



OPEN

Efficient intelligent fault diagnosis method and graphical user interface development based on fusion of convolutional networks and vision transformers characteristics

Chaoquan Mo✉, Ke Huang✉, Houxin Ji, Wenhan Li & Kaibo Xu

Convolutional Neural Networks have been widely applied in fault diagnosis tasks of mechanical systems due to their strong feature extraction and classification capabilities. However, they have limitations in handling global context information. Vision Transformers, by leveraging self-attention mechanisms to capture global dependencies, have shown excellent performance in many visual tasks, but often come with high computational costs. Therefore, this paper proposes a lightweight and efficient intelligent fault diagnosis method based on the fusion of Convolutional Network and Vision Transformer features (FCNVT). This method combines the local feature extraction capability of CNNs with the global dependency capturing ability of ViTs, while maintaining computational efficiency. Random overlapping sampling (ROS) techniques are used to preprocess signals, generating two-dimensional synchronized wavelet transform (SWT) images as inputs to the network. Experimental verification has shown that the proposed method achieves up to 100% classification accuracy, with the model having 7 million parameters and a computational cost of only 0.28 G, outperforming other state-of-the-art methods. Finally, a graphical user interface (GUI)-based mechanical equipment fault detection system was developed using this method, which holds positive implications for advancing the practical application of intelligent fault diagnosis in mechanical equipment.

Keywords Mechanical equipment fault diagnosis, CNN-transformer, Random overlapping sampling, Synchrosqueezed wavelet transform, Graphical user interface

Equipment failure refers to an abnormal state where a component or assembly deteriorates and loses its ability to perform a specified function. Based on this inherent characteristic, it is possible to dynamically monitor, identify, and diagnose the operational model, status parameters, dynamic responses, and signs of failure to achieve the purpose of fault detection and prediction. Due to the presence of working frequency vibration, electrical noise, hydraulic system pulsation noise, and modulation and noise interference in transmission links, the characteristic information in online monitoring is weak, and the robustness of feature extraction methods is poor. Therefore, new solutions for mechanical equipment fault diagnosis have always been a focus and challenge in research both domestically and internationally¹.

Artificial Intelligence (AI) fault diagnosis and prediction hold promise as a potent tool for ensuring the safe operation of equipment: using intelligent algorithms such as artificial neural networks to judge and predict the operational state of equipment to make reasonable maintenance decisions. Smart diagnosis and prediction break away from the traditional reliance on fault mechanisms, diagnosis experts, and professional technicians, making it a core technology of “Smart Industry”^{2,3}. After decades of theoretical research, numerous learning algorithms have emerged within AI, with deep learning being the most prominent⁴. With the success of AlphaGo and ChatGPT, enhanced versions of deep learning^{5,6} have rapidly gained widespread research and application in

College of Mechanical & Electrical Engineering, Wenzhou University, Wenzhou 325035, China. ✉email: 22451439021@stu.wzu.edu.cn; hk125cn@wzu.edu.cn

fields such as natural language processing, computer vision, smart marketing, augmented reality, and fault diagnosis.

Mainstream architectures in deep learning for fault diagnosis include Recurrent Neural Networks (RNNs)⁷ and Convolutional Neural Networks (CNNs)⁸. While RNNs can handle sequential data, they cannot perform parallel computations and are unsuitable for training large datasets. Although Long Short-Term Memory networks (LSTMs)⁹, a variant of RNNs, mitigate some issues with long-range dependencies, they require manual extraction of features in time–frequency domains and lack inherent feature extraction capabilities. Compared to RNN models, CNN models adopt local connectivity and weight sharing designs, making them faster to train, with superior feature learning capabilities and broader applicability. However, CNNs lack the ability to model relationships between targets, treating all parts identically and lacking specificity, making it difficult to process data containing temporal information¹⁰. Additionally, while CNNs excel in deconstructing image information and extracting low-level features, they suffer from limitations such as restricted receptive fields in convolution operations and difficulty capturing global information from images, affecting overall model performance¹¹.

To further improve these issues, researchers worldwide have conducted extensive studies and proposed various modified models. For example, Jia et al.¹² introduced a multi-scale residual attention-based CNN model, utilizing a multi-scale residual attention mechanism to learn distinctive multi-scale features from vibration signals and enhance CNN's effectiveness in identifying fault features by leveraging multi-scale feature denoising. Li et al.¹³ proposed a deep learning bearing fault diagnosis method combining CNN and LSTM, adding attention to input data segments to visualize learned weights and obtain interpretable diagnostic results, effectively mining fault features. Long et al.¹⁴ proposed a motor fault diagnosis method using a multi-sensor information-driven attention mechanism and an improved AdaBoost, enhancing the robustness, generalization capability, and accuracy of fault diagnosis. These methods alleviate some of the problems with CNNs and RNNs to some extent, but there is still a pressing need in actual production for more precise, stable, and faster inference speed methods for mechanical equipment fault diagnosis to minimize loss caused by failures.

The self-attention network based on Transformer architecture was first proposed in 2017¹⁵, and due to its advantages such as parallel computation, capturing long-distance dependencies, and strong global feature learning capabilities, it quickly found wide application in natural language and machine vision fields. The Vision Transformer (ViT) model achieved performance comparable to CNNs¹⁶ and is considered a strong alternative to CNN models. Improved Transformer models like Swin Transformer (SwinT)¹⁷, Shuffle Transformer¹⁸, and Cswin¹⁹ continue to emerge. Consequently, Transformer-based fault diagnosis and prediction research have become a hotspot for scholars globally.

Fang et al.²⁰ proposed the CLFormer self-attention model for bearing fault identification, demonstrating high classification performance under strong noise conditions; Liang et al.¹¹ combined sub-domain adaptation with Vision Transformer, realizing bearing fault diagnosis under varying operating conditions based on the integrated Transformer network. Traditional Transformer models mostly adopt hierarchical frameworks, making feature integration across different layers difficult and weakening the learning capability of local features. Moreover, Transformer models have large parameter scales, requiring substantial computational resources, leading to lower efficiency and high computational costs, making rapid deployment challenging in industrial settings²¹.

As discussed, existing deep learning models such as CNNs and LSTMs excel in local feature extraction but face limitations when it comes to processing global information and long-range dependencies. Moreover, although Transformer models have achieved significant success in fields like natural language processing, vision, and fault diagnosis, their application in mechanical fault diagnosis is still hampered by high computational costs and the need for large amounts of data, making deployment in resource-constrained industrial environments challenging.

Based on the above, this paper proposes a lightweight and efficient intelligent fault diagnosis method (FCNVT) that integrates the characteristics of convolutional networks and vision transformers. The FCNVT constructs a network model that fuses the features of convolutional networks and vision transformers, employing random overlapping sampling techniques for data augmentation of vibration signals, and generating two-dimensional synchronized wavelet transform (SWT) images as inputs to the model. By utilizing the strong learning capabilities of FCNVT to automatically extract temporal and spatial features from the images, the diagnosis of the fault status of mechanical equipment can be achieved. Finally, a graphical user interface (GUI)-based mechanical equipment fault detection system has been developed using this method, contributing to the advancement of practical applications in intelligent fault diagnosis for mechanical equipment.

The main contributions of this paper are as follows:

- (1) The FCNVT model is proposed for fault diagnosis in mechanical equipment, combining convolutional layers and Transformer layers to capture local and global features while maintaining efficiency, utilizing local connection patterns and sparse attention mechanisms to reduce computational load and parameter count.
- (2) The FCNVT model achieves the fusion of multi-scale features through multi-layer convolution operations and multi-head self-attention mechanisms. The convolutional layers are responsible for extracting local features, while the self-attention mechanism captures global dependencies, all while maintaining lower computational costs. This design of multi-scale feature fusion enables the model to more comprehensively understand and recognize complex fault patterns.
- (3) By adopting random overlapping sampling techniques to fully utilize the information in the original signal data and simulate the randomness of signals during fault occurrences, SWT is used to improve the time–frequency resolution and reduce redundant information, endowing the model with good diagnostic capability and generalization performance.

- (4) A GUI-based mechanical equipment fault detection system is developed based on this method, facilitating the practical application and information management of intelligent fault diagnosis in mechanical equipment.

The rest of the paper is organized as follows: Section “[Theoretical foundations](#)” outlines the relevant theoretical foundations; Section “[Proposed method](#)” elaborates on the proposed method; Section “[Fault diagnosis experiment verification and analysis](#)” validates and analyzes the model through fault diagnosis experiments; Section “[Development and application of graphical user interface](#)” discusses the development and application of the graphical user interface; Section “[Conclusion](#)” summarizes the work of this paper.

Theoretical foundations

Convolutional neural networks

The structural characteristics of Convolutional Neural Networks (CNNs) are reflected in three aspects: local connection, weight sharing, and spatial pooling. Unlike fully connected neural networks where neurons in each layer are fully interconnected, CNNs implement local connections between layers through convolutional kernels (filters) that are much smaller than the input size. The size of the convolutional kernel is referred to as the receptive field of that layer. The application of local connections significantly reduces the number of connections in the network and allows it to handle arbitrary-sized input data. Weight sharing means that the weights and biases of the convolutional kernels are shared across the same convolutional layer, meaning the parameter values remain constant during convolution operations at different positions in the input data. Spatial pooling is a down-sampling method that reduces the dimensionality of the input data, thereby decreasing the computational load of the model. These three structural characteristics of CNNs overcome the shortcomings of fully connected neural networks, such as inability to handle large-scale data (like high-definition images), excessive parameter counts, and susceptibility to overfitting, enabling the training of networks with dozens or even hundreds of layers and achieving widespread application in the processing of various types of data²².

The convolutional layer is the core structure in a CNN network responsible for adaptive feature extraction. Typically composed of one or multiple convolutional kernels, the neurons in the convolutional layer connect locally with the preceding feature layer via the convolutional kernel. Each convolutional kernel independently performs convolution operations on the input features and calculates the final output features through linear superposition. The feature calculation process for the r -th layer of the convolutional layer can be expressed as follows:

$$C_n^r = \sum_m f(C_m^{r-1} * w_n^r + b_n^r) \quad (1)$$

where C_m^{r-1} is the output of the previous layer of the neural network, m is the number of convolutional kernels in the r -th convolutional layer, $*$ represents the convolution operation between the input feature and the convolutional kernel, w_n^r is the corresponding parameters of the convolutional kernel, b_n^r is the bias term, and f is the activation function for the output features of the network. Activation functions are usually added after the convolutional network layers to enhance the model's nonlinear modeling capability and the linear separability of features, thus improving the model's learning ability for complex problems. In CNN networks, ReLU is the most popular activation function. The convolution operation is illustrated in Fig. 1.

Vision transformer neural network

The Vision Transformer (ViT)¹⁶ model utilizes a multi-head self-attention mechanism, which enables it to fully learn the global information of the input signal without being limited by local interactions, thereby addressing the issue of limited receptive fields in convolutional kernels²³. Additionally, it can learn long-range dependencies within the signal. The model dynamically generates weights for different nodes through similarity measurements and aggregates information, possessing the ability to flexibly respond to changes in input information. This effectively avoids many inherent limitations of Convolutional Neural Networks. ViT replaces convolutional structures entirely with the Transformer architecture to accomplish classification tasks and has achieved

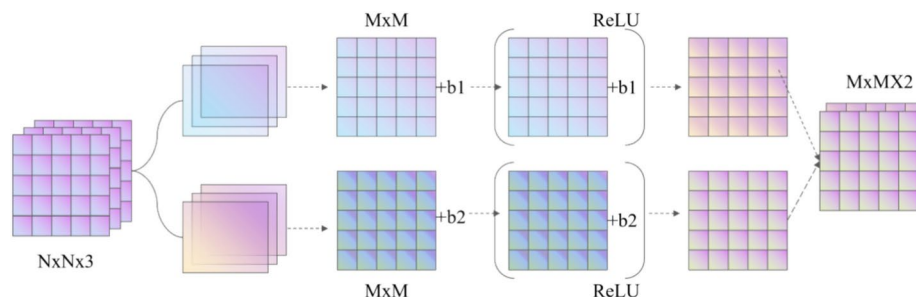


Fig. 1. Convolution operation.

performance surpassing that of CNNs on superscale datasets^{24–26}. As shown in Fig. 2, the Vision Transformer model primarily consists of three parts: input encoding, Transformer encoder, and classification module.

(1) Input Encoding (Image Tokenization)

The ViT encoding consists of patch embedding and positional encoding. The image $x \in \mathbb{R}^{H \times W \times C}$ is divided into patches and linearly transformed into a series of flattened Patches $x_p \in \mathbb{R}^{N \times (p^2 \cdot c)}$, where N is given by:

$$N = HW/P^2 \quad (2)$$

where H is the height of the image, W is the width of the image, and (H, W) represents the image resolution; p is the height and width of a Patch, p^2 is the size of a Patch, and (p, p) represents the Patch resolution.

In Fig. 2, the patch embedding operation is:

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; L x_p^N E] + E_{\text{pos}} \quad (3)$$

In Fig. 2, positional encoding is generated using sine and cosine functions, specifically:

$$\begin{aligned} \text{PE}_{(\text{pos}+k, 2i)} &= \sin[(\text{pos} + k) / 10000^{2i/d_{\text{model}}}] \\ \text{PE}_{(\text{pos}+k, 2i+1)} &= \cos[(\text{pos} + k) / 10000^{2i/d_{\text{model}}}] \end{aligned} \quad (4)$$

where pos is the position in the input sequence; k is any fixed offset; d_{model} is the output dimension of the model's sublayers and embedding layer; i is the dimension of the positional encoding vector.

(2) Transformer encoder

The ViT encoder, as seen in Fig. 2, mainly includes a Multi-Head Self-Attention (MSA) mechanism layer and a Multilayer Perceptron (MLP).

The MSA, as shown in Fig. 2, is composed of linear layers, dot-product self-attention layers, Concatenate layers, and linear layers. It maps queries and a set of key-value pairs to outputs, where the output is a weighted sum of the values, with the weights assigned to each value computed using the keys. The calculation of MSA is:

$$\begin{aligned} \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Concatenate}(\text{Head}_1 \cdots \text{Head}_n) W^T \\ \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d_k}}\right) \times \mathbf{V} \end{aligned} \quad (5)$$

where n is the number of heads; d_k is the dimensionality of the query or key; W is the weight matrix.

The attention function is calculated for a set of queries packed into matrix $\mathbf{Q}$, and keys and values are also packed into matrices $\mathbf{K}$ and $\mathbf{V}$. The Concatenate function is used to concatenate the outputs from multi-head

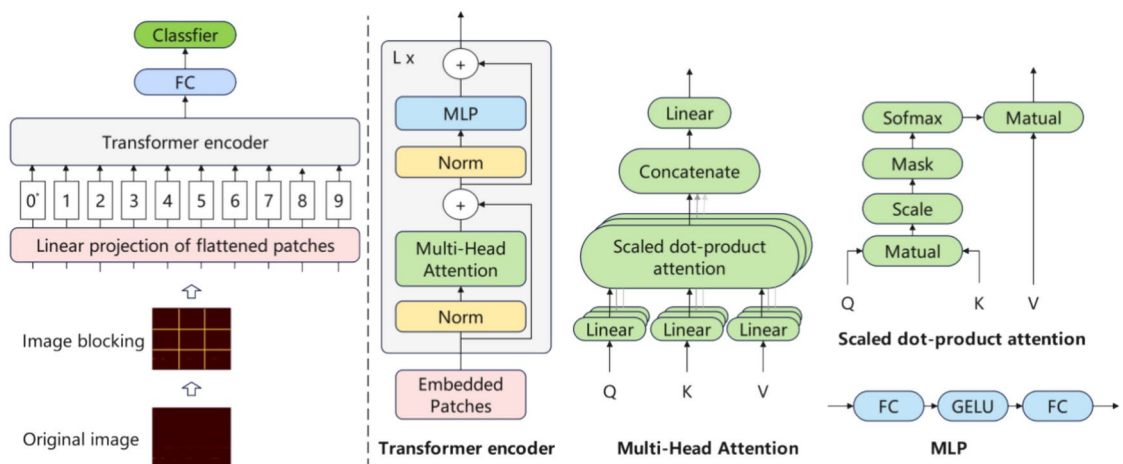


Fig. 2. Structure of the vision transformer model.

self-attention calculations, and Softmax is used to obtain the weights of the values. To enable the network to improve fault diagnosis accuracy by increasing depth, residual connections are used within each sublayer, and layer normalization²⁷ is applied at the end of each sublayer. Therefore, the output of each sublayer is:

$$\text{Out} = \text{LayerNorm}(\text{X} + \text{Sublayer}(\text{X})) \quad (6)$$

where $\text{Sublayer}(\text{X})$ represents the processing function of the self-attention mechanism layer and the 1D convolution layer, and LayerNorm is the layer normalization function.

After passing through the Transformer encoder, the image is transmitted to the MLP. The MLP contains two Fully Connected (FC) layers and a non-linear layer with a GELU (Gaussian Error Linear Unit) activation function. The fully connected layers map the distributed features learned by the multi-head attention mechanism to the sample space, and the feature calculation process for the (l) -th layer is as follows:

$$Z_L = \text{MSA}(\text{LayerNorm}(Z'_L)) + Z'_L \quad (7)$$

where $L = 1, 2, \dots, l$.

The MLP performs a nonlinear mapping of the image matrix output from the Transformer encoder, and after passing through the Softmax layer, it achieves image recognition.

Proposed method

Network architecture description

The proposed lightweight and efficient network architecture that integrates the characteristics of convolutional networks and vision transformers, along with specific parameters, is illustrated in Fig. 3. This network consists of a convolutional module, Transformer Module 1, Transformer Module 2, an average pooling module, and a classifier. The approach aims to enhance the model's efficiency and performance by combining local connectivity patterns with sparse attention mechanisms, while also reducing computational costs.

- (1) Convolutional Module: The convolutional module employs four layers of 3×3 convolutions (with a stride of 2) to reduce the activation maps before feeding them into the Transformer layers. The strong inductive biases of convolutional layers allow the model to learn low-level information effectively at early stages, a capability that Transformers lack. Hence, several convolutional layers are stacked before the Transformer layers.

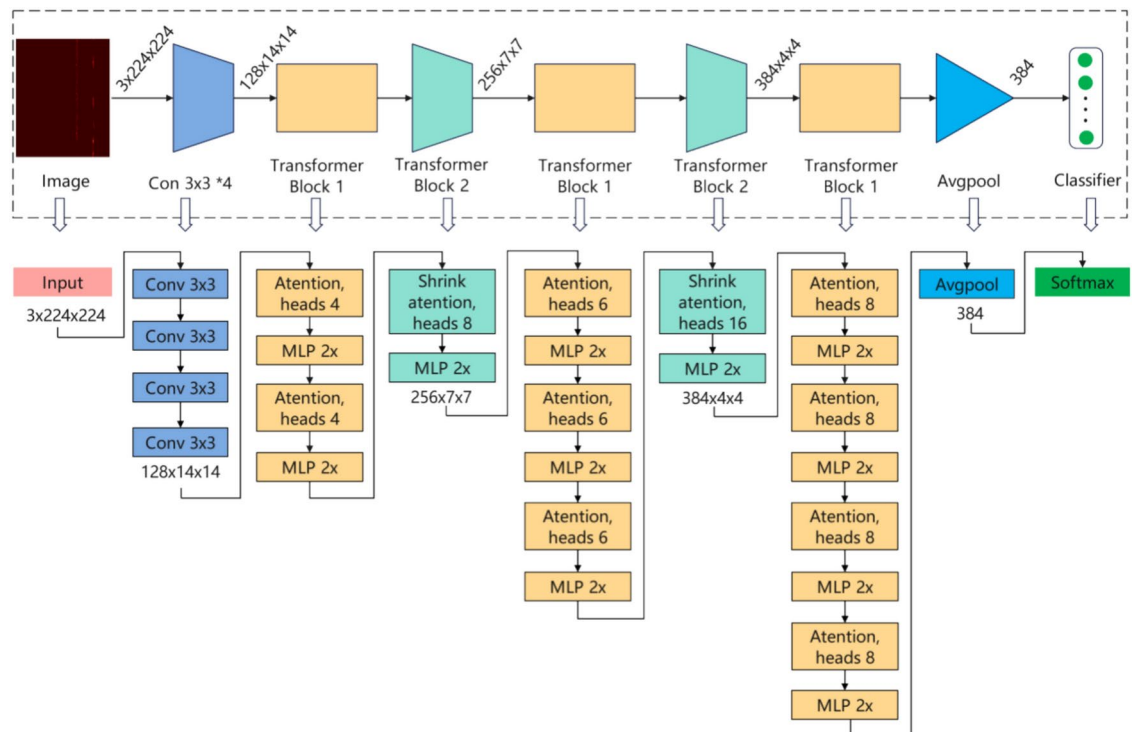


Fig. 3. Proposed model architecture.

- (2) Transformer Layers: The Transformer layers consist of two types of modules: Transformer Block 1 and Transformer Block 2. In the context of Vision Transformers, MLPs are generally more computationally expensive in terms of runtime and parameter count compared to attention blocks. In this paper, the MLP is implemented as a 1×1 convolution followed by standard batch normalization, and all non-linear activation functions are Hardswish. To reduce the computational cost at this stage, the expansion factor of the convolutions is reduced from 4 to 2.

Transformer Block 1 versus Transformer Block 2: The difference lies in the type of attention blocks they use. As shown in Fig. 4, the left diagram represents regular attention, while the right diagram shows shrink attention, which halves the activation maps. Here, $C \times H \times W$ denotes the size of the input activation map, D is the key matrix dimension, and N is the number of heads.

Traditional Vision Transformers and similar architectures use positional embeddings to provide positional information, adding a fixed positional encoding to each input token. However, positional embeddings are only appended to the input sequence, implying that positional information is not dynamically updated in higher-level representations. This could lead to unnecessary consumption of representational capacity by positional information in intermediate representations. Therefore, our structure does not include a class token or positional encoding but introduces a new attention bias mechanism allowing each attention head to learn independent positional biases, thereby directly injecting relative positional information within each attention block.

The scalar attention value between two pixels $(x, y) \in [H] \times [W]$ and $(x', y') \in [H] \times [W]$ in a head $h \in [N]$ is calculated as:

$$A_{(x,y),(x',y')}^h = Q_{(x,y),:} \cdot K_{(x',y'),:} + B_{|x-x'|,|y-y'|}^h \quad (8)$$

where $A_{(x,y),(x',y')}^h$ represents the attention value between positions (x, y) and (x', y') , indexed by the attention head. $Q_{(x,y),:} \cdot K_{(x',y'),:}$ is the classical attention calculation, i.e., the dot product of Query and Key. $B_{|x-x'|,|y-y'|}^h$ is a learnable attention bias that is computed based on the relative position between (x, y) and (x', y') . This bias is translation invariant, meaning it depends only on the relative distance between positions.

Each attention head has $H \times W$ parameters corresponding to different pixel offsets. Here, H and W are the height and width of the activation map, respectively. Through this approach, each attention head can learn specific relative positional relationships, such as those between adjacent pixels.

- (3) Average Pooling: Following the last Transformer layer, average pooling is introduced to progressively reduce the dimensions of the activation maps, thereby reducing the number of parameters and computational complexity of the model. This pooling operation helps maintain the depth of the model while reducing its width, contributing to higher computational efficiency.

Data division and transformation

Random overlapping segmentation

To make full use of the information contained in the original signal data and to simulate the randomness of signals when faults occur, this paper adopts a signal division method known as random overlapping segmentation. By “random,” it is meant that the positions at which samples are segmented are chosen randomly. “Overlapping

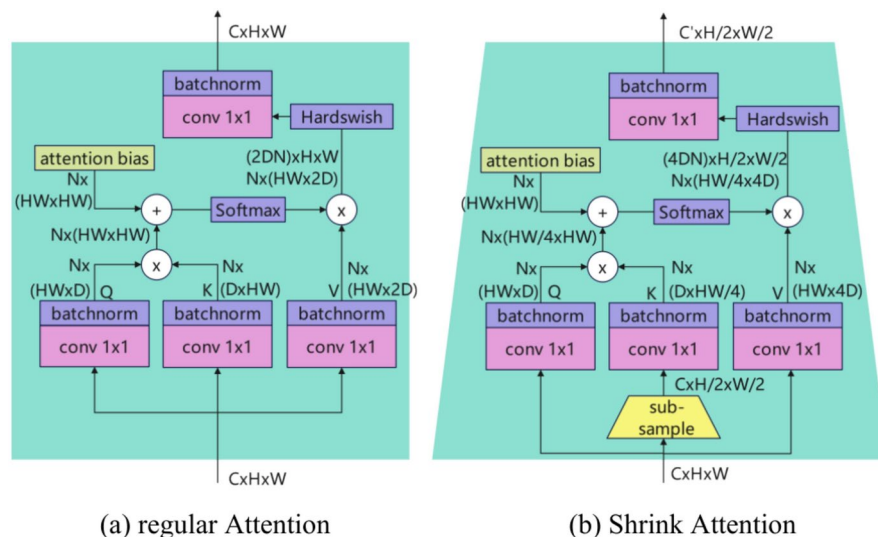


Fig. 4. Two types of attention blocks.

segmentation” refers to the fact that adjacent samples share a portion of the same data points. Using random overlapping segmentation to construct the sample set, different batches of samples are generated each time, as shown in Fig. 5. With sufficient training, theoretically, signal segments from any position can be extracted, leading to better data augmentation effects and thus enhancing the generalization capability of the diagnostic model.

Synchrosqueezed wavelet transform (SWT)

Time–frequency analysis, as an effective method for analyzing time-varying signals, allows for a deeper understanding of the multiple components contained within real engineering signals and calculates the instantaneous frequency and amplitude at each moment²⁸. However, constrained by the Heisenberg uncertainty principle, classical time–frequency analysis methods produce results with limited resolution in the time–frequency domain, which is not conducive to their application in engineering. Synchrosqueezed Wavelet Transform is a special signal rearrangement method that sharpens the time–frequency curves of continuous wavelet transforms and improves the frequency accuracy of these curves by reassigning the values of wavelet coefficients $W_s(t, s)$ at different local points (t', s') near (t, s) .

Synchronous Squeezing Wavelet Transform is based on wavelet transform²⁹. First, Continuous Wavelet Transform (CWT) is performed on the signal $S(t)$ to obtain the wavelet coefficients $W_s(a, b)$. Given a mother wavelet function ψ , the continuous wavelet transform of $S(t)$ is:

$$W_s(a, b) = \int_{-\infty}^{+\infty} s(t) \frac{1}{\sqrt{a}} \overline{\psi\left(\frac{t-b}{a}\right)} dt \quad (9)$$

where $\overline{\psi\left(\frac{t-b}{a}\right)}$ is the conjugate of the wavelet function $\psi\left(\frac{t-b}{a}\right)$; a and b are the scale and translation factors, respectively; $W_s(a, b)$ is the calculated wavelet coefficient.

Based on the one-to-one correspondence between the wavelet scale factor and frequency, the obtained wavelet coefficients $W_s(a, b)$ can be represented in the time–frequency domain, yielding the time–frequency plot of the wavelet transform. By reassigning $W_s(a, b)$, one can extract the instantaneous frequency lines, resulting in a more concentrated time–frequency plot. When the signal is a harmonic wave function, $s(t) = A \cos(\omega t)$ selects a wavelet that is zero in the negative frequency domain. According to the Plancherel theorem, performing continuous wavelet transform on a harmonic wave yields:

$$W_s(a, b) = \frac{1}{2\pi} \int \hat{s}(\xi) \sqrt{a} \overline{\hat{\psi}(a\xi)} e^{ib\xi} d\xi = \frac{A}{4\pi} \sqrt{a} \overline{\hat{\psi}(a\omega)} e^{ib\omega} \quad (10)$$

where, when $\hat{\psi}(\xi)$ is concentrated at $\xi = \omega_0$, the wavelet coefficient $W_s(a, b)$ is correspondingly concentrated at scale $a = \omega_0/\omega$. Taking the partial derivative of the wavelet coefficients, the instantaneous frequency can be obtained:

$$\omega_s(a, b) = \begin{cases} -i(W_s(a, b))^{-1} \frac{\partial}{\partial b} W_s(a, b), & |W_s(a, b)| > 0 \\ \infty, & |W_s(a, b)| = 0 \end{cases} \quad (11)$$

where transforming the wavelet coefficients $W_s(a, b)$ from the time-scale plane to the time–frequency plane becomes $W_s[\omega_s(a, b), b]$. The synchronous squeezing wavelet transform value $T_s(\omega, b)$ of the wavelet coefficients at discrete scales a_k can be obtained by squeezing the values in the frequency band centered at ω_l with a bandwidth of $[\omega_l - \Delta\omega/2, \omega_l + \Delta\omega/2]$, given by the formula:

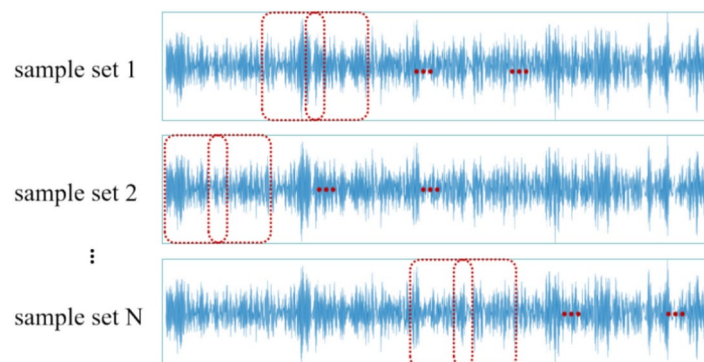


Fig. 5. Schematic of random overlapping segmentation of sample sets.

$$T_s(\omega_l, b) = (\Delta\omega)^{-1} \sum_{a_k: |\omega(a_k, b) - \omega_l| \leq \Delta\omega/2} W_s(a_k, b) a_k^{-3/2} (\Delta a)_k \quad (12)$$

where $a_k - a_{k-1} = (\Delta a)_k$ is the discrete scale interval, and $\Delta\omega = \omega_l - \omega_{l-1}$ is the frequency interval.

As shown in Fig. 6, from top to bottom, the figures represent the time-domain plots, CWT plots, and SWT plots for different health states. It can be seen that the CWT plots generate a lot of redundant information because each scale produces a subplot on the time–frequency plane, and their energy distribution on the time–frequency plane is relatively blurred. The SWT plot, by reassigning the energy of the CWT, reduces redundant information and redistributes the energy to more precise locations on the blurred time–frequency plane, enhancing frequency resolution.

Network diagnosis process

The diagnosis process of the proposed scheme is illustrated in Fig. 7. The diagnostic process of the model can be summarized in three parts: sample set preparation, model training, and model evaluation.

- (1) *Sample set preparation*: Obtain vibration data from mechanical equipment under different operating conditions. Use random overlapping segmentation technology to augment the original vibration data. Further divide the data into training and testing datasets. Finally, generate the corresponding SWT image samples through Synchronized Compressive Wavelet Transform.
- (2) *Model training*: After constructing the proposed model architecture and configuring the parameters, divide the training and testing datasets. Next, set the model training parameters, including the learning rate, loss function, and training batch size. During the training process, the learning rate is set to 0.001, and the Adam optimizer is used for parameter updates. Additionally, the loss function chosen is the categorical cross-entropy loss function; the batch size for each step of the model's training input is 64. After each round of model training, the model's diagnostic accuracy is tested using the test set to further validate the diagnostic capabilities of the model.
- (3) *Model evaluation*: To further examine the model's ability to learn fault features and the degree of mastery, the model is evaluated using values of the loss function, diagnostic classification accuracy, confusion matrices, and t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. The software environment used for the model to perform fault diagnosis is: PyCharm Community V2019.3.3, TensorFlow V2.13. The hardware environment is: Intel i7-12700H @ 2.3 GHz, RAM 32 GB, GPU RTX3060.

Fault diagnosis experiment verification and analysis

Experiment one

(1) Description of dataset I

The Case Western Reserve University (CWRU) bearing dataset is one of the most widely used open-source datasets for mechanical equipment fault diagnosis. To enhance the reference value of our work, we chose to conduct experiments on this public dataset. The experimental setup of the CWRU bearing dataset is shown in Fig. 8. The bearing models include SKF-6205 drive end bearings and SKF-6203 fan end bearings. The sampling frequencies are 12 kHz and 48 kHz, with vibration acceleration signals of faulty bearings collected by accelerometers. The dataset is divided into four operational conditions, with motor loads of 0HP, 1HP, 2HP, and 3HP, corresponding to rotational speeds of 1,797 r/min, 1,772 r/min, 1,750 r/min, and 1,730 r/min, respectively. Fault types are categorized into three major classes: inner race faults (IR), ball faults (B), and outer race faults (OR). Each fault type includes three fault diameters: 0.1778 mm, 0.3556 mm, and 0.5334 mm.

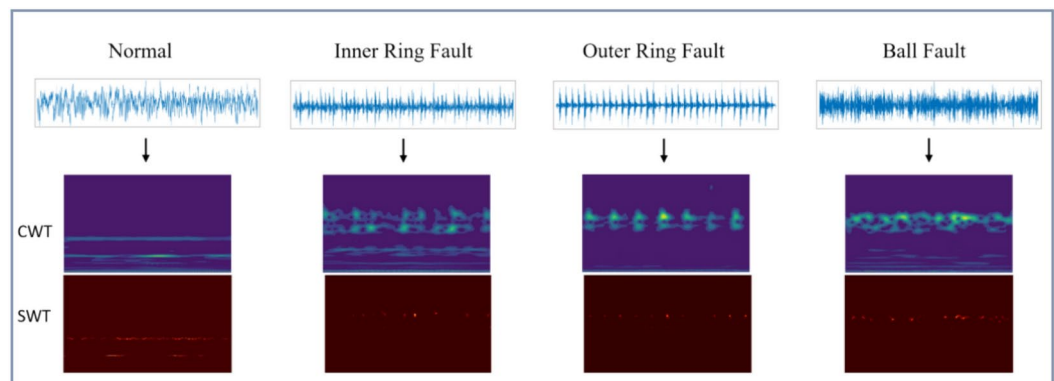


Fig. 6. Time-domain plots, CWT plots, and SWT plots for different health states.

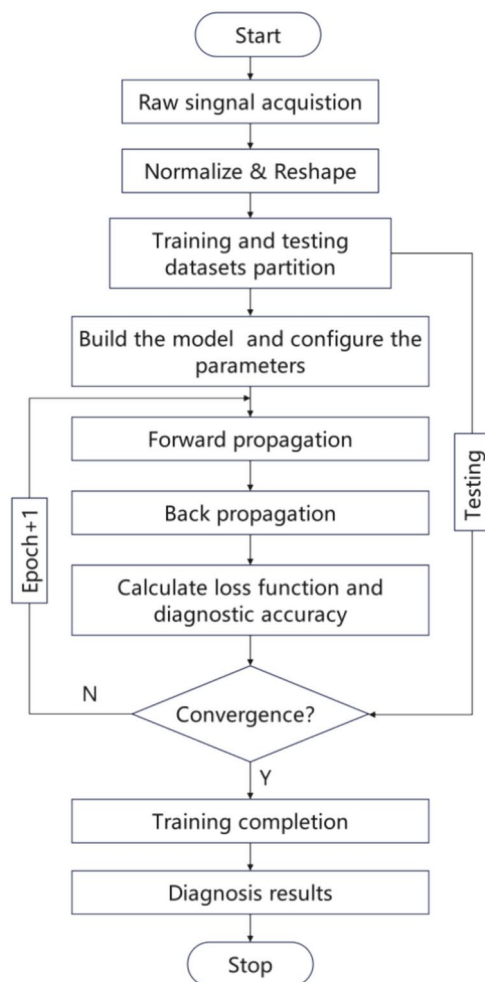


Fig. 7. Diagnostic process of the proposed scheme.

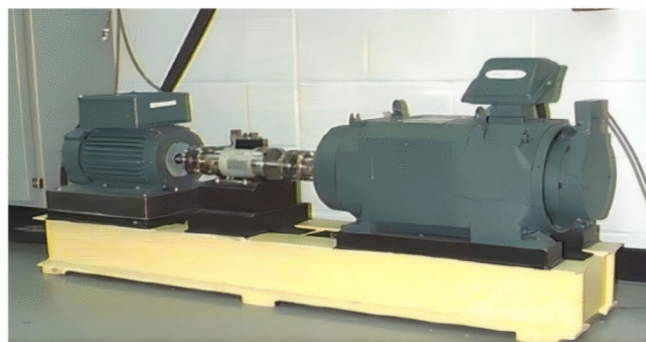


Fig. 8. CWRU bearing fault diagnosis experimental platform.

The experiment selects the drive end bearing with a sampling frequency of 12 kHz and a motor load of 2HP, corresponding to a rotational speed of 1,750 r/min. It includes ten fault states, where labels 0 to 8 represent faulty bearing data, and label 9 represents normal bearing data. Each experiment uses a total of 1,000 samples, with each sample containing 1,024 data points. The data is split into training and testing sets in a 7:3 ratio. The specific experimental data is shown in Table 1.

(2) Diagnosis results and analysis of dataset I.

- Analysis of diagnostic accuracy and loss function values

Label	Load / RPM	Status	Fault sizes(mm)	Training set	Testing set	sample lengths
0	2HP, 1750 r/min	B007	0.1778	70	30	1024
1	2HP, 1750 r/min	B014	0.3556	70	30	1024
2	2HP, 1750 r/min	B021	0.5334	70	30	1024
3	2HP, 1750 r/min	IR007	0.1778	70	30	1024
4	2HP, 1750 r/min	IR014	0.3556	70	30	1024
5	2HP, 1750 r/min	IR021	0.5334	70	30	1024
6	2HP, 1750 r/min	OR007	0.1778	70	30	1024
7	2HP, 1750 r/min	OR014	0.3556	70	30	1024
8	2HP, 1750 r/min	OR021	0.5334	70	30	1024
9	2HP, 1750 r/min	Normal	—	70	30	1024

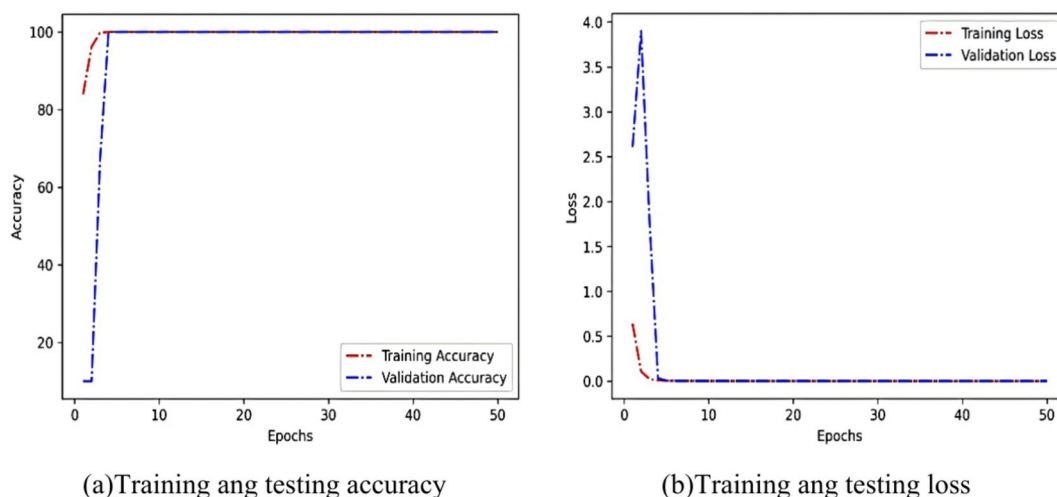
Table 1. Description of bearing experiment data.**Fig. 9.** Loss function values and diagnostic accuracy.

Figure 9 illustrates the changes in accuracy and loss values during the training process. It can be observed that the method proposed in this paper demonstrates excellent diagnostic effectiveness for bearing faults, achieving nearly 100% accuracy after about six iterations, requiring very few iterations to reach a high level of accuracy. Around six epochs, the loss function converges towards 0. Additionally, the consistency in diagnostic accuracy and loss function values between the training and testing sets throughout the training process indicates that the model did not experience overfitting or underfitting during hyperparameter tuning. This shows that the proposed method has excellent diagnostic capability and can achieve accurate and stable recognition results for mechanical equipment faults.

- Analysis of t-SNE results

To verify the feature extraction capability of the proposed method, the t-distributed stochastic neighbor embedding (t-SNE) algorithm was employed to visualize the features of the input and output layers, characterizing the model's feature learning ability. As shown in Fig. 10. From Fig. 10a, it can be seen that the original input data in the two-dimensional space are mixed together in a chaotic manner, with a high degree of confusion. From Fig. 10b, it can be observed that after the model extracts and learns the features of the input signal data, the data representing different health states are well grouped. The intra-class compactness and inter-class separability of the health state features in the output layer are significantly improved. The inter-class distances between different health states become clear and wide, while the data within the same class become more compact, and the confusion between different features disappears. This indicates that the proposed method can effectively extract and learn the features of different health states in bearing data and accurately identify the corresponding fault types.

- Analysis of confusion matrix results

To further illustrate the effectiveness and classification performance of the proposed method, a confusion matrix was used to analyze one of the test results on this dataset. Figure 11 visualizes the classification results of the test set using the confusion matrix. It can be seen that all prediction results are correct, achieving 100% diagnostic

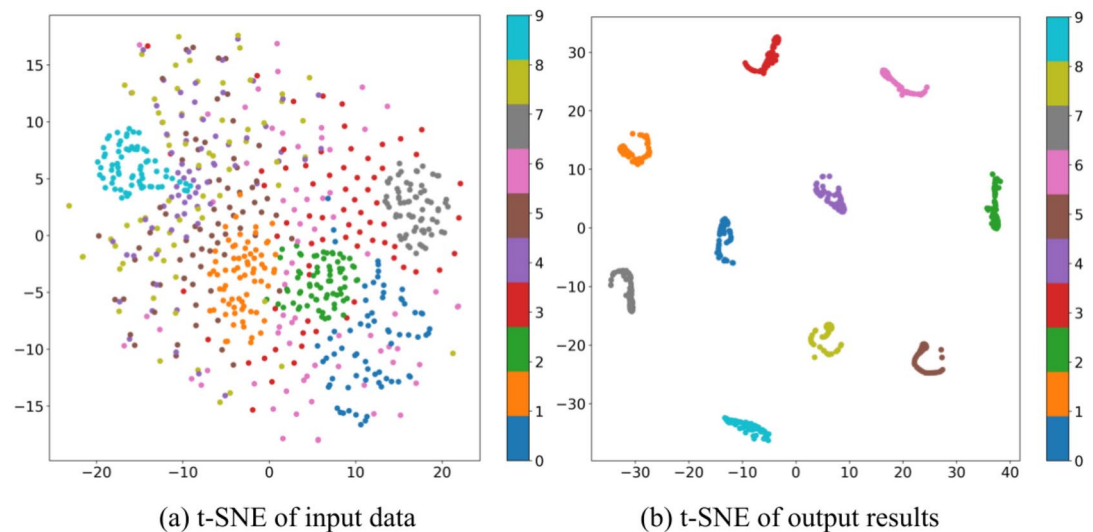


Fig. 10. Feature visualization.

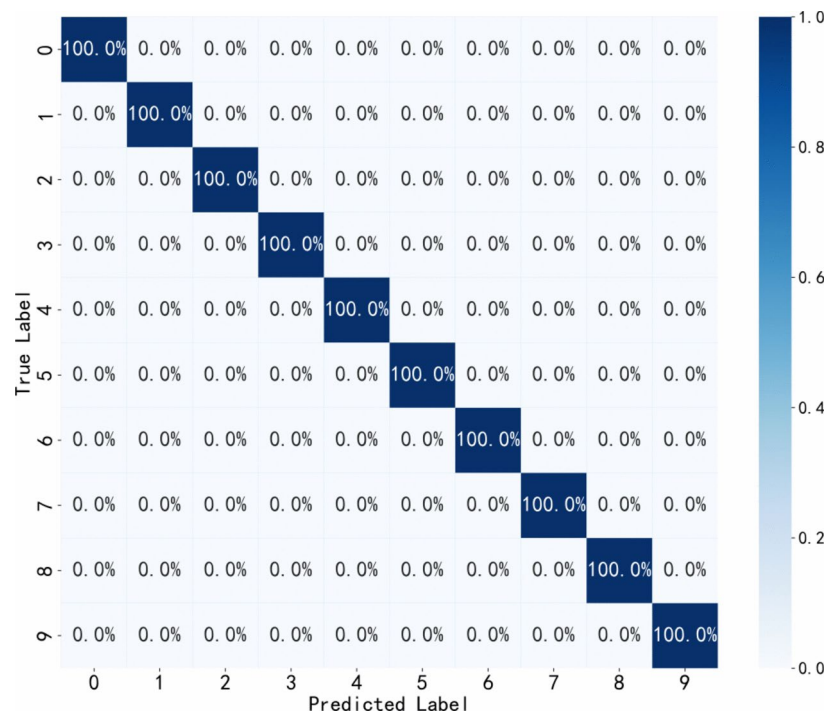


Fig. 11. Confusion matrix.

accuracy in the final diagnostic results, consistent with the results in Fig. 10. This further confirms that the model has excellent fault diagnostic classification performance.

- Verification of training sample extraction at random positions

To determine the optimal overlap ratio and segmentation parameters, we designed a series of comparative experiments. The experimental data came from the CWRU bearing dataset, covering different fault types and operating conditions. The specific experimental settings are as follows: overlap ratios: 0%, 25%, 50%, 75%; segmentation parameters: 128 points, 256 points, 512 points, 1024 points.

First, we fixed the segmentation parameter at 1024 points and varied the overlap ratio to observe its impact on the model's generalization ability and diagnostic accuracy. The experimental results are shown in Table 2. From the table, it can be seen that when the overlap ratio is 50%, the model's classification accuracy reaches the

Overlap ratio (%)	Accuracy (%)	Computational complexity (G)
0	98.2	0.24
25	99.1	0.25
50	100.0	0.28
75	98.9	0.32

Table 2. Impact of different overlap ratios.

Segmentation parameter (Points)	Accuracy (%)	Computational complexity (G)
128	96.3	0.20
256	98.5	0.22
512	99.2	0.25
1024	100.0	0.28

Table 3. Impact of different segmentation parameters.



Fig. 12. Fault diagnosis test bench for gearbox at Southeast University.

highest at 100%, with a computational complexity of 0.28 G. Therefore, an overlap ratio of 50% is a relatively ideal choice.

Next, we fixed the overlap ratio at 50% and changed the segmentation parameters to observe their effect on model performance. The experimental results are shown in Table 3. From the table, it can be seen that when the segmentation parameter is 1024 points, the model’s classification accuracy is the highest, and the computational complexity is relatively moderate. Therefore, in the experiments described in our manuscript, we adopted a segmentation parameter of 1024 points and an overlap ratio of 50%.

Experiment two

(1) Description of dataset II

To further validate the effectiveness of the proposed model, Experiment Two employs the SEU_gearbox dataset from Southeast University for experimental validation, with the experimental platform illustrated in Fig. 12. The SEU_gearbox dataset was collected from the transmission system dynamic simulator under two different operating conditions: one with the speed system load set at 20 Hz–0 V, speed at 1200 rpm, and load at 0 N × m; and the other with the speed system load set at 30 Hz–2 V, speed at 1800 rpm, and load at 10.97 N × m.

This dataset includes four gear fault conditions and one healthy condition (Normal). The gear fault conditions are: crack at the bottom of the gear (Chipped), missing one tooth (Miss), crack at the root of the gear (Root), and wear on the surface of the gear (Surface). The specific experimental data are shown in Table 4.

(2) Diagnosis results and analysis of dataset II.

- Precision, recall, F1 score, and accuracy metrics of the model

Precision, recall, F1 score, and accuracy were selected as evaluation metrics to analyze the diagnostic performance of the model. These metrics are crucial references for assessing the performance of an AI model. As shown in Table 5, the proposed model can achieve up to 100% in all these metrics, which, to some extent, indicates that the model possesses excellent diagnostic classification performance.

- Scatter plot of diagnostic prediction vs. true values

Label	Load	Status	Training set	Testing set	sample lengths
0	20_0/30_2	Chipped	70	30	1024
1	20_0/30_2	Miss	70	30	1024
2	20_0/30_2	Root	70	30	1024
3	20_0/30_2	Surface	70	30	1024
4	20_0/30_2	Normal	70	30	1024

Table 4. Data description of the test gearbox.

Status	Precision	Recall	F1-score	Support
Chipped	1.00	1.00	1.00	10
Miss	1.00	1.00	1.00	10
Root	1.00	1.00	1.00	10
Surface	1.00	1.00	1.00	10
Normal	1.00	1.00	1.00	10
Accuracy	–	–	1.00	50
Macro avg	1.00	1.00	1.00	50
Weighted avg	1.00	1.00	1.00	50

Table 5. Diagnostic performance of the proposed model.

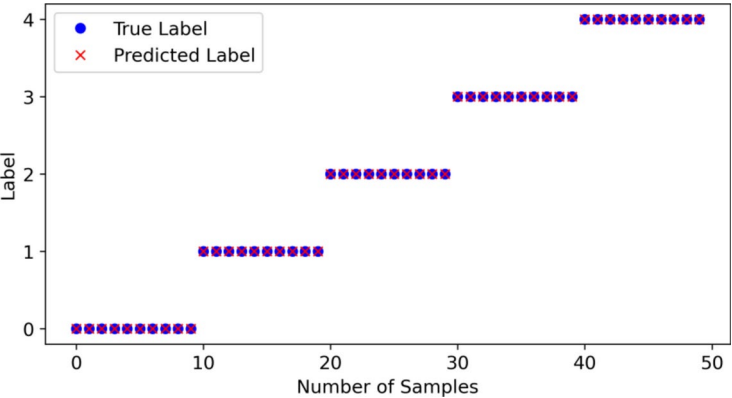


Fig. 13. Scatter plot of diagnostic predicted values versus actual values.

The experimental validation results are shown in Fig. 13, where the model’s recognition and prediction outcomes for various health statuses are consistent with the actual health statuses, and all predictions are correct. This once again demonstrates the superior diagnostic classification capability and stability of the proposed model.

• Confusion matrix analysis

Furthermore, to further illustrate the effectiveness of FCNVT, a confusion matrix analysis was conducted on one of the test results from this dataset, as shown in Fig. 14. The proposed method accurately predicted all three health statuses of the mechanical equipment, achieving a maximum diagnostic accuracy of 100% in the final diagnostic results, which aligns with the findings in Fig. 13. This shows that the FCNVT model has excellent fault diagnosis classification accuracy.

• t-SNE analysis

To verify the feature extraction capability of the proposed method, the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm was employed to visualize the features of the input layer and output layer, characterizing the model’s feature learning ability. As shown in Fig. 15. In Fig. 15a, it can be observed that the original input data of various types are mixed together in the two-dimensional space, with a high degree of confusion. From Fig. 15b, after feature extraction and learning of the input signal data by the model, different health status data are well grouped during the learning process. The intra-class compactness and inter-class separability of the health status features at the output layer have significantly improved, with the inter-class distances between different health statuses becoming clear and wide, and the data within the same class

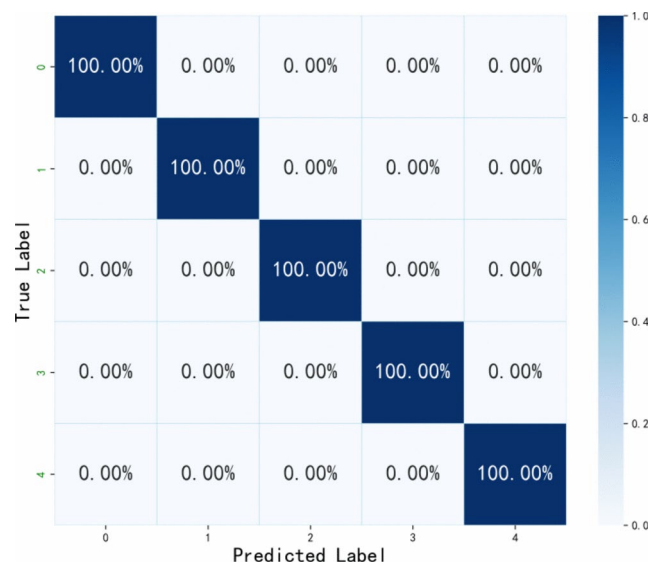


Fig. 14. Confusion matrix.

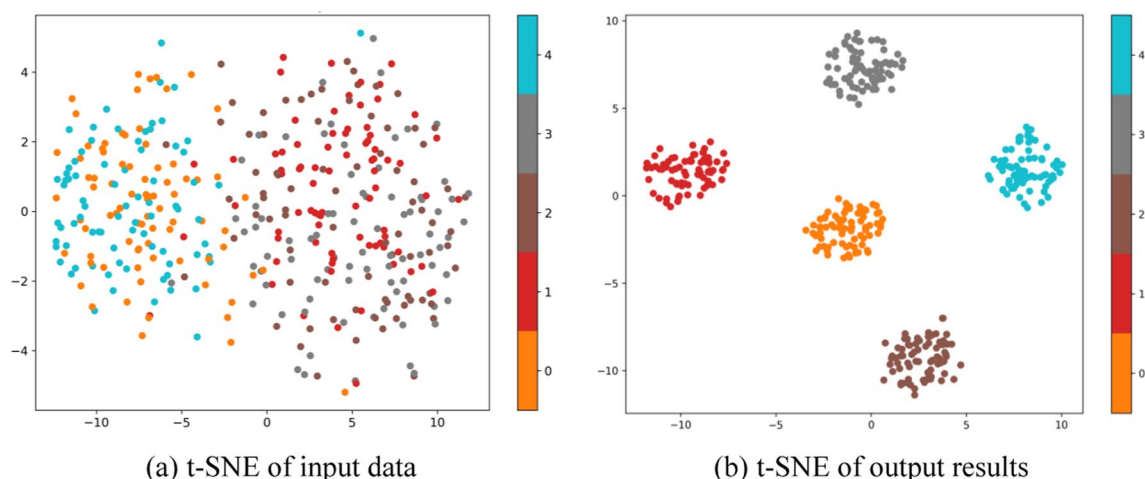


Fig. 15. Feature visualization.

becoming more compact, while the confusion between different features has disappeared. This also indicates that the method proposed in this paper can effectively extract and learn the features of different health statuses from bearing data and accurately distinguish the corresponding fault types.

Comparison of model parameters and computational costs

As shown in Table 6, among the compared models, the proposed method has the smallest number of parameters and computational cost, with the computational cost being only 0.28 G. In contrast, the traditional ViT has the largest number of parameters and computational cost, significantly higher than those of other models. Improved Transformer models such as Swin, ShuffleT, and Cswin, compared to the ViT, exhibit better performance, lower computational costs, and fewer parameters. Conventional Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have notably lower numbers of parameters and computational costs compared to models like ViT and Swin. The aforementioned comparative analysis demonstrates that the proposed method has a significant advantage in terms of both the number of parameters and computational cost.

Comparison of diagnostic accuracy with state-of-the-art algorithm models

Diagnostic accuracy is a core metric in the process of mechanical equipment fault diagnosis. As shown in Fig. 16, we compare the diagnostic accuracy results of the FCNVT model proposed in this paper with other state-of-the-art algorithm models. These algorithms include CNN, LSTM, CNN-LSTM, Dconformer³⁰, TST³¹, Diagnosisformer³², and TFT¹⁰.

Models	Parameters/M	Computational Costs/G
ViT ¹⁶	86	55.4
Swin ¹⁷	29	4.5
ShuffleT ¹⁸	28	4.5
CSwin ¹⁹	23	4.3
CNN	13.4	1.6
RNN	18.0	3.2
Proposed method	7	0.28

Table 6. Parameter counts and computational costs of various models.

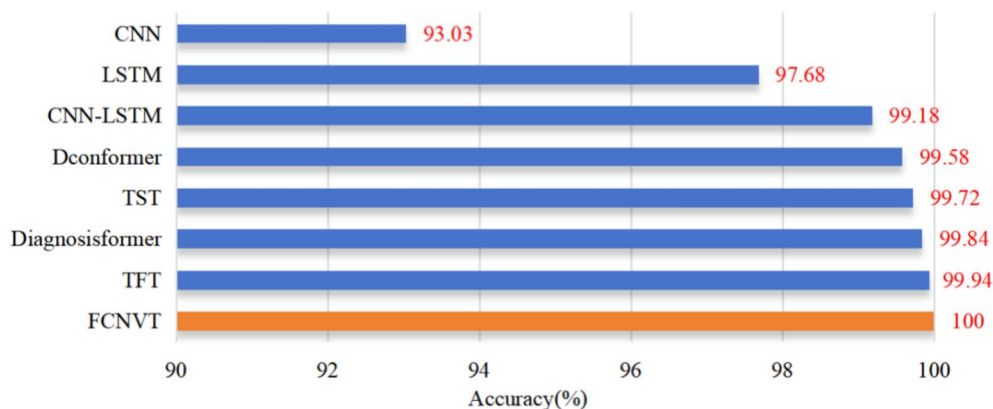


Fig. 16. Comparison of diagnostic accuracy with other state-of-the-art algorithms.

From Fig. 16, it can be observed that among the seven algorithms compared with the proposed model, the TFT model has the highest diagnostic accuracy at 99.94%, while CNN has the lowest diagnostic accuracy at 93.03%. The FCNVT model proposed in this paper has the highest diagnostic accuracy among all models, improving by 6.97% over CNN and still showing an enhancement compared to TFT. Therefore, we can conclude that the proposed model is better at extracting and learning the characteristic information of mechanical equipment faults, giving it a higher diagnostic accuracy and better classification capability.

Development and application of graphical user interface

Graphical user interface (GUI)

A GUI (Graphical User Interface) is a type of user interface that allows users to interact with electronic devices (such as computers, smartphones, tablets, etc.) through graphical elements like icons, buttons, windows, etc. Unlike command-line interfaces (CLIs) that rely solely on text commands, GUIs are more intuitive and user-friendly, enabling users to control and utilize software through visual cues and mouse clicks. This significantly reduces the learning and usage costs for non-professionals, thereby accelerating the dissemination and application of related technologies.

Therefore, based on the fault diagnosis method proposed in this paper, we have developed a GUI-based mechanical equipment fault detection system to promote the practical implementation and information management of intelligent fault diagnosis for mechanical equipment.

GUI-based mechanical equipment fault detection system

- Login interface of the diagnostic system

To ensure data security and manage permissions, the system requires entering an account and password for login. The login interface is shown in Fig. 17.

- Main interface of the diagnostic system

Upon entering the correct account and password, the user logs into the main interface of the mechanical equipment fault diagnosis system, as shown in Fig. 18. The main interface of this system includes functions for data access and reading, intelligent diagnosis, and monitoring of diagnostic results. Below, we will introduce these three functional areas and their implementation methods.

- (1) Data access and reading: The system provides three modes for data access and reading. The first mode is selecting a single piece of data for a single fault diagnosis; the second mode involves accessing and reading

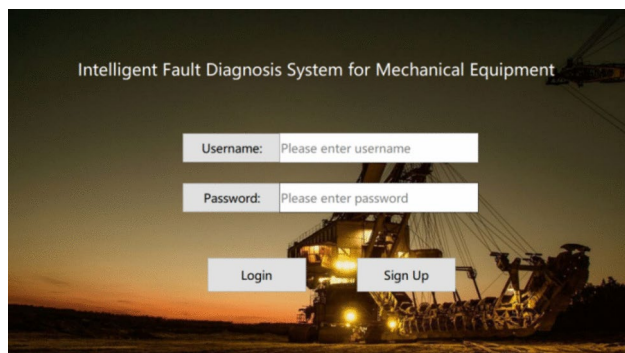


Fig. 17. Login interface of the diagnostic system.

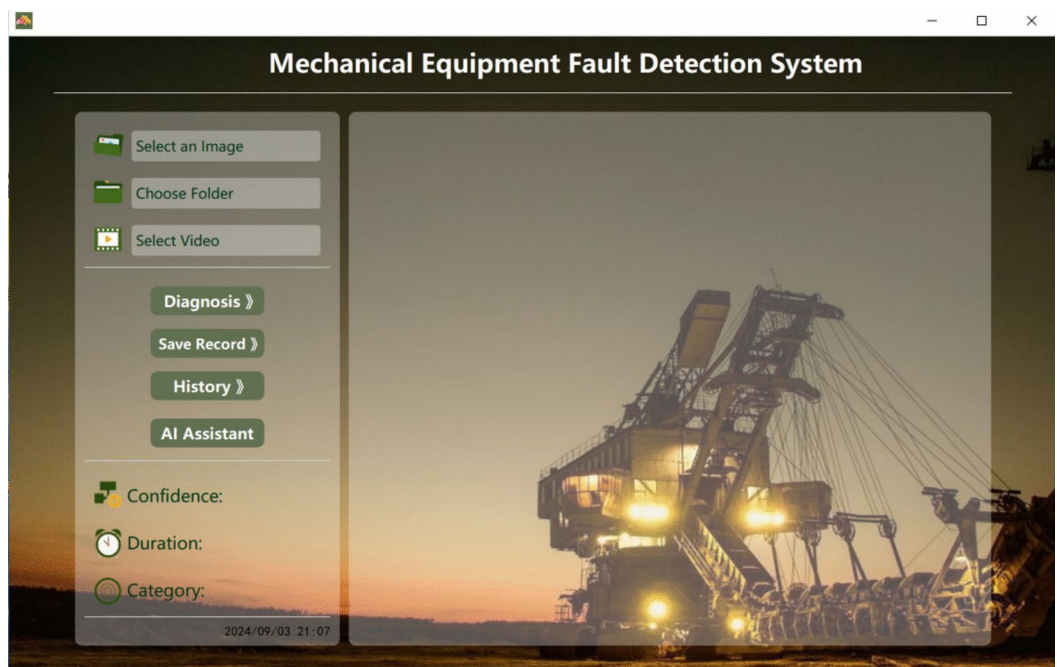


Fig. 18. Main interface of the diagnostic system.

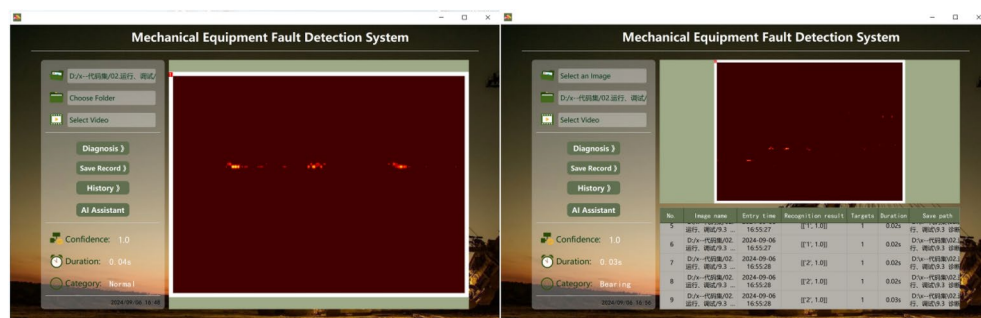
data in folder form to achieve batch detection and identification of faults; the third mode is accessing and reading data in video file form, which allows for dynamic fault detection and thus possesses certain object detection capabilities.

- (2) Intelligent diagnosis and monitoring of diagnostic results: Once the data is loaded into the system, fault diagnosis and recognition are performed. As shown in Fig. 19a, the image data after loading is displayed on the right, with the diagnostic results, duration, and confidence level shown in the lower-left corner. As shown in Fig. 19b, the system can record the detected data and allow viewing of historical diagnostic results individually. Additionally, the system offers the export of recorded diagnostic results, facilitating the information management of diagnostic outcomes. To improve work efficiency, we have integrated an AI assistant function into the system, as shown in Fig. 19c. It leverages powerful AI models like ChatGPT and Kimi to provide assistance.

Conclusion

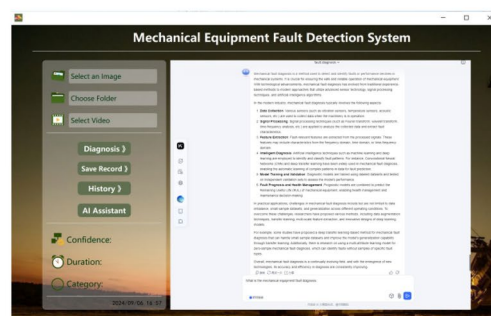
This study proposes a lightweight and efficient intelligent fault diagnosis method (FCNVT) that integrates the characteristics of convolutional networks and vision transformers, and successfully develops a GUI-based mechanical equipment fault detection system. The main conclusions of this paper are as follows:

- (1) The FCNVT model is proposed for the fault diagnosis of mechanical equipment. It combines convolutional layers and Transformer layers, capturing both local and global features while maintaining efficiency. The model utilizes local connection patterns and sparse attention mechanisms to reduce computational costs and the number of parameters.



(a) Diagnostic Function Interface

(b) Historical Record Function Interface



(c) AI Assistant Function Interface

Fig. 19. Demonstration of various functional interfaces of the diagnostic system.

- (2) The FCNVT model achieves the fusion of multi-scale features through multi-layer convolution operations and multi-head self-attention mechanisms. The convolutional layers are responsible for extracting local features, while the self-attention mechanism captures global dependencies, all while maintaining lower computational costs. This design of multi-scale feature fusion enables the model to more comprehensively understand and recognize complex fault patterns.
- (3) The original signal data's information is fully utilized and the randomness of signals when faults occur is simulated through the use of random overlapping segmentation technology. The synchrosqueezed wavelet transform (SWT) is adopted to improve the time–frequency resolution of images and reduce redundant information, thereby enhancing the model's diagnostic capability and generalization performance.
- (4) Based on this method, a GUI-based mechanical equipment fault detection system has been developed. This system facilitates the practical implementation and information management of intelligent fault diagnosis for mechanical equipment.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 15 September 2024; Accepted: 29 January 2025

Published online: 28 February 2025

References

1. Yaguo, L. et al. Opportunities and challenges of intelligent fault diagnosis for machinery under big data. *J. Mech. Eng.* **54**(05), 94–104 (2018) (in Chinese).
2. Lei, Y. et al. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **138**, 106587 (2020).
3. Zhao, R. et al. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **115**, 213–237 (2019).
4. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
5. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
6. Dong, S., Wang, P. & Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **40**, 100379 (2021).
7. Zhang, Y. et al. Fault diagnosis of rotating machinery based on recurrent neural networks. *Measurement* **171**, 108774 (2021).
8. Guan, Y. et al. Rolling bearing fault diagnosis based on information fusion and parallel lightweight convolutional network. *J. Manuf. Syst.* **65**, 811–821 (2022).
9. Hao, S. et al. Multisensor bearing fault diagnosis based on one-dimensional convolutional long short-term memory networks. *Measurement* **159**, 107802 (2020).
10. Ding, Y. et al. A novel time–frequency transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings. *Mech. Syst. Signal Process.* **168**, 108616 (2022).
11. Liang, P. et al. Fault transfer diagnosis of rolling bearings across multiple working conditions via subdomain adaptation and improved vision transformer network. *Adv. Eng. Inform.* **57**, 102075 (2023).

12. Jia, L. et al. Multiscale residual attention convolutional neural network for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **71**, 1–13 (2022).
13. Li, X., Zhang, W. & Ding, Q. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Process.* **161**(8), 136–154 (2019).
14. Long, Z. et al. Motorfault diagnosis using attention mechanism and improved AD boost driven by multi-sensor information. *Measurement* **170**(1), 108718 (2021).
15. Vaswani, A. Attention is all you need. arxiv preprint <https://arxiv.org/abs/1706.03762> (2017).
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A. et al. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
17. Liu, Z., Lin, Y., Cao, Y. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022 (2021).
18. Huang, Z., Ben, Y., Luo, G. et al. Shuffle transformer: Rethinking spatial shuffle for vision transformer. arxiv preprint <https://arxiv.org/abs/2106.03650> (2021).
19. Dong, X., Bao, J., Chen, D. et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12124–12134 (2022).
20. Fang, H. et al. CLFormer: A lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions. *IEEE Trans. Instrum. Meas.* **71**, 1–8 (2021).
21. Yonglin, T. et al. key issues in vision transformer research: Current status and prospects. *Acta Autom. Sin.* **48**(4), 957–979 (2022) (in Chinese).
22. Li, Z. et al. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw. Learn. Syst.* **33**(12), 6999–7019 (2021).
23. Wang, X., Girshick, R., Gupta, A. et al. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803 (2018).
24. Kolesnikov, A., Beyer, L., Zhai, X. et al. Big transfer (bit): General visual representation learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 491–507 (Springer International Publishing, 2020).
25. Touvron, H., Vedaldi, A., Douze, M. et al. Fixing the train-test resolution discrepancy. In *Advances in Neural Information Processing Systems*, Vol. 32 (2019).
26. Xie, Q., Luong, M. T., Hovy, E. et al. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698 (2020).
27. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. arxiv preprint <https://arxiv.org/abs/1607.06450> (2016).
28. Wen, J., Gao, H., Li, S. et al. Fault diagnosis of ball bearings using synchro squeezed wavelet transforms and SVM. In *2015 Prognostics and System Health Management Conference (PHM)* 1–6 (IEEE, 2015).
29. Daubechies, I., Lu, J. & Wu, H. T. Synchro squeezed wavelet transforms: An empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal.* **30**(2), 243–261 (2011).
30. Li, S. et al. Dconformer: A denoising convolutional transformer with joint learning strategy for intelligent diagnosis of bearing faults. *Mech. Syst. Signal Process.* **210**, 111142 (2024).
31. Jin, Y., Hou, L. & Chen, Y. A time series transformer based method for the rotating machinery fault diagnosis. *Neurocomputing* **494**, 379–395 (2022).
32. Hou, Y. et al. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer. *Eng. Appl. Artif. Intell.* **124**, 106507 (2023).

Author contributions

C.M and K.H: Writing—original draft, Writing—review and editing, Formal analysis, Data curation, image drawing. H.J: Writing—review and editing. W.L and K.X: Data curation, image editing. All authors reviewed the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China under Grant no. KZ1410067 and the Zhejiang Provincial Science and Technology Project under Grant no. KZS2101030.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.M. or K.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.