – applying 1×1 convolutions and a softmax activation to compute attention weights, followed by pooling to obtain a global context feature of size C×1×1; (2) Feature Transformation – using two 1×1 convolutions and layer normalization to model inter-channel dependencies; and (3) Feature Aggregation – applying a reweighting operation to integrate global context features into each channel of the feature map. By learning global contextual information across all channels, GCNet assigns varying attention weights, enhancing informative features while suppressing less relevant ones. Figure 2 depicts the fundamental architecture of GCNet.

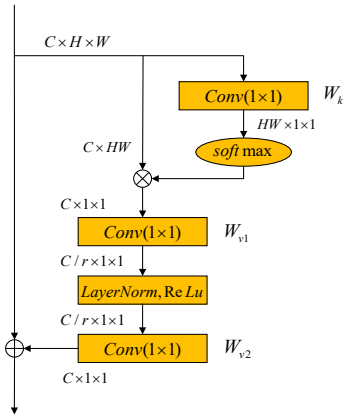


Fig. 2. Basic structure of GCNet

In digital signal processing, soft-thresholding is a fundamental denoising technique for filtering noise-contaminated signal [18]. It typically requires a pre-designed filter to transform the signal into positive or negative values, with noise

components pushed toward zero. A threshold is then applied to suppress signal components whose absolute values fall below this threshold. However, the design of effective filters and appropriate thresholds demands extensive experimentation and expert knowledge, limiting the practical applicability of this method. Neural networks, by contrast, serve as natural self-learning filters. Integrating soft-thresholding with deep learning enables adaptive noise suppression and the extraction of features with strong discriminative ability. The soft-thresholding function is given by equation (1):

$$y = \begin{cases} x - \tau & x > \tau \\ 0 & -\tau \le x < \tau \\ x + \tau & x \le -\tau \end{cases} \tag{1}$$

where the input is denoted as x, the output as y, and τ is a positive-valued threshold vector. The soft-thresholding function sets feature values within the threshold range to zero after activation, thereby preserving salient features while effectively suppressing noise. The derivative of this function is given by equation (2):

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & x > \tau \\ 0 & -\tau \le x < \tau \\ 1 & x \le -\tau \end{cases} \tag{2}$$

In this study, a soft-thresholding strategy and self-attention mechanism are integrated to enhance traditional LSTM networks and residual modules. Based on this, a hierarchically adaptive self-attention LSTM network along with a hierarchically adaptive self-attention residual network have been designed to achieve reliable fault diagnosis of electric locomotive bearings in noisy environments.

2.2 Adaptive hierarchical levels

Due to the complexity of fault state identification in electric locomotive bearings, fixed-depth self-attention fusion networks often struggle to generalize across diverse vibration time-series data. To enhance feature representation and recognition accuracy, the network depth should be adaptively adjusted. To this end, we introduce an intelligent cognition mechanism based on semantic error entropy [19], which defines a confidence metric to guide both feature space optimization and network structure adaptation. By constructing a multi-level, differentiated feature space tailored to vibration data, the proposed method significantly improves diagnostic accuracy.

Denote the training time-series dataset for electric locomotive bearings as $U = \{U_1, U_2, \dots, U_p\}$,where p represents the categories of bearing conditions, including normal and faulty states. Build a q-layer fusion deep network M_q that incorporates self-attention. For the $i-th$ category fault training sample set U_i (where $i \in [1, p]$), take the j-th sample (where $j \in [1, n]$) from U_i and obtain its

fused feature vector throughM_q,denoted as $\boldsymbol{Z}_j^i = \left[z_{j_1}^i, z_{j_2}^i, \ldots, z_{j_k}^i\right]$, here, k denotes the dimension of the fused feature vector. Using latent semantic analysis, $\boldsymbol{Z}_j^i$ is mapped to the fused latent semantic feature vector $\widetilde{\boldsymbol{Z}}_j^i = \left[\tilde{\boldsymbol{z}}_{j1}^i, \tilde{\boldsymbol{z}}_{j2}^i, \ldots, \tilde{\boldsymbol{z}}_{jk}^i\right]$. Similarly, the fused latent semantic feature matrix $\widetilde{\boldsymbol{Z}}^i$ for U_i is formed as presented in equation (3):

$$\widetilde{\boldsymbol{Z}}^i = \left[\widetilde{\boldsymbol{Z}}_1^i, \widetilde{\boldsymbol{Z}}_2^i, \ldots, \widetilde{\boldsymbol{Z}}_n^i\right]^T = \begin{bmatrix} \tilde{z}_{11}^i & \tilde{z}_{12}^i & \cdots & \tilde{z}_{1k}^i \\ \tilde{z}_{21}^i & \tilde{z}_{22}^i & \cdots & \tilde{z}_{2k}^i \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{z}_{n1}^i & \tilde{z}_{n2}^i & \cdots & \tilde{z}_{nk}^i \end{bmatrix} \tag{3}$$

For any sample X in U, its fused latent semantic feature vector is denoted by $\tilde{\boldsymbol{c}} = [\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_k]$. Accordingly, the fused latent semantic error entropy E_i between X and U_i is given by equation (4):

$$E_i = \frac{\sum_{j=1}^{n} \sqrt{\left(\widetilde{\boldsymbol{C}} - \widetilde{\boldsymbol{Z}}_j^i\right) S^{-1} \left(\widetilde{\boldsymbol{C}} - \widetilde{\boldsymbol{Z}}_j^i\right)^T}}{n} \tag{4}$$

Here, S represents the covariance matrix of $[\widetilde{\boldsymbol{C}}; \widetilde{\boldsymbol{Z}}^i]^T$. The fused latent semantic error entropy E between X and all training samples across bearing condition categories is defined in equation (5):

$$E = -\frac{E_i}{\sum_{i=1}^{p} E_i} ln \frac{E_i}{\sum_{i=1}^{p} E_i} \tag{5}$$

A higher value of E signifies greater fused latent semantic error entropy, indicates a lower confidence level in the fault diagnosis result of X under the existing network model. Training samples resulting in a diagnosis confidence level below the established threshold are isolated. To enhance feature representation and mitigate the risk of overfitting to sparse or overly uniform data, an adaptive noise injection (ANI) strategy is applied to these isolated samples for data augmentation. This strategic introduction of noise achieves two critical objectives: it improves model robustness by simulating real-world sensor variability and environmental noise, thus compelling the network to learn essential, noise-resilient fault characteristics, and it enhances generalization by expanding the feature space diversity, which effectively prevents model collapse and ensures sustained high diagnostic accuracy when exposed to unseen vibration signals. Subsequently, a deeper network is built to extract finer details for re-diagnosis. Meanwhile, for training samples meeting the threshold, their corresponding network models are stored in the model repository.

The introduced confidence evaluation metric serves as feedback regulation for the adaptive-layer self-attention fusion feature network. As the m-th model is constructed, the transformation of network depth q is given by equation (6):

$$q(m) = \begin{cases} q_0 & m = 1 \\ m + q_0 - 1 & 1 < m < q_{\max} \\ q_{\max} & m \geq q_{\max} \end{cases} \tag{6}$$

Here, q_0 denotes the initial network depth at model initialization, while q_{max} indicates the maximum network depth achievable after adaptive adjustment. When the system's feedback mechanism triggers a change in network depth, an equal number of layers are added simultaneously to both the self-attention LSTM network and the adaptive self-attention Resnet.

2.3 Adaptive hierarchical self-attention LSTM network

Electric locomotive bearing vibration signals are often contaminated by noise in practical scenarios. Directly inputting these signals into the LSTM network may degrade fault feature extraction and diagnostic performance. To tackle this, the conventional LSTM network is enhanced by integrating a self-attention mechanism, creating a self-attention LSTM network with adaptive layers that mimics human cognitive processes [20].

We borrow from GCNet and introduce a global attention pooling module, shown in the dashed box in Figure 3, to link features from different time steps and learn global time-series context. A time-index filtering threshold vector is then constructed using global context, and soft-thresholding is applied to suppress noise and highlight fault features in the bearing signals. The LSTM-extracted features are short-circuited with the filtered output to prevent gradient vanishing.

When the LSTM network depth is q, this study selects $q = 1$ as the specific depth for the self-attention LSTM network featuring an adaptive hierarchical structure, with its detailed structure illustrated in Figure 3. As the network depth increases, the self-attention LSTM network with adaptive hierarchy is stacked based on the $q = 1$ configuration.

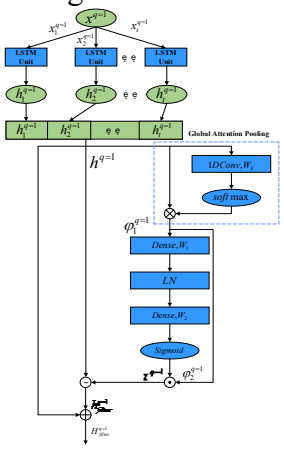


Fig. 3. The adaptive-layer self-attention LSTM network at depth (q =1)

The computation process of the Self-attention LSTM network with adaptive hierarchical levels (q = 1) is as follows:

Step 1: Input the sequential data $x^{q=1} = \{x_1^{q=1}, x_2^{q=1}, \ldots, x_t^{q=1}\}$ into each LSTM unit separately to obtain the hidden layer output at each time step $h^{q=1} = \{h_1^{q=1}, h_2^{q=1}, \ldots, h_t^{q=1}\}$, where the dimension of $h_t^{q=1}$ is $hs \times 1$;

Step 2: After applying the absolute value operation on $h^{q=1}$ (with dimension $C \times 1$), a 1D convolution followed by a softmax activation is used to generate a channel attention weight vector φ_i^{q-1} with dimension $C \times 1$.The weight for the $k-th$ channel is defined in equation (7):

$$\varphi_{i,k}^{q-1} = \frac{exp\left(W_k \cdot abs\left(h_i^{q-1}\right)\right)}{\sum_{j=1}^{C} exp\left(W_j \cdot abs\left(h_i^{q-1}\right)\right)} \tag{7}$$

where W_k represents the weight parameters of the 1D convolution used in the global context pooling operation, and $abs(\,)$ denotes the absolute value operation;

Step 3: The vector $\varphi_1^{q=1}$ is fed sequentially through a series of fully connected (FC) layers, followed by layer normalization (LN), and an additional FC layer. Applying the sigmoid activation function produces a feature vector φ_2^{q-1} with dimensions $C \times 1$, whose values lie between 0 and 1, as shown in equation (8):

$$\varphi_2^{q=1} = sigmoid\left(W_2 LN\left(W_1 \varphi_1^{q=1}\right)\right) \tag{8}$$

Here, W_1 and W_2 represent the weight matrices of the fully connected layers. LN denotes Layer Normalization, a critical component that stabilizes the network training by normalizing the activations across the feature dimension, thereby mitigating internal covariate shift and ensuring robust learning, which is particularly beneficial for the subsequent sequential processing.

Step 4: Multiply φ_2^{q-1} with φ_1^{q-1} element-wise to obtain a timestamp threshold vector τ^{q-1} with dimensions $C \times 1$, as shown in equation (9):

$$\tau^{q-1} = \varphi_2^{q-1} \odot \varphi_1^{q-1} \tag{9}$$

Step 5: Apply soft thresholding strategy to h_i^{q-1} to filter out noise information within the interval $[-\tau^{q-1}, \tau^{q-1}]$ and retain effective fault characteristic information, resulting in filtered feature vector $h_{i,filter}^{q-1}$, as shown in equation (10):

$$h_{i,filter}^{q-1} = \begin{cases} h_i^{q-1} - \tau^{q-1} & h_i^{q-1} > \tau^{q-1} \\ 0 & -\tau^{q-1} \le h_i^{q-1} \le \tau^{q-1} \\ h_i^{q-1} + \tau^{q-1} & h_i^{q-1} < -\tau^{q-1} \end{cases} \tag{10}$$

Step 6: Concatenate $h_{i,filter}^{q-1}$ with h_i^{q-1} to obtain network layer output $H_{i,filter}^{q-1}$ under condition q=1: $H_{i,filter}^{q-1} = h_{i,filter}^{q-1} + h_i^{q-1}$. If additional network layers are added, stack self-attention LSTM networks under condition $q > 1$, using the output from the previous layer H^{q-1} as input for the next layer x^q.

2.4 Adaptive hierarchical self-attention residual network

Deep residual networks, composed of multiple residual modules, can address the gradient vanishing problem caused by increasing network depth. However, residual modules do not have the ability to remove noise from time-series data. To suppress noise in the data and meet the multi-modal feature representation requirements for bearing vibration data, improvements are needed for the ResNet module [21][22].

This paper designs a self-attention ResNet with adaptive layers, enhanced by GCNet and a soft-thresholding strategy within the residual module. A global attention pooling module enables the network to capture contextual channel information and intrinsic inter-channel relationships. As shown in Figure 4's dashed outline, a learned channel threshold vector filters noise, improving fault classification under noisy conditions while reducing parameters. Figure 4 illustrates the self-attention ResNet structure at depth $q = 1$; deeper networks are formed by stacking additional layers accordingly.

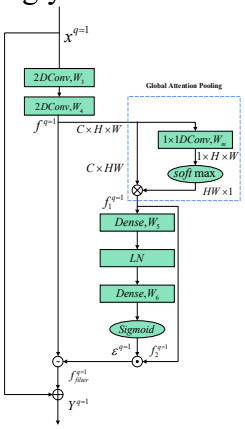


Fig. 4. The self-attention residual network with adaptive layers (q =1)

The computation process of the Self-attention Residual Network with adaptive hierarchical levels ($q = 1$) is as follows:

Step 1: Input the temporal data $x^{q=1}$ sequentially into two 2D convolutional layers for residual mapping, producing feature maps of size $C \times H \times W$, denoted as $f^{q=1}$.

Step 2: Apply the absolute value operation to $f^{q=1}$, then perform a 1×1 convolution followed by a softmax activation to generate intermediate feature

vectors of size $HW \times 1 \times 1$. Subsequently, calculate the dot product between these vectors and the absolute-valued $f^{q=1}$ to obtain global context features of size C×1×1, denoted as $f_1^{q=1}$.

Step 3: Sequentially input $f_1^{q=1}$ into fully connected layers, layer normalization operations, another fully connected layer, and finally utilize sigmoid activation. This process yields feature vectors with dimensions of $C \times 1 \times 1$, denoted as $f_2^{q=1}$, where the numerical values fall within the range of (0,1).

Step 4: Multiply $f_2^{q=1}$ and $f_1^{q=1}$ element-wise to produce the channel threshold vector $\varepsilon^{g=1}$.

Step 5: Filter $f^{q=1}$ using multi-channel noise characteristics based on $\varepsilon^{q=1}$, resulting in the filtered feature map $f_{filter}^{q=1}$;

Step 6: Perform a shortcut connection by adding $f_{filter}^{q=1}$ to $x^{q=1}$ to obtain the output of the self-attention residual module at network level $q = 1$, expressed as $Y^{q=1} = f_{\text{filter}}^{q=1} + x^{q=1}$. When the network depth increases, stack multiple self-attention residual modules at level $q = 1$, using the output Y^q from the previous layer as the input x^q for the subsequent layer.

Concatenate H^q sequentially, then apply global average pooling to Y^q. After unfolding and joining these vectors, the full fault feature vector derived from the self-attention fusion deep neural network's output is obtained.

2.5 Stochastic configuration network classifier

This paper proposes a self-attention fusion deep network for the fault diagnosis of electric locomotive bearings. Time-series fault data are initially processed by a feature extraction network to construct a comprehensive feature space, which is subsequently mapped by a Stochastic Configuration Network (SCNs) classifier, leveraging its universal approximation capability. The objective of the combined architecture is to achieve the complex nonlinear mapping from the constructed fault features to the designated fault categories.

The feature extraction network and the SCNs classifier are treated as an integrated architecture, optimized holistically via gradient descent. While the feature extraction network utilizes standard deep learning methodologies, the optimization pathway from the feature layer to the output requires an alternating process between the configuration of the SCNs and the update of the feature extraction parameters.

During backpropagation, the SCNs classifier's output error is first propagated to the feature extraction layer and subsequently to the input layer. The weights matrix β connecting the SCNs hidden layer to the output layer is updated based on equation(13), while the input parameters (w_j, b_j) of the SCNs hidden layer

and the upstream feature vector Z are updated according to equations (14)–(16). The overall training process is illustrated in Figure 5.

$$e_j = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(O_i - t)^2} \tag{11}$$

$$g(x) = \frac{1}{1 + exp(-x)} - \frac{1}{1 + exp(-(w^T Z + b_j))} \tag{12}$$

$$\frac{\partial e_j}{\partial \beta_j} = \frac{\partial e_j}{\partial O_j} \cdot \frac{\partial O_j}{\partial g_j} \cdot \frac{\partial g_j}{\partial \beta_j} = \frac{(O_j - t)}{\sqrt{(O_j - t)^2}} \cdot g_j \tag{13}$$

$$\frac{\partial e_j}{\partial w_j} = \frac{\partial e_j}{\partial O_j} \cdot \frac{\partial O_j}{\partial g_j} \cdot \frac{\partial g_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial w_j} = \frac{(O_j - t)}{\sqrt{(O_j - t)^2}} \cdot \beta_j \cdot x_j(1 - x_j) \cdot Z \tag{14}$$

$$\frac{\partial e_j}{\partial b_j} = \frac{\partial e_j}{\partial O_j} \cdot \frac{\partial O_j}{\partial g_j} \cdot \frac{\partial g_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial b_j} = \frac{(O_j - t)}{\sqrt{(O_j - t)^2}} \cdot \beta_j \cdot x_j(1 - x_j) \tag{15}$$

$$\frac{\partial e_j}{\partial Z} = \frac{\partial e_j}{\partial O_j} \cdot \frac{\partial O_j}{\partial g_j} \cdot \frac{\partial g_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial Z} = \frac{(O_j - t)}{\sqrt{(O_j - t)^2}} \cdot \beta_j \cdot x_j(1 - x_j) \cdot w_j^T \tag{16}$$

Here, $O = [O_1, O_2, \dots, O_L]$ represents the network's output matrix, and e_j is the error measured by the root mean square error (RMSE) for the $j - th$ node. L denotes the size of the current training batch. Z is the feature vector output by the feature extraction network, serving as the input to the SCNs hidden layer. Within the SCNs framework: w_j and b_j are the input weights and bias of the $j - th$ hidden node; $g_j = g(w_j^T Z + b_j)$ is the output of the $j - th$ hidden node (defined by the basis function $g(x)$ in Equation (12)); β_j is the output weight connecting the $j - th$ hidden node to the output layer; O_j is the $j - th$ component of the output vector O, and t is the target output. The term x_j in Equations (13)–(16) represents the activation of the Sigmoid function itself, $x_j = g(w_j^T Z + b_j)$, which is essential for computing the chain rule derivatives.

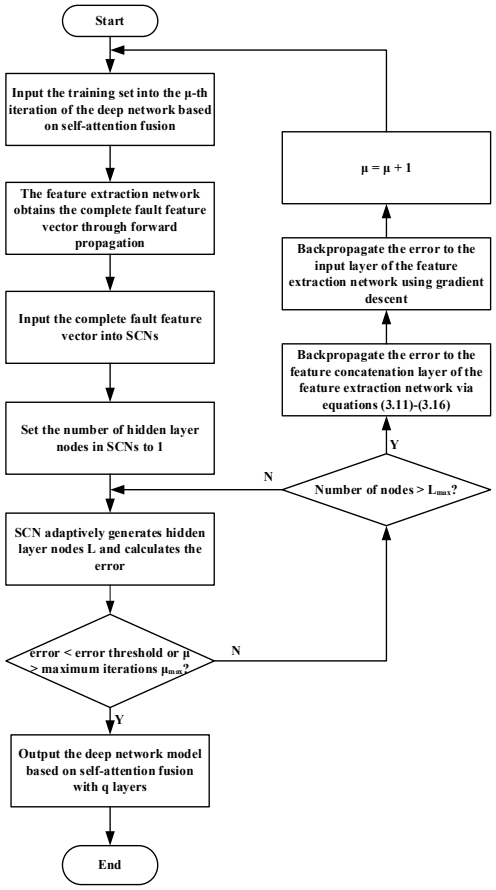


Figure. 5. The training flow chart of the self-attention fusion deep network model based on the stochastic configuration network classifier

3. Experimental verification

3.1 Data description

An experimental setup was established to validate the proposed method on electric locomotive bearings. A 100 mV/g accelerometer measured vibration signals on the load module. Data were sampled at 12.8 kHz for 32 seconds under a 9800 N load. Eight fault conditions with different bearing defects were tested, with characteristic fault frequencies calculated from parameters and speed. Each condition included 500 samples—400 for training and 100 for testing. Figure 6 overviews the experimental test rig to facilitate verification and replication. Figure 7 details the transducer mounting and DAQ connections.

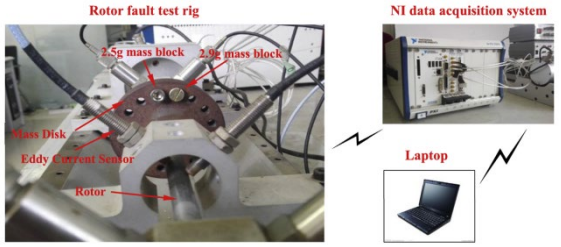


Fig. 6. Rotor fault test rig.

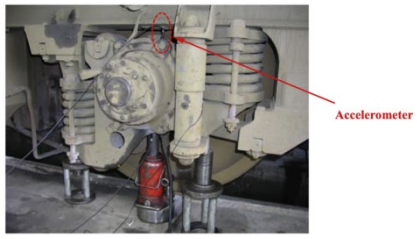


Fig. 7. Electrical locomotive bearing experimental setup

3.2 Experimental setup

To validate our method, a comparison with other existing algorithms was conducted under the same experimental conditions. Three different models were compared: 1-D CNN, LSTM, and Self-attention-based Deep Network (SADN). Their detailed parameter settings are shown in Table 1. Representative faulty components are shown in Figure 8.

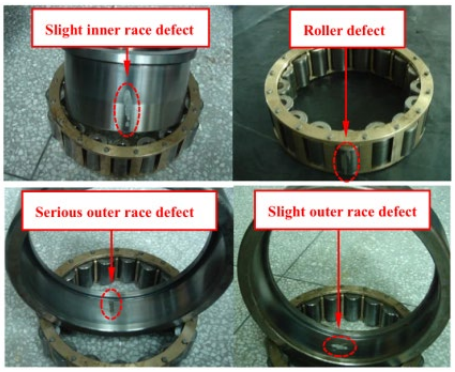


Figure. 8. Faults in the electrical locomotive bearings.

(1) 1D CNN (One-Dimensional Convolutional Neural Network)

The 1D-CNN [23] is a conventional convolutional network for time series analysis, used to extract time-domain features from bearing vibration signals. Convolutional layers capture local patterns, while pooling layers reduce dimensionality and prevent overfitting. During training, kernel parameters are optimized by minimizing the loss to enhance classification accuracy.

(2) LSTM (Long Short-Term Memory Network)

LSTM, a specialized recurrent neural network, excels at modeling long-term dependencies in sequential time-series data. Its gating mechanism overcomes the gradient vanishing issue of traditional RNNs, enabling effective learning of time-dependent vibration features for electric locomotive fault diagnosis.

Table 1

The parameter setting for the three models

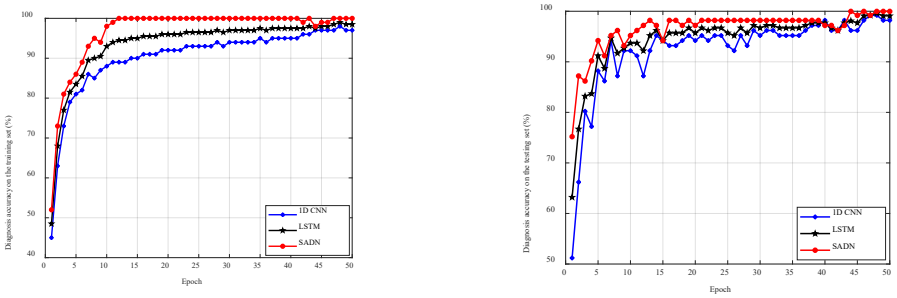
Network Type	Layer Structure	Number of Convolutions/Neurons	Activation Function	Optimizer	Learning Rate	Batch Size
1D CNN	4 layers	The kernel size: 3, 3, 5, 5.	ReLU	Adam	0.001	64
LSTM	5 layers	LSTM units: 32, 32, 32, 64, 64.	Tanh	Adam	0.001	64
SADN	5 layers	Each layer contains one LSTM-Attention and one ResNet-Attention.	ReLU	Adam	0.001	64

3.3 Results and analysis

(1) Performance Comparison

Figure 9 compares the fault diagnosis accuracy of 1D-CNN, traditional LSTM, and the proposed SADN on the same training and test datasets. The 1D-CNN comprises 4 layers, and the LSTM has 5 layers. As shown in Figure 9 (a), all models improve with training. By epoch 30, 1D-CNN reaches 98.42% accuracy, while SADN converges faster, achieving 99.95% at epoch 20. The LSTM lags behind, reaching 97.36% at epoch 50. Figure 9 (b) shows test performance after 50 epochs: SADN achieves 93.95% accuracy, outperforming LSTM (90.52%) and 1D-CNN (91.64%). All models exhibit some overfitting, with test accuracy drops of 6.2% (1D-CNN), 7.1% (LSTM), and 4.8% (SADN) after 30 epochs.

In summary, SADN offers superior diagnostic accuracy and better generalization under noise, outperforming both 1D-CNN and traditional LSTM networks.



(a) Training set (b) Testing set

Fig. 9. Diagnosis accuracies of 1D CNN, LSTM and SADN on the training set and testing set

(2) Algorithm Robustness Analysis

To evaluate the robustness of the proposed method, Gaussian white noise was added to the signals, with the SNR ranging from 0 to 20 dB with a step of 2. Figure 10 displays the fault diagnosis results of the proposed approach compared

to other algorithms across different SNR levels. The average recognition accuracy was employed to measure each algorithm's performance under noisy environments.

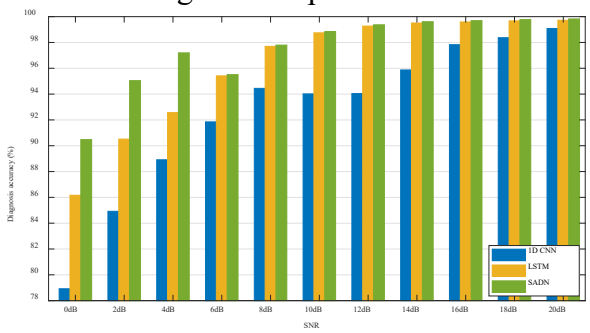


Fig. 10. Diagnosis accuracy

As shown in Figure 10, the proposed method consistently outperforms other algorithms as noise intensity increases. By effectively extracting and integrating global temporal and local spatial features, it mitigates noise impact and leverages complementary strengths, achieving higher classification accuracy than single-network approaches. Experimental results confirm its robust and stable fault diagnosis performance for electric locomotive bearings under high noise conditions.

We further compare the proposed SADN with the VMD-hdCSP [24], which relies on signal decomposition and manually engineered feature extraction. Although VMD-hdCSP achieves commendable accuracy under ideal conditions, its performance is inherently sensitive to parameter configuration and environmental noise fluctuations. Contrarily, SADN adopts an end-to-end learning paradigm augmented with self-attention and adaptive depth mechanisms. The strategy of injecting noise into training samples, which acts as a form of regularization and data augmentation, compels SADN to learn noise-invariant features, thereby enhancing its resilience. The comparison results are listed in Table 2. Specifically, SADN achieved a diagnosis accuracy of 84.1% at SNR=0dB (a substantial increase compared to the 96.4% attained by VMD-hdCSP) and consistently maintained an accuracy exceeding 97% across the entire SNR range of 0dB–20dB. This performance profile unequivocally demonstrates SADN's superior noise robustness and enhanced adaptability compared to traditional feature-engineering-based methods.

Table 2

Accuracy comparison at different SNR levels

Method	SNR = 0 dB (%)	SNR = 10 dB (%)	SNR = 20 dB (%)
VMD-hdCSP	82.7	97.5	98.9
1D CNN	78.4	94.3	99.2
LSTM	82.2	96.8	99.4
SADN	84.1	98.6	99.7

4. Conclusions

This study presents an intelligent diagnostic framework based on a self-attention fusion deep network, targeting key challenges in electric locomotive bearing fault diagnosis, such as strong noise, multi-fault coupling, and limited generalization of traditional models. Designed and validated using real-world vibration data from locomotive bearing test platforms, the method is suitable for online monitoring and intelligent diagnosis of electric locomotives and other high-speed rotating machinery, ensuring practical applicability in railway scenarios. The framework features a spatiotemporal dual-branch architecture: a self-attention-enhanced LSTM captures global temporal features, while a self-attention residual network extracts local spatial features. A hierarchical feature fusion mechanism preserves multi-scale information and improves class separability. To enhance adaptability in complex sample conditions, the method introduces a dynamic depth expansion strategy based on semantic error entropy, alongside an incremental stochastic configuration network classifier for robust generalization in constrained parameter spaces. Experiments confirm its effectiveness and robustness. Under severe noise (SNR = 0 dB), it achieves 84.1 % accuracy, outperforming 1D-CNN and LSTM by 12.7% and 9.5%. Across SNRs from 0 to 20 dB, accuracy ranges from 84.1% to 99.7%, demonstrating strong noise resilience. This work offers both theoretical and practical support for intelligent locomotive fault diagnosis. Future work will explore extensions to multi-source fault coupling, cross-device transfer, and real-time deployment on embedded systems to enhance engineering applicability.

REFERENCES

[1] *Xu, B., Chen, F., Chen, X., Yang, Z., Xie, Q., Zhang, H., & Ye, Y.* (2016). A rolling bearing fault diagnosis method based on VMD–multiscale fractal dimension/energy and optimized support vector machine. Journal of Vibroengineering, 18(6), 3581-3595.

[2] *Qiao, Y., Ma, X., Chen, X., Wang, R., & Jia, L. (*2024). Rolling Bearing Fault Diagnosis Method Based on GWO-VMD-SVM. In Proceedings of the 6th International Conference on Electrical Engineering and Information Technologies for Rail Transportation (EITRT) (pp. 468-479). Springer.

[3] *Zhang, M., Jiang, F., & Zhang, Y*. (2024). Bearing fault diagnosis based on POA-VMD with GADF-Swin Transformer. Measurement, 202, 111666.

[4] *LI, Z. and Zhang, N.*, Rolling bearing fault diagnosis based on BRB AND PSO-SVM.

[5] *Chegini, Ghasemloonia A, Sun Q*. Automatic band selection algorithm for envelope analysis. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2018,233(5):1641-1654.

[6] *X. Li, H. Jiang, M. Niu, R. Wang,* An enhanced selective ensemble deep learning method for rolling bearing fault diagnosis with beetle antennae search algorithm, Mech. Syst. Sig. Process. 142 (2020) 106752.

[7] *H. Shao, H. Jiang, H. Zhang, T. Liang*, Electric locomotive bearing fault diagnosis using a novel convolutional deep belief network, IEEE Trans. Ind. Electron. (2017) 2727–2736.

[8] *Z.H. Wu, H.K. Jiang, S.W. Liu, C.X. Yang,* A Gaussian-guided adversarial adaptation transfer network for rolling bearing fault diagnosis, Adv. Eng. Inf. 53 (2022) 101651.
[9] *Z. He, H. Shao, L. Jing, J. Cheng, Y. Yang*, Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder, Measurement 152 (2019) 107393.
[10] *K. Zhao, H. Jiang, X. Li, R. Wang,* An optimal deep sparse autoencoder with gated recurrent unit for rolling bearing fault diagnosis, Meas. Sci. Technol. 31 (1) (2020) 015005.
[11] *Barbelian, Mihai Alexandru, and Casandra Venera Bălan*, Fault tree event classification by neural network analysis, UPB Sci. Bull. Series D, 79, no. 1 (2017): 165-176.
[12] *M.S. Rathore, S.P. Harsha*, Prognostics analysis of rolling bearing based on bidirectional LSTM and attention mechanism, J. Fail. Anal. Prev. 22 (2) (2022) 704–723.
[13] *X. Li, W. Zhang, Q. Ding,* Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism, Signal Process. 161 (AUG.) (2019) 136–154.
[14] *K. Xu, J. Ba, R. Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, et al.* Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Computer Science, 2015:2048-2057.
[15] *F. Wang, M. Jiang, Q. Chen, S Yang, C Li, H Zhang, et al*., Residual Attention Network for Image Classification[J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6450-6458.
[16] *Y. Cao, J. Xu, S. Lin, Fangyun Wei, Han Hu,* GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond[J]. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 1971-1980.
[17] *Joy J, Peter S, John N*. Denoising using soft thresholding. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2013, 2(3): 1027-1032.
[18] *Rong Y, Nan G, Zhang M, S Chen, S Wang, X Zhang, et al*. Semantic entropy can simultaneously benefit transmission efficiency and channel security of wireless semantic communications[J]. IEEE Transactions on Information Forensics and Security, 2025.
[19] *Yuan X, Jia Z, Xu Z, N Xu, L Ye, K Wang, et al.,* Hierarchical self-attention network for industrial data series modeling with different sampling rates between the input and output sequences. IEEE Transactions on Neural Networks and Learning Systems, 2024.
[20] *Zhang K, Sun M, Han T X, X Yuan, L Guo, T Liu*. Residual networks of residual networks: Multilevel residual networks. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(6): 1303-1314.
[21] *Alaeddine H, Jihene M*. Deep residual network in network. Computational intelligence and neuroscience, 2021, 2021(1): 6659083.
[22] *Li W, Deng Y, Ding M, D Wang, W Sun, Q Li.,* Industrial data classification using stochastic configuration networks with self-attention learning features. Neural Computing and Applications, 2022, 34(24): 22047-22069.
[23] Ince, Turker, et al. Real-time motor fault detection by 1-D convolutional neural networks. IEEE Transactions on Industrial Electronics, 2016, 63(11): 7067-7075.
[24] *Li Z, Lv Y, Yuan R, et al.,* An intelligent fault diagnosis method of rolling bearings via variational mode decomposition and common spatial pattern-based feature extraction. IEEE Sensors Journal, 2022, 22(15): 15169–15177.