Early Anomaly Detection of Offshore Wind Turbine Gearbox Based on SLFormer Neural Network

Zekun Wang [a, b], Xue Jiang [c], Zifei Xu [d, e, *], Chang Cai [b], Xiaodong Wang [a], Jianzhong Xu [a, b], Xiaohui Zhong [b], Wei Yang [f], Qing 'an Li [b, *]

[a] *School of Energy Power and Mechanical Engineering, North China Electric Power University, Beijing 102206, P. R. China*

[b] *Institute of Engineering Thermophysics, Chinese Academy of Sciences, Beijing 100190, P. R. China*

[c] *Department of Naval Architecture and Marine Engineering, Zhejiang Ocean University, Zhoushan 316022, P. R. China*

[d] *School of Engineering, Liverpool John Moores University, Liverpool, L3 3AF, United Kingdom*

[e] *State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai, 200240, P. R. China*

[f] *CSSC Haizhuang Windpower Co., Ltd., Chongqing 401122, P. R. China*

Corresponding author. *E-mail address:* Z.Xu@ljmu.ac.uk (Zifei Xu), liqingan@iet.cn (Qing 'an Li)

Keywords**:** Offshore Wind Turbine, SLFormer, Early Anomaly Detection, Gearbox, SCADA

Abstract

Offshore wind turbines (OWTs) have become a leading strategic equipment for the large-scale development of deep offshore wind energy. Developing efficient fault detection methods is crucial for ensuring the reliability of OWTs. However, the widely used fault detection techniques are mostly based on traditional deep learning frameworks, which are limited in handling long-term dependencies, leading to constraints in global feature extraction. Additionally, current studies often build models based on the operational performance of an individual wind turbine, limiting the generalizability of the models. Therefore, this study introduces an improved model called SLFormer, which integrates Long Short-Term Memory into the Transformer encoder for early fault detection in the gearbox. Simultaneously, a novel fault detection strategy based on SCADA data analysis and transfer learning is proposed. Case studies from three wind farms indicate that the SLFormer model significantly outperforms six other popular prediction models in terms of stability and accuracy in modeling normal behavior and exhibits high robustness to random disturbances in SCADA data. The SLFormer model can predict gearbox anomalies twenty-one days in advance and effectively avoids false fault reports. The proposed strategy can help the model acquire fault knowledge from multiple wind farms, thus creating a framework with generalizability.

1 Introduction

Wind energy, as a clean and renewable energy source, is attracting more and more attention from countries around the world due to its low pollutant emissions and abundant resources. At present, the development of offshore wind power is considered to be an effective means to increase the utilization rate of wind energy as well as to reduce the cost of electricity. According to the Global Wind Energy Council (GWEC), the total installed capacity of OWTs in the world has reached 64.3 GW by 2022 [1]. Compared to 2021, installed capacity has increased by 12.4%. However, the rapid development of wind power generation has also brought about the problem of excessive failure rates. The operation and maintenance costs of OWTs account for about 23% of all wind farm revenues [2]. According to the research of Ref. [3], gearbox failure is the main cause of excessive OWTs downtime. The gearbox is the core equipment for the operation of OWTs. Among them, bearings, as its key components, play a vital role in the stable operation of OWTs. However, due to the instability of wind and long-term operation of OWTs, the bearings need to withstand complex and alternating loads and stresses [4]. In particular, the large size of OWTs makes gearbox bearings often operate under high loads. In addition, OWTs need to face salt spray and high humidity, which exacerbate the corrosion and wear of bearings. These factors will lead to damage, fatigue and other problems in OWT bearings, which will lead to failures. Li *et al.* [5] showed that bearing failures account for 70% of the total OWT gearbox failures. Bearing failures can seriously affect the performance and lifetime of wind turbines (WTs). Once a failure occurs, it may lead to turbine downtime or even damage, affecting the operational efficiency of the whole wind farm. Therefore, fault detection of OWT gearbox bearings is of great practical significance. Effective early fault detection can prevent accidents from occurring, reduce maintenance costs, and improve the operational efficiency and reliability of WTs [6].

Earlier fault detection techniques have mainly used the development of mathematical models close to the WT operating conditions to predict future behavior. This approach is usually based on a solid theoretical foundation, so that the results have good interpretability [7]. However, it is challenging to construct accurate mathematical models for complex systems. Especially for WT, the operation is difficult to be described using physical models. In addition, the presence of unknown inputs leads to the poor adaptability of the mathematical model-based approach to the environment with multiple operating conditions.

With the development of computer technology, the huge amount of data generated during the operation of WT has received great attention from researchers. Among them, as a standard configuration of WT operation, the Supervisory Control and Data Acquisition (SCADA) system contains a large number of parameter laws that have not been fully explored. For newer wind farms, the vibration data collected by the Condition Monitoring System (CMS) can also be used as the main target for analysis. This data-based analysis can predict WT bearing failures in real time, reducing the dependence on specialized knowledge and improving the efficiency and reliability of wind power generation. Early research on data-driven fault detection was dominated by approaches combining signal processing with traditional machine learning algorithms. These approaches have shown effectiveness in different scenarios. Shanbr *et al.* [8] addressed the problem of signal mixing by using envelope analysis after spectral kurtosis to provide early fault detection. In this relatively early study, they did not use machine learning techniques. Similarly, Ref. [9] explored the problem of WT bearing fault detection under non-stationary conditions. They combined variational modal decomposition (VMD) with blind-source separation to form a set of bearing fault detection methods that can be adapted to WT's drastically variable operating conditions. They are also using only signal processing techniques. However, this route has limited adaptability and flexibility, and feature extraction is not comprehensive enough. So, some researchers have chosen to combine it with machine learning. Saari *et al.* [10] used fault-specific features extracted from vibration signals to detect WT bearing faults. They developed detection models with different sensitivities based on SVM. The results showed that their proposed model detects faults earlier than conventional methods. Yan *et al.* [11] used VMD to divide the raw bearing temperature data into multiple subsequences. Subsequently, stacked autoencoder and group method of data handling were combined to accomplish the bearing temperature prediction task. Ref. [12] proposed a new monitoring method considering the effect of ambient temperature on bearings. They input ambient temperature variables into the model and combined swarm intelligence optimization and extreme learning machine methods to predict bearing temperature failures. In addition to this, wavelet denoising methods [13], artificial neural networks [14], energy entropy [15] and other techniques have been applied to the task of fault detection in WT bearings. However, these fault detection methods based on traditional machine learning or signal processing generally have limitations such as low prediction accuracy and poor convergence. These problems seriously restrict the application of these methods in wind farms. Early research on data-driven fault detection was dominated by approaches combining signal processing with traditional machine learning algorithms. These approaches have shown effectiveness in different scenarios.

Unlike the above methods, Deep Learning (DL) techniques have more powerful data mining capabilities. For the complex and nonlinear problem of WT fault detection, DL obtains deeper data characteristics by stacking more layers of networks. Moreover, DL has the ability to construct and learn features automatically, which reduces the workload of feature engineering. This is more applicable to the WT online fault detection problem. Ref. [16] proposed a health detection framework based on clustering and deep belief network by considering the multiple operating conditions characteristic of WT operation. They used DL to model the normal behavior of the WT and captured the nonlinear correlation between different features. Ref. [17] used data from normal operation of WT to accomplish the task of bearing fault detection in real wind farms by building a gated recurrent unit neural network. Since their proposed method does not require the involvement of fault data, it can be applied to old wind farms. Yang *et al.* [18] discussed the vibration mechanism and fault characteristics of WT bearing faults and summarized different bearing fault detection models. Besides, they applied deep neural network algorithm to the bearing fault detection task. The results showed that their proposed method can effectively detect the early failure of WT bearings. Zhang *et al.* [19] proposed a cascade deep learning method for WT bearing fault detection. Their proposed model based on the combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) is better adapted to the main bearing fault detection task than the traditional LSTM. Ref. [20] performed semi-supervised learning based on SCADA data during normal WT operation to avoid the problem of data imbalance. They established a prediction metric based on the mean square error, showing adaptability to different seasons and operating behaviors. The results show that their proposed method can warn bearing failures months in advance. Wu *et al.* [21] proposed a WT bearing fault detection method that combines Kullback-Leibler divergence and LSTM. Different from the above researchers' methods, they categorized the abnormal states of WTs into alarms and faults, and regarded the alarms as early warnings. Ref. [22] proposed a WT bearing condition detection method with built-in visualization, taking into account the interpretability of the DL model. The case study showed that this method can warn bearing failures in advance and effectively improve the credibility of the DL model. In summary, DL techniques have been widely used in WT fault prediction. However, for long time series such as SCADA data, it is difficult to deal with the problem of long-distance dependency using traditional DL models alone. In addition to that, most of the current studies are unfolded based on Recurrent Neural Network (RNN), which cannot use parallel computing on large-scale datasets, resulting in low computational efficiency.

In order to solve the above problem, researchers have proposed a technique called attention mechanism. The DL algorithm based on the attention mechanism focuses on the parts that are more important to the current task when parsing the input information. Therefore, the attention mechanism can utilize the long-range information in SCADA data more effectively. Yan *et al.* [23] developed a Gated Recurrent Unit (GRU) network with a self-attention mechanism for the problem of early fault detection in WT main bearings. They used this method to learn the data during normal operation of the WT and mine the nonlinear correlation between features in multiple states. In addition to this, they incorporated a changepoint detection method to automatically identify fault states. Ref. [24] proposed an attention mechanism combined with CNN-LSTM network for fault detection in WT. In their proposed method, the attention mechanism is used to enhance the effect of important information on the results, thus improving the prediction accuracy. Yang *et al.* [25] used a model combining a two-stage attention mechanism with an RNN for learning WT normal behavior. In addition to that, they designed an input information-based attention mechanism for assigning the importance of different variables. The results show that their proposed method can effectively detect WT bearing faults. Therefore, the above studies can demonstrate the effectiveness of attention mechanism in deep learning. However, despite the efforts of these studies to integrate attention mechanism into existing DL architectures by incorporating them as add-ons, they do not essentially overcome the original limitations of these classical models completely.

As a result, researchers have begun to look for new models that can utilize the attention mechanism at a deeper level than just as an add-on module. This thinking led to the emergence of a new model based entirely on the attention mechanism - Transformer. The original Transformer was designed for sequence prediction problems such as machine translation, and demonstrated superior performance across a variety of tasks. However, for the fault prediction task of WT, Transformer do not explicitly model the continuity and periodicity of time series. Therefore, they are usually combined with traditional DL algorithms for fault prediction. Ref. [26] used a combination of Resnet and Transformer to extract features of bearing faults. The results showed that the method is effective in predicting bearing faults and is adapted to high-noise environments. Jiao *et al.* [27] embedded a binary tree filter into Transformer to obtain a better feature representation. The results showed that their proposed method can effectively detect bearing faults.

In summary, most of the current research focuses on applying algorithms for modeling the normal behavior of WTs and learning the correlation between features by the model for fault detection. This approach has been proven to be effective by researchers, but its accuracy for normal data prediction still needs to be improved [28]. Low accuracy will lead to false and missed alarms of faults, which will affect the economy of wind farms. On the other hand, most of the existing fault detection methods mainly focus on the relationship between variables at the same point in time, while the self-correlation and inter-correlation of the features are not well attended to. These methods need further improvements to fully utilize these important properties of time series data. Besides, how to maximize the advantages of traditional deep learning models and Transformer models is also a question worth exploring. Therefore, it is necessary to propose a new fault detection model to address the above challenges.

Moreover, obtaining operational data from OWTs, especially fault data, poses a significant challenge due to the usual confidentiality surrounding these data. To address this issue, some researchers have opted to use simulated data for detecting faults in OWTs [29, 30]. However, simulated data may not fully replicate the complex environments and conditions present during actual faults. This discrepancy can lead to decreased performance of detection models when confronted with real faults. Therefore, leveraging the vast of actual amounts operational data from onshore wind turbines to establish a universal fault detection framework becomes a viable solution to this problem [31]. Admittedly, OWTs are subject to greater loads, leading to stronger component vibrations and correspondingly higher temperatures. However, the gearbox shaft systems of both types of wind turbines are usually similar, which means the fault characteristics are alike [3, 5]. The advantage of deep learning algorithms lies in their ability to capture potential fault patterns within the data. In conclusion, by training deep learning models with the abundant actual operational data from onshore wind turbines, it is possible to learn and identify these common gearbox fault characteristics, and then by devising detection strategies, effectively detect gearbox faults in OWTs.

Consequently, this paper proposes a novel transfer learning-based online detection process aimed at the early detection of anomalies in WT gearbox bearings. Firstly, this paper utilizes overlapping sliding windows to segment the multidimensional features in SCADA data into uniform-length samples, thereby capturing both self-correlation and inter-correlation among each feature. Subsequently, we proposed a SLFormer model, specifically designed for time series data analysis. The SLFormer model integrates LSTM into the Transformer's computation process, enhancing its ability to capture the temporal dependencies within the input data. Our experimental results demonstrated that the SLFormer outperforms six other widely-used models in accurately modelling the normal operational behavior of WTs. Lastly, the fault detection performance of the proposed model is validated using SCADA data from two other wind farms. The results underscored that the SLFormer can effectively identify early gearbox faults while avoiding false alarms.

The main contributions of this paper are as follows:

a) This paper completes the online detection of early anomalies in WT gearbox bearings by using limited SCADA data features. The sliding window sampling retains the self-correlation and inter-correlation among each feature.

b) A SLFormer model suitable for fault prediction is proposed. This model integrates the advantages of LSTM and Transformer to fully exploit the deep-seated features of the input information.

c) A transfer-learning-based process for online detection of early anomalies in WT bearings is proposed, which provides an effective solution for wind farm applications. Specifically, the knowledge learned from the normal operational data of the WT is migrated to the new WT to detect potential faults.

This study is organized as follows: Section II describes the mathematical principles in the framework proposed in this paper. Section III describes the overall detection process. Section IV carries out three case studies, mainly including normal behavior modeling, early anomaly detection and false alarm verification. The last section summarizes the whole paper.

2 Proposed SLFormer Architecture

The complex environment faced during WT operation poses a great challenge to effective fault detection methods, which makes it difficult to adapt the traditional Transformer Encoder. Therefore, in this paper, we use LSTM to improve the original Transformer, combined with Overlapping Sliding Window Sampling to compose a novel model architecture for early anomaly detection, as shown in Fig. 1. The theoretical foundations in the proposed method are described in this chapter.

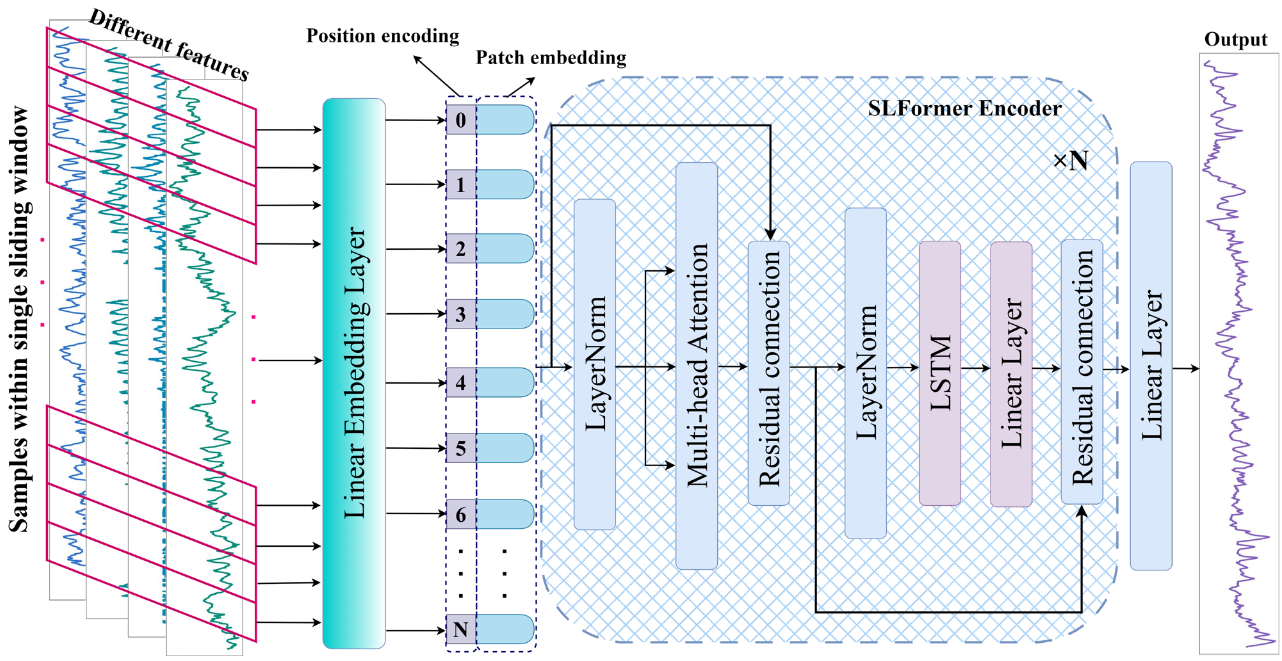

Fig. 1. The framework of the proposed method.

2.1 Overlapping Sliding Window Sampling (OSWS)

Early failure detection of WT bearings based on SCADA data requires mining the trends and patterns of the data over time. In practical scenarios, relying only on data from a single point in time may not be sufficient to fully reveal the dynamic characteristics of features over time. Therefore, this paper adopts the OSWS method, which can retain both self-correlation and intercorrelation of SCADA data, as shown in Fig. 2. The OSWS method slices the SCADA data by setting the window length and sliding step size. Here, the window length S defines the number of time points in each sample, while the sliding step W_s denotes the distance the window moves over the time series. In this paper, the sliding step is set to be smaller than the window length. In this case, the center sample of each window will be able to contain the information of the samples before and after it at the same time, thus reflecting the trend and periodicity of the data in a more comprehensive way. It also increases the diversity of samples.

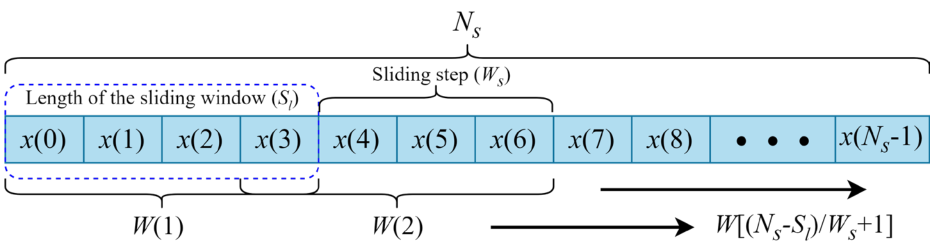

Fig. 2. The sampling process of OSWS method.

For a time series X of length N_s :

$$X = \left[x(0), x(1), x(2), ..., x(N_s - 2), x(N_s - 1)\right] \tag{1}$$

The data $W(i)$ in the ith window is:

$$W(i) = \left[x\left[(i-1)W_s\right], ..., x\left[S + (i-1)W_s\right]\right] \tag{2}$$

Thus, the total number of windows N_w can be expressed as:

$$N_w = (N_s - S) / W_s + 1 \tag{3}$$

In this paper, each feature is divided into samples of length S to be processed by the above operation.

2.2 Linear Embedding

The input to SLFormer is defined as multiple tokens. K features from SCADA data are selected as input. The sample of length S after processing above is:

$$X_0 = (x_1, x_2, x_3, x_4, ..., x_{S-1}, x_S) \in \mathbb{R}^{K \times S} \tag{4}$$

where each x_r at time step r is a vector of dimension K. Specifically, the jth feature in a sample of length S with each time index starting at 1 can be expressed as:

$$x_{1:S}^{(j)} = \left(x_1^{(j)}, x_2^{(j)}, x_3^{(j)} ..., x_{S-1}^{(j)}, x_S^{(j)}\right) \tag{5}$$

Where j=1, 2, 3, …, K.

Before the Linear Embedding operation, the input multidimensional features need to be split evenly into patches of length L_s, each labeled $z_s^i \left(i = 1, 2, 3, ..., N_s\right)$. The divided tensor Z is:

$$Z = \left[z_s^1, z_s^2, z_s^3, ..., z_s^{N_s}\right] \in \mathbb{R}^{B_s \times L_s \times N_s \times K} \tag{6}$$

where N_s is the number of patches with a value equal to S/L_s and B_s is the batch size.

Subsequently, each patch is mapped to the D_m dimension by a linear transformation:

$$Z^* = \left[z_s^1, z_s^2, z_s^3, ..., z_s^{N_s}\right] \cdot W_s \in \mathbb{R}^{B_s \times N_s \times K \times D_m} \tag{7}$$

where W_s is a trainable linear mapping matrix. Notably, all tokens share the same linear mapping matrix. Linear embedding can effectively extract important features from the data and enhance the SLFormer model's understanding of the data. By mapping the original data into a new space, the model can better learn the relationships and patterns between these features [32].

2.3 Position Encoding

Since the SLFormer model adopts an approach that decomposes the sequence into multiple segments for parallel training, this approach may lead to the loss of position information of each sub-sequence. Therefore, in order to preserve this important positional information, a trainable Position Encoding is introduced in this paper. This design ensures that the model is able to understand not only the content of each element, but also their respective positions in the overall sequence when processing the sequence data [33]. Thus, the sequence can be redefined:

$$Z^* = \left[z_s^1, z_s^2, z_s^3, ..., z_s^{N_s}\right] \cdot W_s + Z_p \in \mathbb{R}^{B_s \times N_s \times K \times D_m} \tag{8}$$

There are two main forms of positional coding: one-dimensional (1D) and two-dimensional (2D).1D positional coding considers only relative positions in the sequence, while 2D positional coding takes into account the two-dimensional spatial structure of the data such as images. Since SCADA data is a linear sequence over time, its intrinsic structure dictates the use of 1D positional coding to capture the temporal information of the features. This one-dimensional coding approach allows the model to capture the temporal correlation between data points, which is critical for predicting future system states and identifying trends.

2.4 SLFormer Encoder Layer

By stacking the SLFormer Encoder, it allows the model to learn complex and abstract features and representations at deeper levels. At each layer, the model learns and extracts different features and concepts, resulting in a hierarchical understanding. This layer-by-layer learning approach better helps the model understand the data and thus make more accurate predictions.

2.4.1 Layer Normalization

Layer normalization (LN) makes the model less sensitive to parameter initialization as it reduces layer to layer dependency by normalizing the inputs. In addition to this, LN helps to improve the flow of gradients and reduces the problem of vanishing or exploding gradients, thus improving training efficiency [34]. Compared with Batch Normalization, LN does not depend on batch size and sequence length. Therefore, in this paper, we used LN to normalize the input data with the following formula:

$$\begin{cases} \mu_s = \sum_{i=1}^{D_m} Z_i^* \Big/ D_m \\ \sigma_s = \sqrt{\sum_{i=1}^{D_m} \left(Z_i^* - \mu_s\right)^2 \Big/ D_m + c} \\ LN\left(Z^*; \lambda, \eta\right) = \left[\left(Z^* - \mu_s\right)/\sigma_s\right] \odot \lambda + \eta \end{cases} \tag{9}$$

where $\odot$ is the Hadamard product. λ and η are the scale and bias of the LN, respectively. c is a smaller constant used to improve the stability of the calculation.

2.4.2 Multi-Head Self-Attention

Selective attention in the attention mechanism helps the model to learn the features in the data more efficiently. In particular, as a representative study of the attention mechanism, the self-attention mechanism allows each element in the data to be attended to all other elements in the data. This feature allows the self-attention mechanism to capture the global dependencies in the data, thus dramatically improving the expressive ability of the model [35]. The three components Query(Q), Key(K), Value(V) are used in the self-attention mechanism to compute the attention. Where Q represents the data at a specific point in time. K is a representation of all time points in the features. Each time point has a corresponding key vector. V is an alternative representation of each time point in the time series, usually generated along with the key. Scaled dot-product is used in the self-attention mechanism to calculate and assign attention scores, as shown in Fig. 3. In this paper, each token is linearly transformed to obtain three vector matrices Q_m, K_m, V_m. i.e:

$$Queries: \quad Q_m = Z_s W_m^q \tag{10}$$

$$Keys: \quad K_m = Z_s W_m^k \tag{11}$$

$$Values: \quad V_m = Z_s W_m^v \tag{12}$$

where W_m^q、W_m^k、W_m^v are the parameters that can be trained. The dimensions of both Q_m and K_m are d_k. In single-head scaled dot-product attention, the dot product of all Queries and Keys is computed to obtain the attention weight matrix. In order to maintain the stability of the gradient, each Key needs to be divided by the scaling factor $\sqrt{d_k}$. Dividing by $\sqrt{d_k}$ provides some assurance that the resulting dot product values are within a reasonable range, thus making the model training more robust. The output is then normalized using the softmax function so that all the attentional weights add up to 1. This process is shown in Equation 13.

$$AT = \mathrm{softmax}(Q_m K_m^T / \sqrt{d_k})V_m \tag{13}$$

where AT is the normalized attention weight.

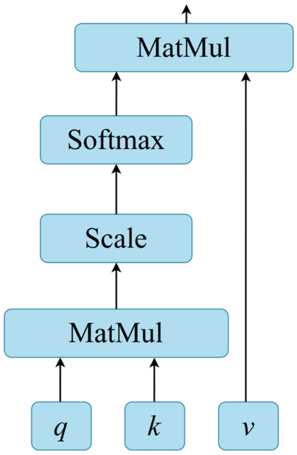

Fig. 3. Scaled dot-product attention.

Models can only learn dependencies in a sequence from one perspective when self-attention is applied alone. However, in SCADA data, a sequence may have many different types of dependencies. Therefore, the use of Multihead self-attention in the model allows for a more diverse representation to be learned. The Multihead self-attention mechanism uses parallelization to compute each projection about q, k and v. As a result, h outputs are obtained. Each output is labeled as $Head_i$ (i=1, 2, 3, ..., h). Finally, the output of each head is spliced to obtain the computation result of multi-head self-attention, as shown in Fig. 4. The calculation process of multi-head self-attention is as follows:

$$\begin{aligned} &MultiHead(Q,K,V) = \mathrm{Conact}(Head_1, Head_2, \ldots\ldots, Head_h)W^o \\ &\text{where } Head_h = AT\left(QW_i^q, KW_i^k, VW_i^v\right) \end{aligned} \tag{14}$$

where $\mathrm{Concat}(\square)$ splices the output from different heads. W_i^q, W_i^k, W_i^v and W_i^o are learnable linear projections. Usually, $d_k = d_v = D_m / h$ 。

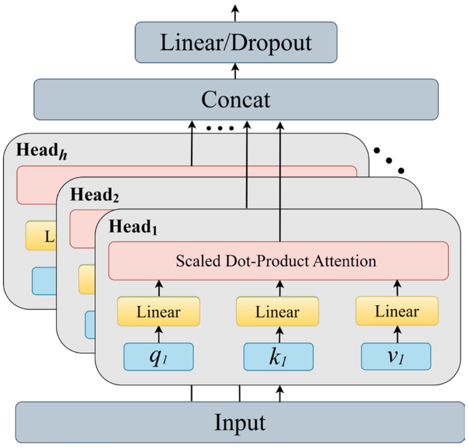

Fig. 4. Schematic diagram of Multihead self-attention.

2.4.3 LSTM layer

In this study, LSTM is employed to process the output results of the Multihead self-attention and is successfully integrated into the core components of the SLFormer. The LSTM layer is introduced mainly to capture long-term dependencies and dynamic temporal information in the data. The LSTM consists of three main gate structures: Forget Gate (FG), Input Gate (IG) and Output Gate (OG) [36]. The role of the FG is to decide what information to discard from the long-term state. It uses a sigmoid function to generate an output value between 0 and 1 based on inputs from the previous hidden state and the current timestep to indicate the proportion of information to retain. The IG is responsible for filtering new information to be added to the long-term state, and is supported by a sigmoid function to determine what information to update, and a hyperbolic tangent (tanh) function to generate new candidate values [37]. The core of LSTM is the cell state, which conveys information at each time step of the sequence. The OG determines the next hidden state based on the cell state and thus plays a key role in sequence prediction and processing. The update process of LSTM is as follows:

$$\begin{cases} f_k = sig\left(W_f * [S_{k-1}, x_k] + h_f\right) \\ i_k = sig\left(W_i * [S_{k-1}, x_k] + h_i\right) \\ o_k = sig\left(W_o * [S_{k-1}, x_k] + h_o\right) \\ \hat{C}_k = \tanh\left(W_c * [S_{k-1}, x_k] + h_c\right) \\ C_k = f_k * C_{k-1} + i_k * \hat{C}_k \\ S_k = o_k * \tanh\left(C_k\right) \end{cases} \tag{15}$$

where $sig(\cdot)$ is the sigmoid function. S_{k-1} is the hidden state at the k-1st time step. x_k is the input at the kth time step. f_k, i_k, and o_k are the outputs of FG, IG, and OG at the kth time step, respectively. $\hat{C}_k$ is the generated new candidate value and C_k is the updated cell state.

Finally, the hyperparameter selection results of SLFormer in this paper are shown in Table 1.

Table 1. The selection result of hyperparameters in SLFormer.

Parameters	Value
Window size	S_l=1024
Sliding step	W_s = 768
Number of features	K = 4
Batch size	B_s = 16
Learning rate	L_r = 1e-4
Optimizer	Adam
Patch size	L_s =64
Depth of SLFormer Encoder	4
Number of attention heads	h = 6
Embedding dimension	D_m =256

Position encoding	1D
Number of LSTM layer	3

3 Detection framework

In order to improve the accuracy of early fault detection of WT bearings, this paper proposes a detection framework based on transfer learning considering the practical applications in wind farms. This framework is developed on the basis of OSWS and SLFormer, and is mainly divided into three tasks, as shown in Fig. 5. The specific content is as follows:

a) Task-1 is performed offline with the aim of obtaining a pre-trained model of the WT's normal behavior, a task that usually takes a long time. First, the normal operation SCADA data of any WT in the target wind farm need to be obtained and preprocessed. The preprocessing mainly includes removing abnormal zeros and missing values. Second, feature correlation analysis is performed to select features with high correlation as inputs to the model. Since the order of magnitude of each variable in the SCADA data is not consistent, it is necessary to normalize the data. Finally, the dataset processed using the OSWS method was input into the SLFormer model to obtain the pre-training weights.

b) Task-2 needs to be performed when the model is deployed. Since it is too inefficient to train a model for each WT when applied to wind farms. Therefore, only a small portion of normal operation data of the target WT needs to be extracted for fine-tuning the pre-trained model in Task-2. This approach allows the deployed model to learn more knowledge, which improves the performance and generalization ability of the model. Finally, the residuals of the predicted and actual values are calculated using the fine-tuned model to determine the control limits of the exponential weighted moving average (EWMA).

c) Task-3 is performed online. First, the online SCADA data are acquired and simply preprocessed, and the same variables normalized as in Task-1 are selected as inputs to the SLFormer model. Second, the residual values were calculated from the output of the model with the measured values and smoothed using EWMA. Finally, the calculated results are compared with the control limits to determine whether the target WT is in a fault state.

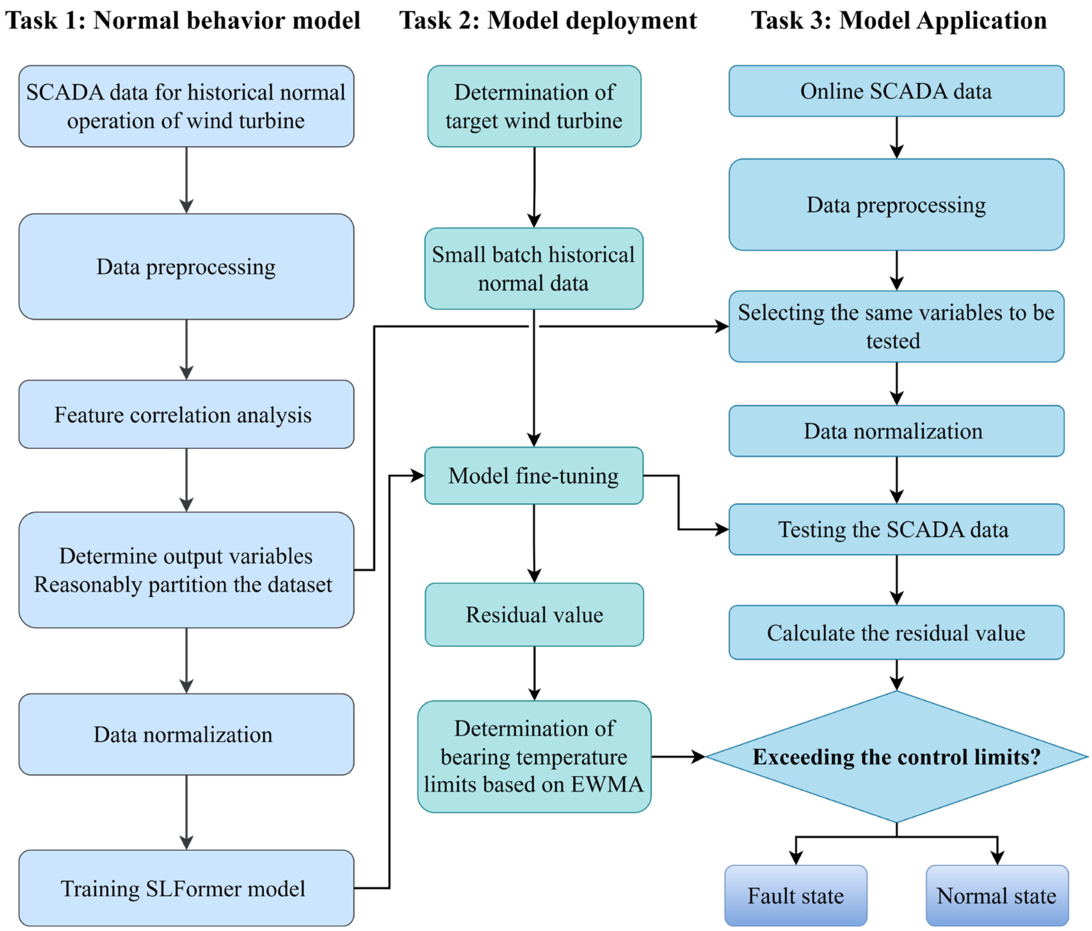

Fig. 5. Flowchart for early anomaly detection of WT gearbox.

4. Case study

Due to the difficulty in obtaining relevant data from offshore wind farms, this paper uses datasets from three onshore wind farms to validate and analyze the effectiveness of the proposed method and its application to early bearing failure detection. The fault detection strategy proposed in this paper is developed based on transfer learning, so it can be applied to OWTs in the future. In addition to this, the fault detection method proposed in this paper requires only the normal operation data of the OWT to fine-tune the model, thus greatly enhancing the adaptability of the model to different units, including the OWT. The three datasets are all SCADA data during WT operation, with a sampling interval of 1 minute. The Boolean variables in the SCADA data are used to determine whether the WT is in normal operation or not. The program is written in the Python language and runs in a Windows 11 environment. The hardware configuration for running the program is Intel i7-13700F, NVIDIA RTX 3060 Ti. The deep learning framework used is Pytorch 1.13.1. The results of bearing fault detection based on the proposed SLFormer will be discussed and analyzed in detail later, including normal behavior modeling, early bearing anomaly detection, and false alarm verification.

4.1. Case 1: Normal behavior model

4.1.1 Dataset Description

In this section of the study, we selected the normal operation data of a WT in eastern China for the period from March to August 2019. This dataset (dataset-1) originally contains data from 200,000 time points, and 195,773 valid samples are obtained after preprocessing. In order to assess the operational status of the gearbox bearing as well as to ensure the future application of the model to OWTs, this paper chose several variables related to bearing temperature, namely: main shaft front bearing temperature (MSFBT), main shaft rear bearing temperature (MSRBT), environmental temperature (ET), nacelle temperature (NT), hub temperature (HT). By analyzing the correlation of these five variables, the results are displayed in Fig. 6. The results of the analysis showed that there was a significant correlation between MSRBT, ET, NT, HT and MSFBT, with correlation coefficients of 0.949, 0.825, 0.884, and 0.908, respectively. Based on this finding, the study chose the four variables mentioned above as the input variables for the establishment of a predictive model to forecast the changes in MSFBT. In addition, the trend of changes in the original data over time is is exhibited in Fig 7. From this figure, it can be observed that as the ambient temperature rises, the other four parameters also show a corresponding upward trend. However, these changes are still in a controllable state as a whole and show some cyclical characteristics.

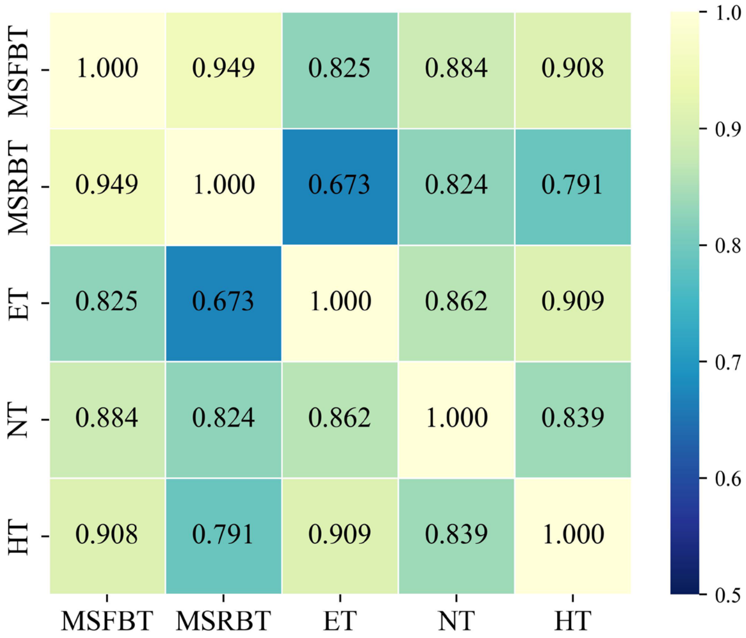

Fig. 6. Results of feature correlation calculation.

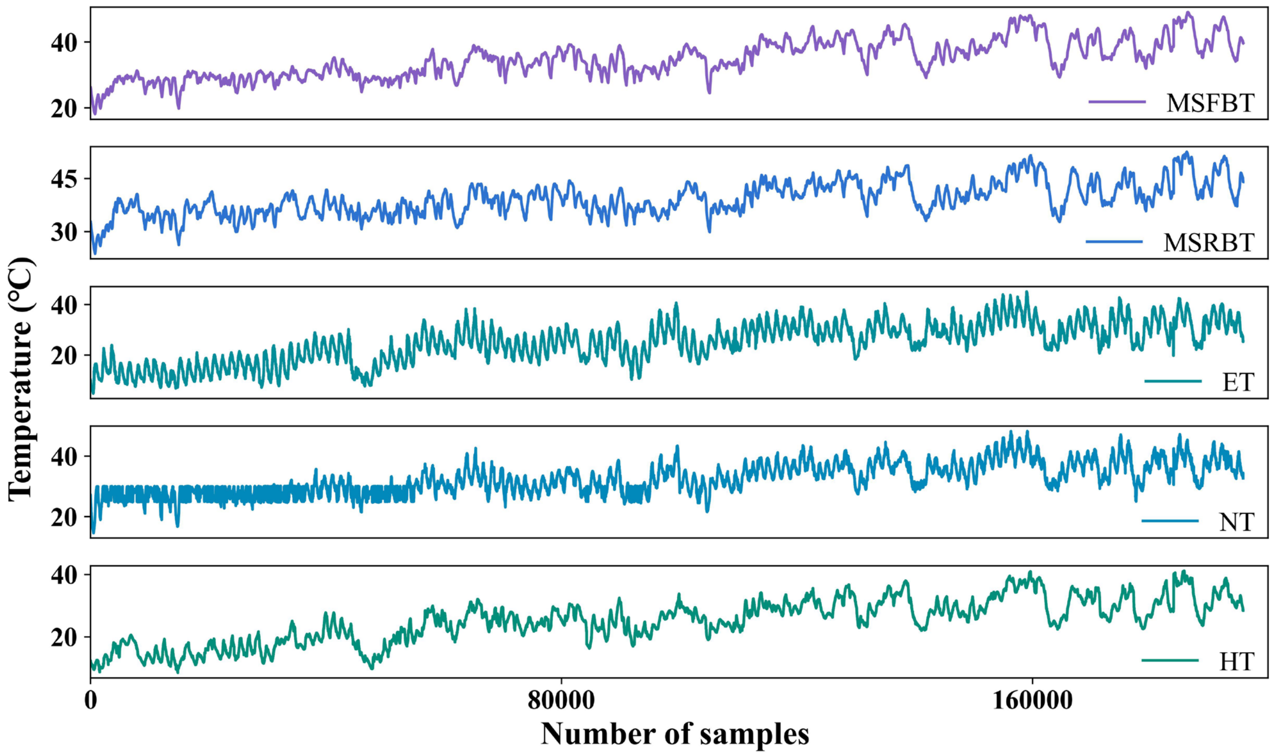

Fig. 7. Time-domain variation of features in dataset-1.

4.1.2 Performance evaluation indicators

In this paper, four evaluation metrics are computed for evaluating the performance of SLFormer using true and predicted values: mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and R-squared (R^2). The four performance evaluation indicators (PEI) are defined as follows [37, 38]:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{17}$$

$$MAPE = \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{18}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2} \tag{19}$$

where m is the length of the prediction. y_i and $\hat{y}_i$ are the true and predicted value of the ith point, respectively.

4.1.3 Model performance verification

This section focuses on evaluating the effectiveness of the proposed model in modeling WT normal behavior. The original data are split into training and test sets, which are distributed in the ratio of 7:3. The number of training epochs was set to 20, and other related parameter settings are detailed in Table 1. The optimal model weights were selected for later model deployment. During the entire training process, we recorded and analyzed the changes of loss values and R^2 coefficients on the training and test sets, as shown in Fig. 8. During the initial training phase, significant fluctuations in the loss on the training and test sets occurred due to the random initialization of the model parameters. Especially during the first epoch of training, a large gap between the results of the training and test set appeared. However, as the training proceeds further, the performance of the model on these two datasets gradually converges, showing good generalization ability. After about 10 epochs, the R^2 and loss values of the training and test sets are gradually stabilized and no longer show large fluctuations, indicating that the method proposed in this study has strong convergence. In the final stage of the training process, the model remains largely stable in terms of R^2 and loss values, further demonstrating the strong generalization ability of the proposed network structure.

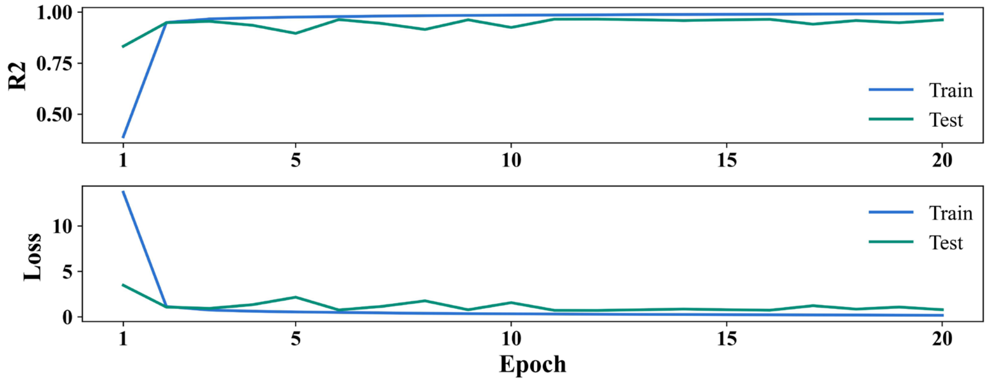

Fig. 8. Variation of R^2 and loss in each epoch of training and testing.

Fig. 9 shows the prediction results and residual distribution of the SLFormer model proposed in this paper on the test set. The results show that the model is able to accurately predict the bearing temperatures and most of the predicted values are very close to the actual values. The residual values are mainly concentrated in the range of -3 to 3, which indicates that the prediction error of the model is small and the residual distribution is uniform. This tight residual distribution further proves the efficiency and accuracy of the SLFormer model. In addition, it can be observed from the curves that the residuals do not show any obvious systematic deviation in the whole test set, which indicates that the model is well adapted and robust to the data under different conditions. Overall, the SLFormer model shows excellent performance in handling complex time series data, especially in terms of accuracy and stability.

Fig. 10 shows in detail the prediction results of the model based on the original Transformer architecture on the test set and its residual distribution. The analysis results show that the Transformer model exhibits good prediction accuracy in the vast majority of cases. However, the model's prediction performance is poor under specific operating conditions, i.e., elevated bearing temperatures due to random perturbations. This performance deviation may stem from the insufficient amount of data or uneven data distribution of the model in the high-temperature interval, resulting in limited learning capability of the model in this interval. Such random perturbations are frequent for the complex operating environment of WT. Therefore, this finding suggests a limitation of the Transformer model in dealing with extreme values in the dataset. In contrast, the prediction results of the model proposed in this study in the same situation show a much milder deviation, demonstrating its superior adaptability in dealing with extreme data points.

Fig. 11 shows a demonstration of the prediction efficacy and its residual distribution when using the SLFormer model and samples from a single time point. From the illustration, it can be observed that the prediction performance of the model significantly decreases when dealing with higher bearing temperature values, which can be largely attributed to the neglect of temporal information in the selection of data samples. Since the model mainly learns the interaction law between features at a single point in time and fails to adequately consider the dynamic changes in time latitude, its adaptive ability in the high bearing temperature condition is limited, thus failing to achieve accurate temperature prediction. The residual curves reveal a significant high error bias in the model's prediction, which may lead to false alarm problems in practical applications. Therefore, the limitations of the prediction method based on a single time-point data sample in practical deployment cannot be ignored. This also demonstrates the effectiveness of the OSWS methodology used in this paper.

In order to deeply validate the performance advantages of the proposed SLFormer model, in this study, after optimizing the data preprocessing using the OSWS method, the model is further comprehensively compared with six current mainstream state-of-the-art models. As shown in Table 2, the scores are based on four PEIs: RMSE, MAE, MAPE and R^2. The evaluation results show that the SLFormer model proposed in this paper achieves the best performance in all of these metrics, with specific values of 0.8464 (RMSE), 0.6147 (MAE), 1.5286 (MAPE), and 0.9656 (R^2), respectively. This result strongly demonstrates the significant effectiveness of the SLFormer model in modeling WT normal behavior.

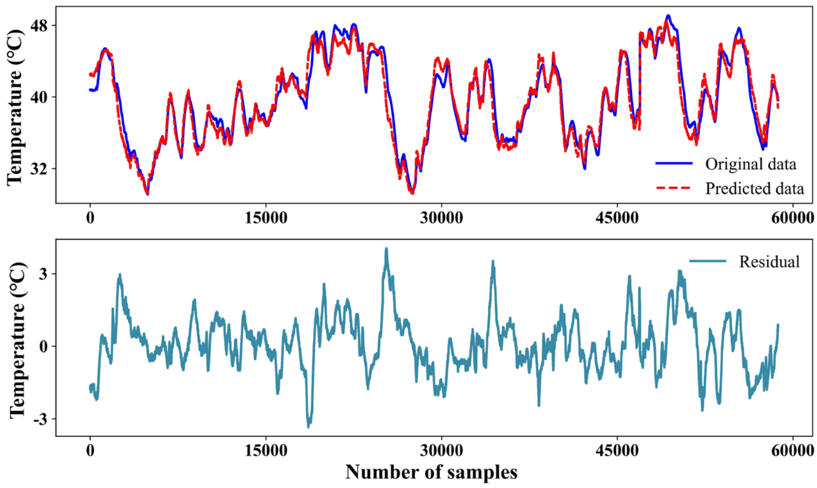

Fig. 9. The prediction results and residual distribution using the OSWS and SLFormer on the test set (dataset-1).

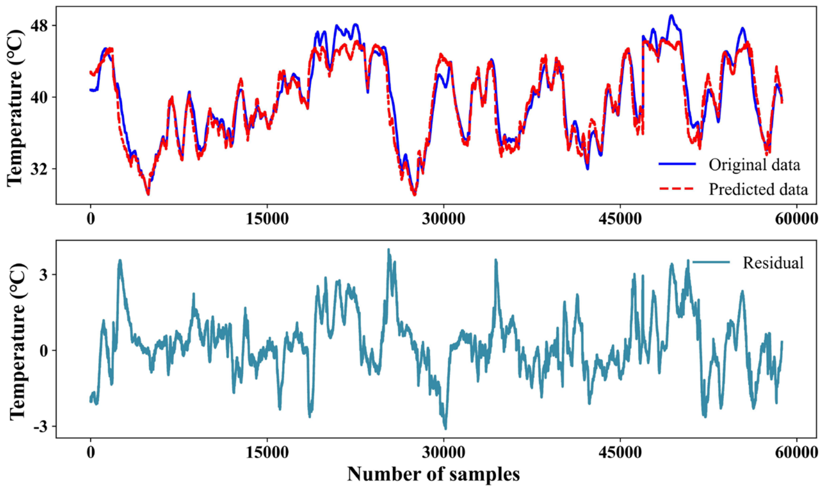

Fig. 10. The prediction results and residual distribution using the OSWS and Transformer on the test set (dataset-1).

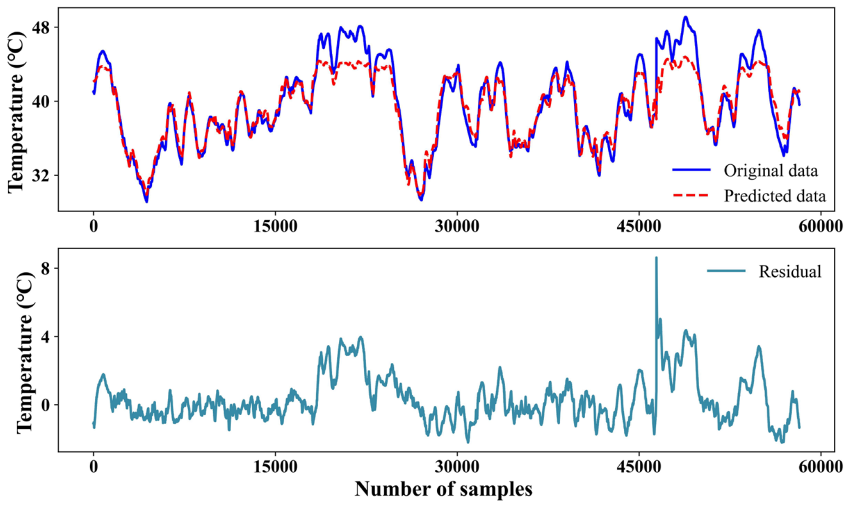

Fig. 11. The prediction results and residual distribution using single time point sample and SLFormer on the test set (dataset-1).

Table 2. Prediction performance of different models.

Methods/PEIs	RMSE	MAE	MAPE [%]	R^2
SVR	1.5891	1.1753	2.8047	0.8775
RFR	1.7855	1.3268	3.2238	0.8467
CNN-LSTM	1.1242	0.8674	2.1681	0.9387
Bi-LSTM	1.4723	1.3017	3.3378	0.8958
PSO-LSTM	1.3944	0.9940	2.3753	0.9065
Transformer	0.9695	0.6936	1.6836	0.9548
SLFormer	**0.8464**	**0.6147**	**1.5286**	**0.9656**

4.2. Case 2: Early fault detection

4.2.1 Dataset Description

In this section, the SCADA data of a WT in North China for the period of February to May 2019 is selected as the dataset to be analyzed (dataset-2). On May 23, 2019, the WT alarmed due to a serious abnormality in the gearbox bearing, leading to a shutdown for inspection. This original dataset contains data from 107,321 time points, and 105,501 valid samples are obtained after preprocessing. The input variables of the model are the same as in Section 4.1.1, and the variable to be predicted is MSFBT. Fig. 12 demonstrates the trend of these five variables over time. As can be seen from the graphical information, it is not possible to directly observe early anomalies in the bearings from the raw data of any single feature. Therefore, further analysis and research are necessary to determine a reasonable alarm threshold.

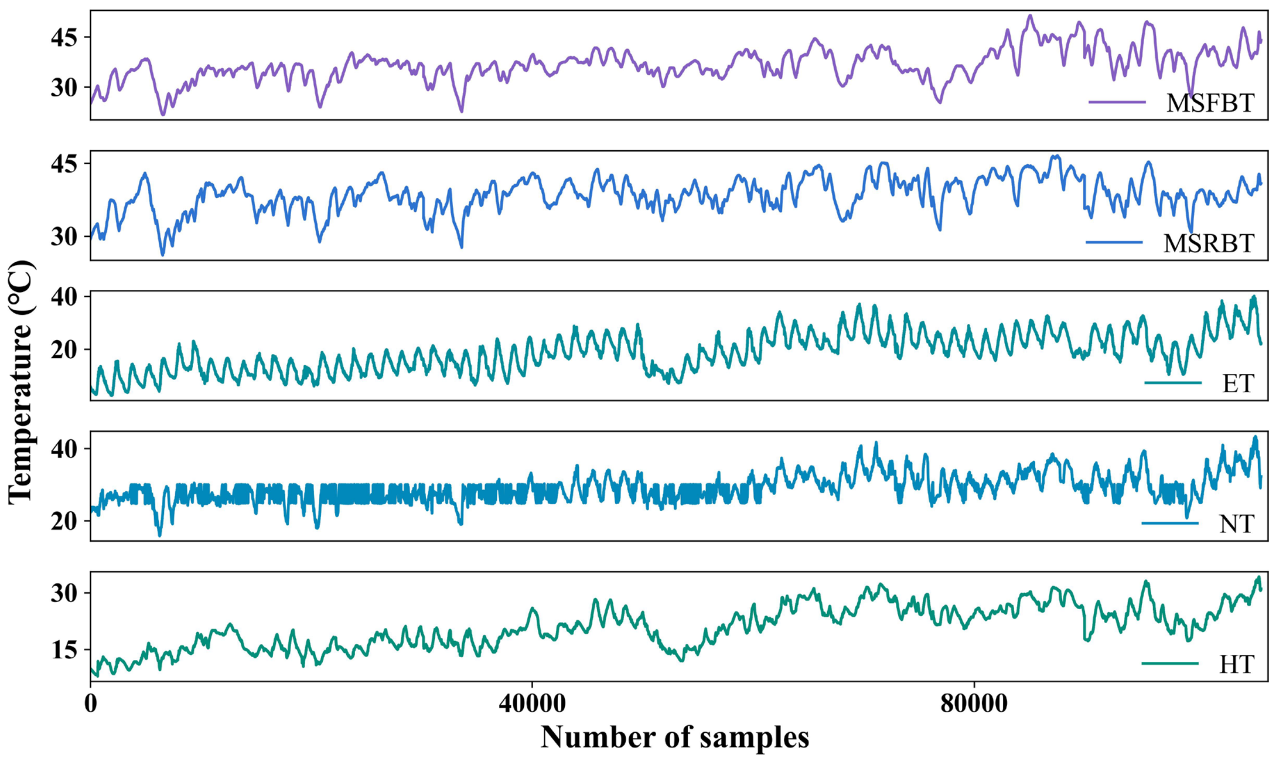

Fig. 12. Time-domain variation of features in dataset-2.

4.2.2 Method for setting alarm threshold

During WT operation, the monitoring data of each component will be affected by gusts of wind, frequent switching of operating conditions, etc. As a result, the fluctuation of the bearing temperature will also be drastic. If the residual difference between the measured and predicted values is set to a fixed threshold, this will lead to frequent alarms in the system due to random disturbances. Therefore, this paper uses the EWMA method to set a threshold line. It gives greater weight to data that are close, while data that deviate more receive a smaller weight [24]. The fluctuations of the residuals can be effectively detected by utilizing the thresholds set by EWMA to monitor the operating status of the WT in real time. For the residual sequence R:

$$R = \left(r_1, r_2, r_3, \ldots r_{j-1}, r_j\right) \tag{20}$$

The ith statistic U_i in the EWMA control chart can be expressed as:

$$U_i = \alpha r_i + (1-\alpha) U_{i-1} \tag{21}$$

where α is the weight of the historical residuals on the current statistic.

The formula for calculating the upper limit of EWMA is as follows:

$$U_l = \mu_l + \frac{\beta \sigma_l}{N} \sqrt{\alpha / (2-\alpha)} \tag{22}$$

where U_l is the upper threshold of the control chart. μ_l and σ_l are the mean and standard deviation of the residuals, respectively. β is a constant related to the upper control limit, which is obtained in this paper by training the WT normal behavior model. N is the capacity of the residual sequence.

4.2.3 Result analysis

The main objective of this section of the study is to evaluate the performance of the model proposed in this paper in detecting early anomalies in WT bearings. In the original dataset, the first 20,000 samples were first selected as normal operation data and used to fine-tune the pre-trained model. The remaining data were then used as a test set in order to obtain predictive results. This approach allows the model to learn and master the operating characteristics of the target WT. The other parameter settings remain the same as described in the previous section. Fig. 13 shows the performance of SLFormer model on the test set. The illustrated results show that the model performs excellently with normal operational data early in the test set. However, after about 61,000 samples, the actual values of the bearing temperatures are significantly higher than the predicted values, which suggests that there are significant anomalies in the bearing temperatures at this point. However, it is still impossible to determine exactly when the early anomaly occurred.

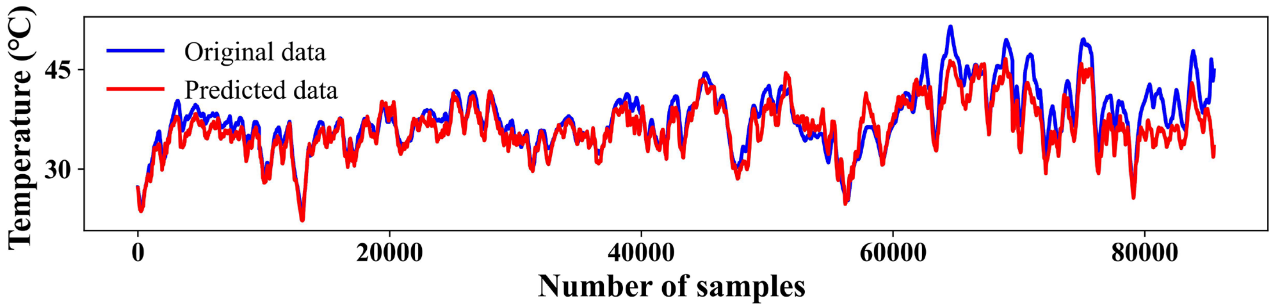

Fig. 13. The prediction results using the OSWS and SLFormer on the test set (dataset-2).

Fig. 14 illustrates the EWMA control chart of the residuals between the true and predicted values of the bearing temperature. Analyzing the data in the graph, it can be seen that during the normal operation of the WT, all the residuals were located within the preset control limits, indicating that the model proposed in this study did not trigger any alarm mechanism at this stage. However, on May 2, 2019, the residuals violated the threshold for the first time. Thereafter, this threshold was breached several times and continued until May 23 when the WT was shut down for inspection due to a malfunction. This phenomenon can indicate that the potential failure of this WT has reached a certain level. The timing of the maximum level of failure predicted by the model is consistent with the downtime. Comprehensive analysis shows that the model proposed in this paper is able to provide early warning 21 days before downtime, thus confirming the efficiency and practicality of the model in detecting early anomalies in WT bearings.

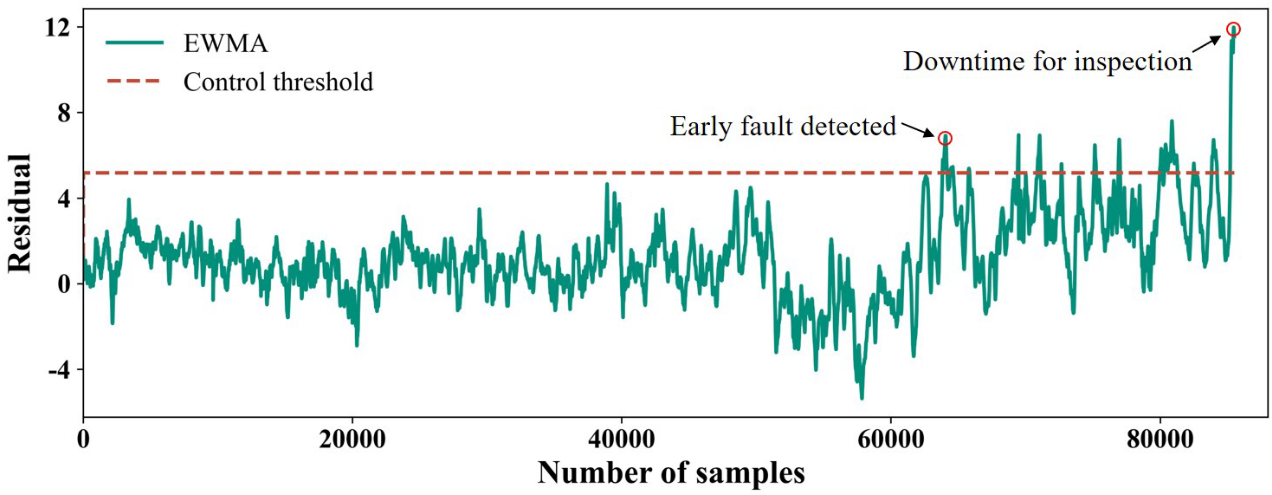

Fig. 14. EWMA control charts (dataset-2).

4.3. Case 3: False Alarm Verification

The dataset for Case 3 is derived from a normally operating WT in a wind farm in Northern China (dataset-3). The dataset contains operational data from November 2019 to January 2020. After data preprocessing, a total of 49,552 samples were obtained. Among these data, the first 12,000 samples were selected as the training set for fine-tuning the pre-trained model. The remaining samples were used as the test set. The performance of SLFormer on the test set is shown in Fig. 15. The number of training samples and seasonal factors lead to high prediction error of the model. However, the model's prediction of the overall trend was not significantly affected. The EWMA control chart of the residuals of the true and predicted values is shown in Fig. 16. From the figure, it can be observed that the residual values fluctuate within the set threshold range without abnormal alarms. Therefore, the SLFormer model proposed in this paper can effectively prevent the occurrence of false alarm events.

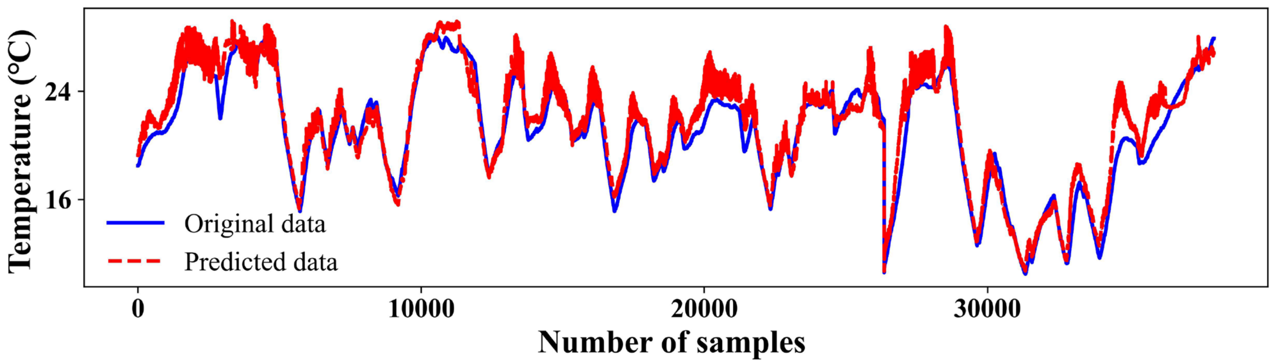

Fig. 15. The prediction results using the OSWS and SLFormer on the test set (dataset-3).

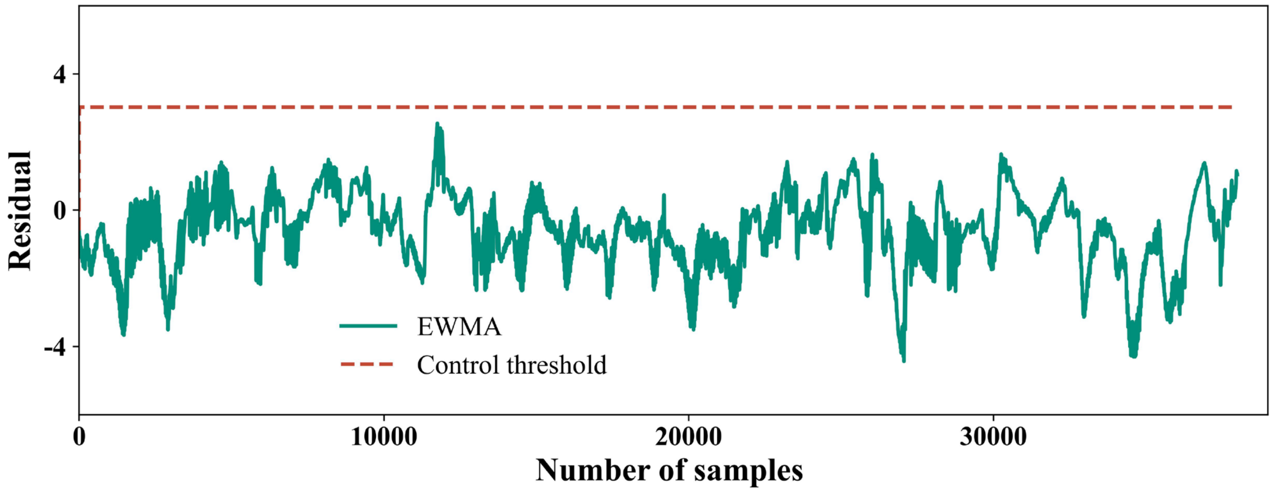

Fig. 16. EWMA control charts (dataset-3).

5 Conclusion

In view of the complex working conditions faced by OWTs during operation, this study designs a model called SLFormer, specifically designed for early anomaly online detection of wind turbine gearboxes. Additionally, combined with the theory of transfer learning, this study proposes a fault detection framework process with wide applicability. Operational data from three onshore wind farms were used in the paper to verify the validity of the model. In the future, the model can be applied to gearbox condition monitoring of OWTs using the strategy proposed in this paper. The main conclusions of this study include:

1) In this paper, we utilized the OSWS approach to process multidimensional SCADA data, effectively capturing self-correlation and inter-correlation in features. This processing demonstrates high adaptability to random disturbances in the prediction results.

2) The SLFormer model proposed in the study incorporates a self-attention mechanism, capable of recognizing important relationships between different parts of the sequence. Meanwhile, the incorporation of LSTM layers helps to capture long-term dependencies in the time series. The case study showed that compared with six other popular models, SLFormer demonstrates significant advantages in both stability and accuracy, especially in modeling WT normal behavior.

3) By analyzing three different SCADA datasets, the detection framework proposed in this paper can effectively detect early anomalies in gearboxes. And it is able to issue a fault warning 21 days in advance, while avoiding the occurrence of false alarms. This universal detection framework provides strong support for practical applications in both onshore and offshore wind farms.

Declaration of Competing Interest:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Statement:

This work is supported by the National Nature Science Foundation of China (No. 52101329), the Marie Skłodowska-Curie Postdoctoral Fellowship (No. EPSRC EP/Y014235/1), the National Engineering Research Center for Offshore Wind power (No. HSFD22004), the High-tech industrialization special fund project of scientific and technological cooperation between Jilin Province and Chinese Academy of Sciences (No. 2023SYHZ0001), and the State Key Laboratory of Mechanical System and Vibration (No. MSV202411).

Acknowledgement:

We would like to thank all those who have reviewed and contributed to this paper for their valuable assistance.

References

1. Global Wind Energy Council (GWEC), Global Wind Report 2023, GWEC, Brussels, 2023.

2. Leite, G.D.P., A.M. Araújo, and P.A.C. Rosas, *Prognostic techniques applied to maintenance of wind turbines: a concise and specific review.* Renewable & Sustainable Energy Reviews, 2018. **81**: p. 1917-1925.

3. Bhardwaj, U., A.P. Teixeira, and C.G. Soares, *Reliability prediction of an offshore wind turbine gearbox.* Renewable Energy, 2019. **141**: p. 693-706.

4. Feng, K., et al., *A novel order spectrum-based Vold-Kalman filter bandwidth selection scheme for fault diagnosis of gearbox in offshore wind turbines.* Ocean Engineering, 2022. **266**.

5. Li, H. and C.G. Soares, *Assessment of failure rates and reliability of floating offshore wind turbines.* Reliability Engineering & System Safety, 2022. **228**.

6. Li, H., C.G. Soares, and H.Z. Huang, *Reliability analysis of a floating offshore wind turbine using Bayesian Networks.* Ocean Engineering, 2020. **217**.

7. Márquez, F.P.G., et al., *Condition monitoring of wind turbines: Techniques and methods.* Renewable Energy, 2012. **46**: p. 169-178.

8. Shanbr, S., et al., *Detection of natural crack in wind turbine gearbox.* Renewable Energy, 2018. **118**: p. 172-179.

9. Li, Z.X., et al., *Multi-dimensional variational mode decomposition for bearing-crack detection in wind turbines with large driving-speed variations.* Renewable Energy, 2018. **116**: p. 55-73.

10. Saari, J., et al., *Detection and identification of windmill bearing faults using a one-class support vector machine (SVM).* Measurement, 2019. **137**: p. 287-301.

11. Yan, G.X., C.Q. Yu, and Y. Bai, *Wind Turbine Bearing Temperature Forecasting Using a New Data-Driven Ensemble Approach.* Machines, 2021. **9**(11).

12. Hou, Z.N., X.X. Lv, and S.X. Zhuang, *Optimized Extreme Learning Machine-Based Main Bearing Temperature Monitoring Considering Ambient Conditions' Effects.* Energies, 2021. **14**(22).

13. Sun, H.L., Y.Y. Zi, and Z.J. He, *Wind turbine fault detection using multiwavelet denoising with the data-driven block threshold.* Applied Acoustics, 2014. **77**: p. 122-129.

14. Zhang, Z.Y. and K.S. Wang, *Wind turbine fault detection based on SCADA data analysis using ANN.* Advances in Manufacturing, 2014. **2**(1): p. 70-78.

15. Chen, X.J., et al., *Vibration fault diagnosis of wind turbines based on variational mode decomposition and energy entropy.* Energy, 2019. **174**: p. 1100-1109.

16. Wang, H., et al., *Early Fault Detection of Wind Turbines Based on Operational Condition Clustering and Optimized Deep Belief Network Modeling.* Energies, 2019. **12**(6).

17. Encalada-Dávila, A., et al., *Early Fault Detection in the Main Bearing of Wind Turbines Based on Gated Recurrent Unit (GRU) Neural Networks and SCADA Data.* Ieee-Asme Transactions on Mechatronics, 2022. **27**(6): p. 5583-5593.

18. Yang, B.Y., A.M. Cai, and W.R. Lin, *Analysis of early fault vibration detection and analysis of offshore wind power transmission based on deep neural network.* Connection Science, 2022. **34**(1): p. 1005-1017.

19. Zhang, F.H., et al., *Abnormality Detection Method for Wind Turbine Bearings Based on CNN-LSTM.* Energies, 2023. **16**(7).

20. Tutivén, C., et al., *Detecting bearing failures in wind energy parks: A main bearing early damage detection method using SCADA data and a convolutional autoencoder.* Energy Science & Engineering, 2023. **11**(4): p. 1395-1411.

21. Wu, Y.Q. and X.D. Ma, *A hybrid LSTM-KLD approach to condition monitoring of operational wind turbines.* Renewable Energy, 2022. **181**: p. 554-566.

22. Oliveira, A., et al., *Early Detection and Diagnosis of Wind Turbine Abnormal Conditions Using an Interpretable Supervised Variational Autoencoder Model.* Energies, 2023. **16**(12).

23. Yan, J.S., Y.Q. Liu, and X.Y. Ren, *An Early Fault Detection Method for Wind Turbine Main Bearings Based on Self-Attention GRU Network and Binary Segmentation Changepoint Detection Algorithm.* Energies, 2023. **16**(10).

24. Xiang, L., et al., *Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism.* Measurement, 2021. **175**.

25. Yang, Q.M., et al., *Fault Detection of Wind Turbine Generator Bearing Using Attention-Based Neural Networks and Voting-Based Strategy.* Ieee-Asme Transactions on Mechatronics, 2022. **27**(5): p. 3008-3018.

26. Hou, S.X., A. Lian, and Y.D. Chu, *Bearing fault diagnosis method using the joint feature extraction of Transformer and ResNet.* Measurement Science and Technology, 2023. **34**(7).

27. Jiao, Z.Y., et al., *Partly interpretable transformer through binary arborescent filter for intelligent bearing fault diagnosis.* Measurement, 2022. **203**.

28. de Azevedo, H.D.M., A.M. Araújo, and N. Bouchonneau, *A review of wind turbine bearing condition monitoring: State of the art and challenges.* Renewable & Sustainable Energy Reviews, 2016. **56**: p. 368-379.

29. Wu, P., et al., *Floating offshore wind turbine fault diagnosis via regularized dynamic canonical correlation and fisher discriminant analysis.* Iet Renewable Power Generation, 2021. **15**(16): p. 4006-4018.

30. Zhang, X.J., et al., *Floating offshore wind turbine fault diagnosis using stacked denoising autoencoder with temporal information.* Transactions of the Institute of Measurement and Control, 2021.

31. Rahimilarki, R., et al., *Convolutional neural network fault classification based on time-series analysis for benchmark wind turbine machine.* Renewable Energy, 2022. **185**: p. 916-931.

32. Jin, Y.H., L. Hou, and Y.S. Chen, *A Time Series Transformer based method for the rotating machinery fault diagnosis q.* Neurocomputing, 2022. **494**: p. 379-395.

33. Ding, Y.F., et al., *A novel time-frequency Transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings.* Mechanical Systems and Signal Processing, 2022. **168**.

34. Tang, J., et al., *Signal-Transformer: A Robust and Interpretable Method for Rotating Machinery Intelligent Fault Diagnosis Under Variable Operating Conditions.* Ieee Transactions on Instrumentation and Measurement, 2022. **71**.

35. He, Q.C., et al., *A Siamese Vision Transformer for Bearings Fault Diagnosis.* Micromachines, 2022. **13**(10).

36. Ma, L.L., et al., *Fault Prediction of Rolling Element Bearings Using the Optimized MCKD-LSTM Model.* Machines, 2022. **10**(5).

37. Xiao, X.C., et al., *Condition Monitoring of Wind Turbine Main Bearing Based on Multivariate Time Series Forecasting.* Energies, 2022. **15**(5).

38. Emeksiz, C. and M. Tan, *Wind speed estimation using novelty hybrid adaptive estimation model based on decomposition and deep learning methods (ICEEMDAN-CNN).* Energy, 2022. **249**.