



Large model construction method for fault diagnosis of heavy gas turbine based on LoRA low-rank modulation

Hang Ruan^{1,*}, Nailiang Sun¹ and Hong He¹

¹ CNOOC Gasand Power Group Co., Ltd. Chaoyang District, Beijing 100028

SUMMARY: Aiming at the problems of high dimension of heavy gas turbine operation data, rapid change of working conditions, unbalanced fault samples and high training cost of traditional deep models, a large fault diagnosis model construction method based on LoRA low-rank modulation was proposed. The method takes multi-source operation data as input, establishes a fault semantic representation module, and maps time series signals such as temperature, pressure, flow, speed, vibration and control feedback into high-dimensional representations with context correlation. On this basis, the LoRA low-rank modulation adaptation mechanism is introduced to efficiently fine-tune the parameters of the pre-trained time series large model, and the feature fusion of time domain, frequency domain and working condition is combined to improve the complex fault recognition ability. Experimental results show that the accuracy, recall rate and macro-average F1 value of the proposed model on the test set reach 96.8%, 95.1% and 95.7% respectively, which are better than those of SVM, CNN, LSTM and standard Transformer models. At the same time, the number of trainable parameters is reduced from 88.4M to 9.6M, which reduces the computational overhead while ensuring the diagnosis performance. The research shows that the proposed method has good accuracy, robustness and deployment feasibility in the intelligent diagnosis scenario of heavy gas turbine.

Povzetek: Raziskava predlaga metodo velikega modela z LoRA nizko-rang modulacijo za diagnostiko okvar težkih plinskih turbin. Model učinkovito združuje večvirovne časovne podatke in dosegla visoko natančnost, priklic in F1, ob zmanjšanem številu parametrov.

KEYWORDS: heavy-duty gas turbine; Fault diagnosis; LoRA low-rank modulation; Large model

1 Introduction

Heavy-duty gas turbine is the core power equipment in large combined cycle unit and industrial drive system, and its running state directly affects the power generation efficiency, peak shiller ability and the safety and stability of the whole system. Due to the long-term high temperature, high pressure, high speed and strong coupling load environment, there will be a complex dynamic association between compressor, combustion chamber, turbine and its auxiliary control unit. Once the performance degradation of local components occurs, it will be gradually transmitted through the gas path parameters, hot end state and vibration response. It is manifested as temperature discharge offset, pressure imbalance, efficiency decline and even unit unplanned shutdown [1-3]. From the perspective of engineering phenomena, many faults do not appear in the form of mutation, but gradually accumulate in the way of weak anomaly,

*rhqditc1978@126.com

<https://doi.org/10.65102/is2026543>

slow drift and multi-variable linkage deviation, which makes fault diagnosis evolve from a single threshold judgment problem to a typical complex timing pattern recognition problem.

With the continuous improvement of industrial control systems, distributed sensor networks and unit monitoring platforms, a large number of multi-source data can be collected in real time during the operation of heavy gas engines, covering multiple dimensions such as temperature, pressure, flow rate, speed, gas parameters, vibration signals and control commands [2,4]. This kind of data has obvious characteristics of high dimensionality, heterogeneity, time correlation and operating condition dependence. It not only contains equipment health state evolution information, but also contains sensor noise, sampling error and operation disturbance. For computer modeling, the real difficulty is not only the limited classification labels, but how to organize the discrete, continuous, strongly correlated and cross-time scale data into a learnable fault representation, and maintain the generalization ability and diagnostic stability of the model under complex operating conditions.

The existing research has gradually shifted from the traditional performance diagnosis to the intelligent analysis framework combining data-driven and mechanism-aided. The method based on compressor characteristic diagram modification, state observer, tracking filter and probabilistic graphical model has certain advantages in component performance degradation analysis, thermal channel fault inference and uncertainty expression [5-8]. However, these methods usually rely on strong prior mechanism and parameter tuning, and the model adaptability is still limited when facing multi-fault coupling, nonlinear drift and complex operating boundaries. In contrast, data-driven methods such as support vector machines, convolutional neural networks, transfer learning and interpretable convolutional models have shown better flexibility in complex feature extraction and pattern classification [9-12]. However, most of the existing models are designed around a single task, fixed working conditions or local features, and the processing of cross-working condition knowledge transfer, long-term temporal dependency modeling and unified representation of multi-source information is still insufficient. Especially in the scenarios of class imbalance, small sample fault and composite anomaly recognition, the model performance fluctuates greatly [13, 14].

From the perspective of the development trend of computer methods, fault diagnosis tasks are shifting from shallow recognition dominated by traditional feature engineering to deep intelligent analysis with representation learning, context modeling and efficient parameter transfer as the core. Transformers and their derived models have shown strong global modeling ability in natural language processing, visual understanding and time series analysis, which can capture long-distance dependencies through the self-attention mechanism and construct the association expression between complex data in a unified encoding space [15-18]. For heavy-duty gas turbine fault diagnosis, this capability is particularly critical because gas turbine anomalies are often not triggered by a single measurement point in isolation, but by the combined deviation of multiple sensing variables formed in different time Windows. If local convolution or short window statistics are still used for modeling, it is easy to miss the remote dependence and implicit coupling relationship in the fault evolution process.

However, directly applying the general large model to industrial fault diagnosis will also face a series of practical constraints at the computer implementation level. On the one hand, the number of high-quality labeled fault samples in the field of heavy gas turbine is limited, the distribution of fault categories is uneven, and the problems of sample scarcity and long-tail distribution are prominent. On the other hand, large model parameters have huge scale, if full parameter fine-tuning is used, not only the training cost is high, but also the computing power, video memory and training time requirements are high, and the existing general representation capabilities may be overcovered, resulting in problems such as low transfer efficiency and unstable domain adaptation [19]. Therefore, the key to building a large fault diagnosis model in

industrial scenarios is not only to "use a large model", but also to solve "how to complete domain adaptation at a lower cost, how to make the model truly understand the timing semantics of the equipment, and how to improve the transferability and deployability of the model under complex working conditions".

LoRA low-rank modulation provides a more practical technical path for this problem. By introducing a trainable low-rank increment into the key weight matrix of the pre-trained model, this method converts large-scale parameter update into small-scale structural adaptation, and can realize task transfer with fewer training parameters and lower computational overhead [19]. Introducing LoRA into the fault diagnosis of heavy gas turbine not only helps to reduce the training threshold of large models in industrial scenarios, but also helps to retain the general sequence representation ability and make targeted modulation for the specific distribution of gas turbine operation data. If multi-source signal coding, fault semantic representation and feature fusion mechanism are further combined, the model is likely to obtain more stable performance in condition switching, weak fault recognition and composite anomaly discrimination.

Based on the above understanding, this paper proposes a large model construction method for heavy gas turbine fault diagnosis fused with LoRA low-rank modulation. The research idea is not to stay in the traditional sense of feature classification, but from the perspective of computer modeling, the multi-source operation data of the heavy-duty gas turbine is organized as a learnable fault semantic sequence, and the large model is used to realize long-term dependence modeling and context correlation representation, and then the LoRA low-rank modulation is used to complete the lightweight domain adaptation. At the same time, the multi-source operation feature extraction and fusion mechanism is introduced. The ability of the model to identify complex fault patterns was enhanced. The core issues of this paper include: how to construct a semantic representation module for gas turbine faults, how to achieve parameter-efficient diagnosis of large model adaptation, and how to ensure the accuracy of diagnosis while balancing training efficiency and engineering deployment feasibility. The related research can provide a new computer method reference for the application of large model in industrial equipment health management.

2 Operation data and fault characteristics analysis of heavy-duty gas turbine

2.1 Data source and collection

The training effect of large model for fault diagnosis of heavy gas turbine depends on the integrity of data source and the stability of acquisition link. The data used in this study mainly come from the distributed control system of heavy gas turbine units, the online state monitoring system and the historical operation and maintenance record library, covering the key operating parameters of compressor, combustion chamber, turbine and its auxiliary control unit. The collected variables include inlet and exhaust gas temperature, pressure at all levels, speed, gas flow rate, exhaust temperature distribution, bearing temperature, vibration amplitude, lubric oil parameters and control command feedback. The types and sampling characteristics of multi-source monitoring data collected in this paper are shown in Table II. In order to ensure the continuity of time series required for subsequent computer modeling, the signals of on-site sensors of the unit are collected by PLC, remote I/O station and edge acquisition gateway, uploaded to the monitoring platform through industrial Ethernet, OPC UA or Modbus communication protocol, and then written into the timing database and fault event database. A data acquisition link of "field perception - edge convergence - platform storage" was formed.

After the above processing, the data from different sources can be aligned under a unified time benchmark, which provides a more stable data basis for the subsequent input construction of large models.

Table 1: Multi source feature dimensions and sampling instructions

Variable Category	Typical Variables	Unit	Sampling Frequency	Data Source Module
Thermal Parameters	Inlet Temperature, Exhaust Temperature, Pressure Ratio	°C / kPa	1 Hz	Compressor / Combustion Chamber
Flow Parameters	Fuel Gas Flow Rate, Lubricating Oil Flow Rate	kg/s	1 Hz	Combustion Chamber / Lubrication System
Mechanical Parameters	Rotational Speed, Bearing Temperature, Vibration Amplitude	r/min / °C / mm/s	10 Hz	Rotor / Bearing / Body Vibration Sensors
Control Commands	Fuel Valve Opening, Regulating Valve Signal	% / Signal Unit	1 Hz	Control System

From the sample composition, the operation data of heavy-duty gas turbine have the characteristics of multivariate, high frequency, strong coupling and obvious dependence on operating conditions. Different operation stages, such as normal load, start-stop switching, power up and down and disturbance adjustment, will make the temperature, pressure and vibration sequence show different change slope and fluctuation rhythm. Once the compressor efficiency decreases, combustion instability, hot end component deterioration or sensor anomalies occur, the relevant parameters do not synchronously mutate, but more often show local offset, slow drift and cross-variable linkage imbalance. For computer diagnosis, this kind of data is not suitable to be simply regarded as a set of isolated measurement points, but should be organized as multi-source time series samples with context relations. Based on this consideration, the original sampling values, sliding time window statistics, alarm logs and maintenance label information are synchronously retained in the acquisition stage, and are stored in segments according to the working condition segment, which provides computable and traceable data input for subsequent fault semantic representation, LoRA low-rank modulation adaptation and multi-source feature fusion. This data organization method not only retains the original state information of the engineering site, but also enhances the learning ability of the model for the complex fault evolution process.

2.2 Analysis of operating conditions and fault characteristics

The fault characteristics of heavy-duty gas turbine operation data usually appear gradually with the change of working conditions, and do not show as a simple single point mutation. From the perspective of time domain, variables such as inlet and exhaust temperature, pressure ratio, speed, vibration amplitude and exhaust temperature dispersion can directly reflect the dynamic state of the unit under different loads. In normal operation, most parameters fluctuate smoothly around a given interval, and the mean, variance and peak values change in an orderly manner. When the compressor efficiency decreases, the combustion deviation increases, or the hot end components deteriorate, the relevant series will show local uplift, fluctuation expansion, and short-time spike increase. This kind of change often does not appear in isolation, but multiple measurement points are synchronized offset, so it has strong linkage. The key operating

parameters and fault response characteristics of heavy gas turbine under typical working conditions are shown in Figure 1.

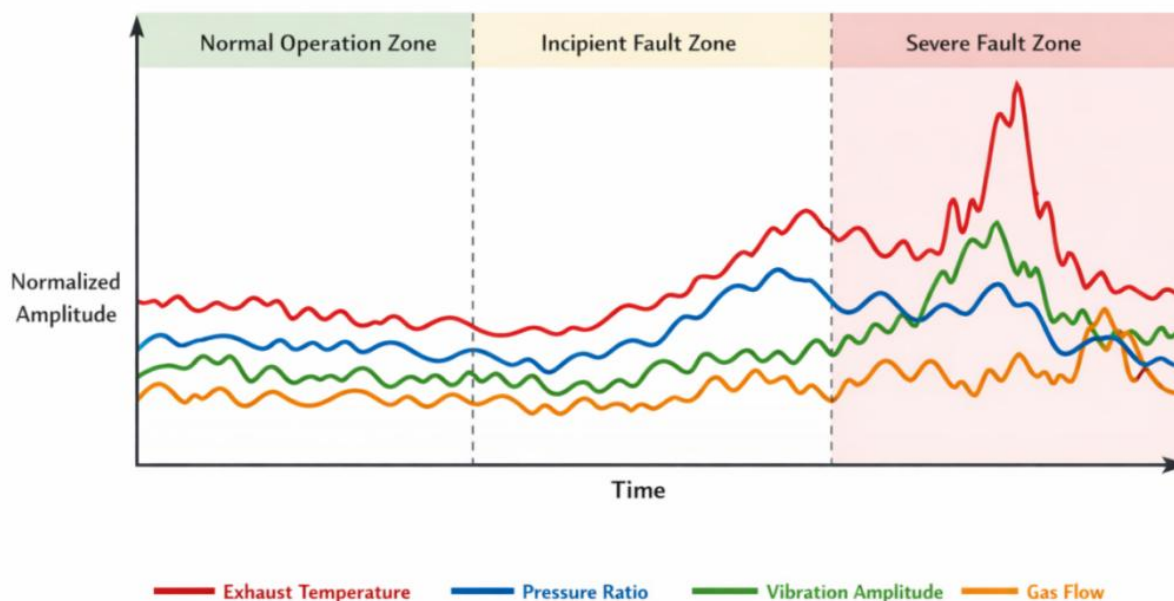


Figure 1: Schematic diagram of typical operation parameters and fault response characteristics of heavy-duty gas turbine

From the frequency domain characteristics, vibration signals and periodic thermal parameters often contain more detailed fault information. Problems such as bearing wear, rotor imbalance, and blade damage can cause energy concentration or amplitude enhancement in specific frequency bands, so that the fault modes show distinguishable response differences in the spectral space. Compared with simply observing the original waveform, frequency domain analysis is more conducive to identifying weak anomalies masked by time-domain fluctuations, and it is also convenient for subsequent computer models to extract stable fault discriminant features. For large models, this composite representation composed of time domain and frequency domain is helpful to improve the identification ability of complex fault modes. At the same time, the data of heavy gas turbine also has obvious non-stationary and nonlinear characteristics. The parameter distribution will drift with time when the unit starts and stops, the load lifting and lowering, the fuel disturbance and the control command adjustment. There are complex coupling between different variables, and temperature, pressure, flow and vibration are not simple linear correspondences. The traditional linear analysis method is difficult to completely describe the anomaly evolution process that is transmitted across time scales and components. Therefore, this paper introduces fault semantic representation, LoRA low-rank modulation and multi-source feature fusion mechanism in the subsequent modeling, so that the model can not only see the parameter changes, but also recognize the structural relationship and fault meaning behind the changes.

2.3 Data preprocessing and sample construction

The operation data of heavy-duty gas turbine need to be carefully preprocessed before entering into the large fault diagnosis model. The signal collected in the field is affected by sensor drift, communication jitter and working condition disturbance, and the original sequence often contains abnormal spikes, random noise and local distortion. In this regard, this paper combines threshold discrimination and sliding window statistical method to identify outliers, and uses the

combination of sliding mean filtering and median smoothing to suppress high-frequency noise, try to retain the real change trend of key variables such as temperature, pressure and vibration, so that the data is closer to the actual operation state of the unit.

Because different monitoring parameters have obvious differences in dimension, value range and fluctuation range, it is easy to cause feature weight imbalance and affect the training convergence effect if they are directly input into the model. In order to enhance the unified representation ability of the computer model for the multivariate series, the continuous monitoring variables are standardized, and the data of different scales are mapped to a comparable numerical space, while preserving the relative fluctuation relationship between the variables. In this way, the model will not cover up other weak fault information due to a single high-magnitude variable when performing timing coding and feature fusion.

For the missing values in the running data, this paper adopts a differentiated treatment strategy according to the length of the missing interval and the change characteristics of the variable. Linear interpolation and neighborhood completion are used to restore the continuity of the sequence for short-term misses, and long-term misses or key label gaps are directly removed to avoid introducing false patterns. After cleaning, the sample segments are constructed according to the fixed time window and sliding step, and the fault labels are generated by combining the alarm records, maintenance logs and expert annotations to form a time series sample set that can be used for LoRA low-rank modulation and large model training. Such a sample construction method not only improves the data availability, but also provides a clearer input basis for the subsequent fault semantic representation.

3 Large model construction method for fault diagnosis of heavy-duty gas turbine based on LoRA low-rank modulation

3.1 Overall framework of the algorithm

The large model of heavy gas turbine fault diagnosis fused with LoRA low-rank modulation is oriented to complex fault recognition tasks under the condition of multi-source monitoring data. The core idea is not to statically classify a single measurement point, but to construct a continuous computing link around "data organization, semantic representation, efficient parameter adaptation, feature fusion, fault output". The overall framework of the model is shown in Figure 2, which is mainly composed of the input and preprocessing layer, the fault semantic representation module, the LoRA low-rank modulation adaptation module, the multi-source operation feature extraction and fusion module, and the fault diagnosis output module. After the operation data of the heavy-duty gas turbine enters the system, the temperature, pressure, flow rate, speed, vibration and control command sequences are first time aligned, window segmented and standardized mapped, and then organized into time series samples that can be processed by the large model.

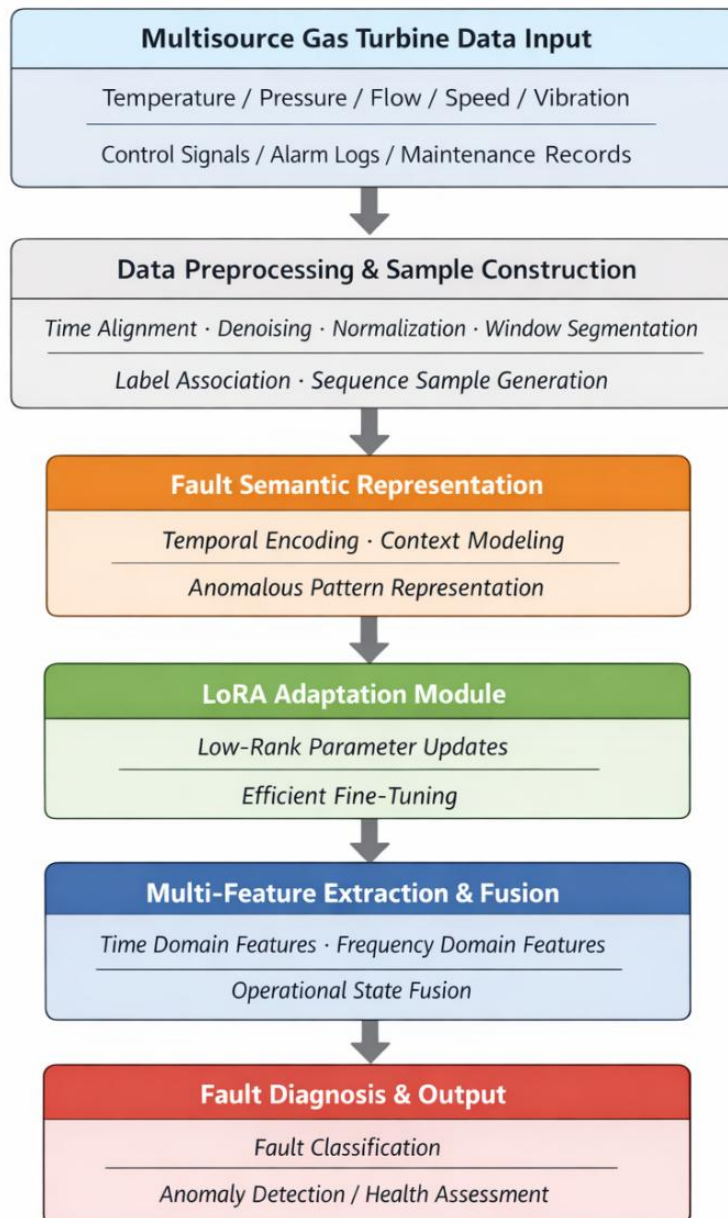


Figure 2: Overall framework of large model for fault diagnosis of heavy-duty gas turbine fused with LoRA low-rank modulation

The fault semantic representation module is responsible for transforming the original numerical sequence into a high-dimensional representation with context relations, so that the model can not only see the amplitude changes of variables, but also identify the linkage structure between parameters and abnormal evolution trajectories. The LoRA low-rank modulation adaptation module is embedded in the key linear layer of the large model, and the lightweight parameter update is completed by introducing a low-rank incremental matrix, which makes the pre-trained model adapt to the heavy-duty gas turbine fault scenario faster without significantly increasing the training overhead. This design is more suitable for small samples, long sequences and multi-condition tasks, and also helps to alleviate the video memory pressure and the risk of overfitting caused by full parameter fine-tuning.

The multi-source running feature extraction and fusion module takes the role of cross-variable information integration. On the one hand, it retains the characteristics of time domain fluctuation, frequency domain response and working condition. On the other hand, it

combines the semantic coding results for weighted fusion, so that the information from different sources such as vibration anomaly, temperature displacement and pressure ratio change can be associated in a unified feature space. The fused features are sent to the fault diagnosis output module to complete the fault category discrimination, abnormal state recognition and health status scoring. On the whole, the framework takes into account the representation ability of large models and the parameter efficiency of LoRA, which not only highlights the learning advantages of computer models for long time series dependencies, but also retains the engineering characteristics of multivariate coupling in the operation mechanism of heavy gas turbine.

3.2 Fault semantic representation module of heavy-duty gas turbine

The task of the fault semantic representation module of heavy-duty gas turbine is not to simply retain the original monitoring values, but to encode multi-source operating signals such as temperature, pressure, flow, speed, vibration and control feedback into contextual temporal semantic representations. For heavy gas turbine, many early faults are not manifested as single variable mutation, but as the linkage imbalance between exhaust temperature offset, pressure ratio fluctuation, vibration enhancement and control quantity compensation. If the isolated feature input method is still used, the model is easy to learn only local numerical differences, and it is difficult to grasp the structural information in the fault evolution. Based on this consideration, the multivariate samples within the sliding time window are denoted as in this paper

$$X = [x_1, x_2, \dots, x_T], \quad x_t \in \mathbb{R}^d \quad (1)$$

Here, T is the time step length and d is the feature dimension at a single instant. After linear mapping, we obtain the query vector, key vector, and value vector:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \quad (2)$$

where, W_Q, W_K and W_V are trainable parameter matrices. Then the self-attention mechanism is used to calculate the correlation strength between different moments and different variables within the sequence:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (3)$$

Here, d_k is the key vector dimension. The calculation process can describe the long-distance dependence in a unified representation space, so that the model not only focuses on the current sampling point, but also understands the context of the anomaly formation by combining the previous and previous states. The fault semantic representation process of heavy gas turbine is shown in Figure 3.

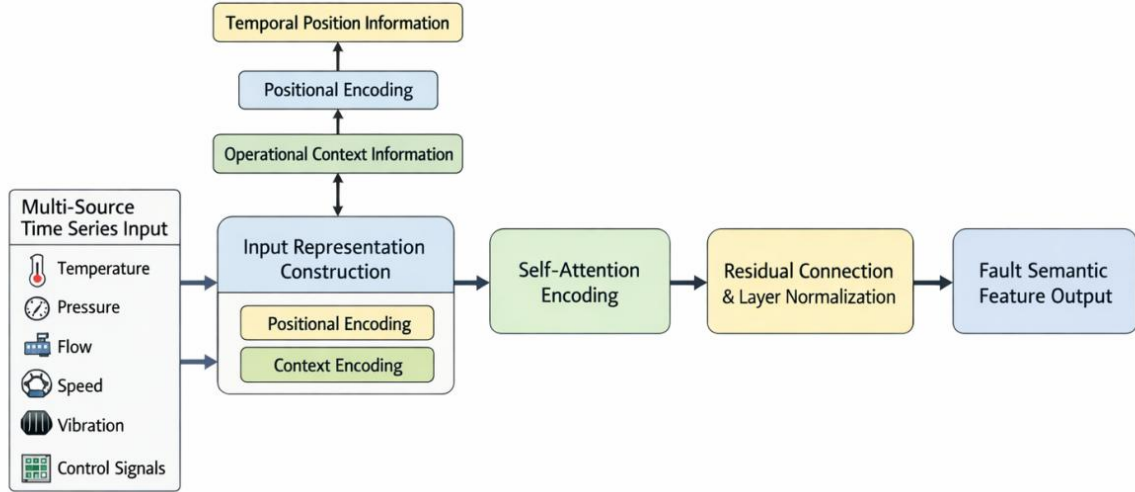


Figure 3: Schematic Diagram of Semantic Characterization Process for Heavy-duty Gas Turbine Failures

In the specific implementation, this paper embeds data from different sources of the heavy-duty gas turbine into the same feature space, and superimposes time position encoding and operating condition identification encoding to form an input representation that combines numerical information, time sequence, and operational scenario information. After this processing, the model can distinguish the semantic differences between "the same temperature fluctuation occurring during the steady-state full-load stage" and "occurring during the load-up transition stage". For the hidden state $H = [h_1, h_2, \dots, h_T]$ after attention aggregation, residual connections and layer normalization are introduced to obtain an enhanced representation:

$$Z = \text{LN}(H + X) \quad (4)$$

The resulting Z is no longer merely the compressed result of the original signal; instead, it incorporates cross-variable correlations, temporal dependencies, and operational background information, representing the fault semantic features. The output of this module will be directly sent to the subsequent LoRA low-rank modulation adaptation layer to complete the lightweight domain transfer for the task of heavy-duty gas turbine fault diagnosis. As a result, the large model can accurately identify complex fault patterns such as compressor degradation, combustion anomalies, and deterioration of thermal end components with fewer parameter updates.

3.3 LoRA Low-Rank Modulation Adaptation Module

The setting of the LoRA low-rank modulation adaptation module is not merely aimed at compressing the training parameters; its more significant purpose is to enable the pre-trained large model to adapt specifically to the fault scenarios of heavy-duty gas turbines at a lower cost. Heavy-duty gas turbine operation data has characteristics such as long time series, multiple variables, small samples, and uneven class distribution. If the large model is directly subjected to full parameter fine-tuning, it will not only result in high memory usage and training costs, but also easily cause the original general representation ability of the model to be overly covered, leading to unstable transfer, fluctuating convergence, and local overfitting problems. Based on this consideration, this paper introduces the LoRA structure into the key linear layers of the large model, restricting the weight update to the low-rank subspace, thereby transforming the

process of injecting domain knowledge from "massive overall rewriting" to "small-scale targeted correction".

Let the original weights of the pre-trained linear mapping layer be $W_0 \in \mathbb{R}^{d \times k}$, and the conventional transformation of the input feature x can be written as

$$y = xW_0 \quad (5)$$

LoRA does not directly update W_0 . Instead, it freezes the original parameters and only learns a low-rank incremental matrix ΔW , whose expression form is

$$\Delta W = BA, B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}, r \ll \min(d, k) \quad (6)$$

Thus, the adapted output can be expressed as

$$y = xW_0 + \frac{\alpha}{r} xBA \quad (7)$$

Among them, r represents the low-rank dimension, and α is the scaling coefficient, which is used to control the modulation intensity of the incremental branches on the main output. This processing method retains the main knowledge structure of the original model while compressing the number of trainable parameters to a lower level, making it more suitable for industrial scenarios such as heavy-duty gas turbines where the annotation cost is high and the number of real fault samples is limited.

Considering that the task of this paper is not only a general time series classification but also involves abnormal identification under the condition of working condition switching, the standard LoRA still needs to be further adapted to the data characteristics of gas turbines. Therefore, this paper adds a condition-aware modulation mechanism to the low-rank update branch, encoding the load level, environmental temperature, control mode, and key operating states into a condition vector c , and generating the modulation coefficient γ through a gating function:

$$\gamma = \sigma(W_c c + b_c) \quad (8)$$

Among them, $\sigma(\cdot)$ is the Sigmoid function. The modulated low-rank update term is written as $\gamma \odot \frac{\alpha}{r} xBA$. This means that the same set of low-rank parameters do not have the same amplitude effect under different operating conditions; instead, their contribution ratios are dynamically adjusted according to the operating background. The significance of this processing lies in that the model can distinguish the differences between "slight temperature fluctuations under steady-state full-load conditions" and "abnormal temperature rise during load increase", thereby reducing the interference of operating condition drift on fault judgment.

In terms of structure deployment, this paper embeds LoRA low-rank modulation preferentially in the query mapping layer, key mapping layer, value mapping layer of the self-attention module, as well as the core linear transformation layer of the feedforward network. The former is responsible for establishing long-term temporal dependencies and cross-variable correlations, while the latter undertakes the task of nonlinear feature reorganization, integrating both into the adaptation scope. This can improve the model's perception ability for various fault modes such as temperature dispersion, pressure ratio deviation, vibration enhancement, and control compensation anomalies without significantly increasing training complexity. During training, the backbone parameters are frozen, and only

the low-rank matrix, modulation parameters, and classification head are updated, thereby reducing the backpropagation overhead and shortening the model convergence time.

From the perspective of computer implementation, this module has both parameter efficiency and deployment friendliness. On the one hand, low-rank decomposition significantly reduces the scale of trainable parameters, alleviating the problem of GPU memory limitations in industrial scenarios; on the other hand, the modulation branch only introduces additional computation on the local linear layer, without significantly disrupting the original model's inference process, making it convenient for subsequent migration deployment on edge servers or unit monitoring platforms. For the fault diagnosis large model constructed in this paper, LoRA low-rank modulation is not an ancillary technique but a key bridge connecting "general pre-training ability" and "heavy-duty gas turbine domain knowledge". Through this module, the model can complete more robust domain adaptation under limited samples and complex operating conditions, laying the foundation for subsequent multi-source feature fusion and fault classification output.

3.4 Multi-source Operating Feature Extraction and Fusion

The setting of the multi-source operating feature extraction and fusion module aims to gather the fault information scattered at different measurement points, with different sampling frequencies and on different physical levels, into a unified representation space during the operation of the heavy-duty gas turbine. For such large rotating thermal equipment, the degradation of compressor efficiency, unstable combustion, deterioration of thermal end components, and abnormal rotor vibration often do not manifest along a single path but are simultaneously projected onto multiple variables such as temperature, pressure, flow rate, discrete temperature deviation, vibration response, and control compensation. If the model relies only on a single sequence feature, it is prone to misjudge the operational disturbances as fault signals and is difficult to identify the transmission relationship of weak anomalies between different variables. Therefore, after LoRA low-rank modulation adaptation, this paper introduces a multi-branch feature extraction and dynamic fusion mechanism to jointly model the time-domain information, frequency-domain information, and operational context information to enhance the discriminative stability of the fault diagnosis large model.

During the feature extraction stage, this paper divides the input samples into three types of information sources: one is the time-domain feature $X^{(t)}$ composed of the original operating sequence, which mainly describes the fluctuation trend, mutation positions, and local drift of the variables over time; another is the frequency-domain feature $X^{(f)}$ obtained from the Fast Fourier Transform and sliding spectral analysis, used to capture the frequency band responses in vibration, temperature deviation fluctuations, and periodic disturbances; and the third is the operational feature $X^{(c)}$ encoded by the load level, start-stop stage, control mode, and environmental conditions, reflecting the operating background of the sample. These three types of features are sent to independent mapping networks respectively, and the corresponding latent vector representations are obtained:

$$H^{(t)} = \phi_t(X^{(t)}), \quad H^{(f)} = \phi_f(X^{(f)}), \quad H^{(c)} = \phi_c(X^{(c)}) \quad (9)$$

Among them, $\phi_t(\cdot)$, $\phi_f(\cdot)$ and $\phi_c(\cdot)$ respectively represent the time-domain encoder, frequency-domain encoder and operating condition encoder. This branch-based processing approach can prevent different source features from being masked from each other during the early mixing stage, and is more in line with the hierarchical extraction rules of heterogeneous data by the computer model.

Based on this, this paper does not adopt the direct classification method after fixed proportion concatenation, but introduces a state-aware gated mechanism to achieve dynamic fusion. Let the semantic features output by the previous fault semantic representation module be $H^{(s)}$, and the preliminary aggregation representation of multi-source features be $H^{(m)} = [H^{(t)}; H^{(f)}; H^{(c)}]$. According to the current sample's operating state, the model calculates the gated vector g :

$$g = \sigma(W_g[H^{(s)}; H^{(m)}] + b_g) \quad (10)$$

Among them, $\sigma(\cdot)$ represents the Sigmoid activation function, and W_g and b_g are trainable parameters. Subsequently, the fused feature representation is obtained:

$$H^{(F)} = g \odot H^{(s)} + (1 - g) \odot W_m H^{(m)} \quad (11)$$

In this formula, \odot represents weighted by elements, and W_m is the dimension mapping matrix. The key of this mechanism does not lie in "summing up the features", but in "deciding which type of features should be emphasized based on the current operating condition". For example, under steady-state full-load conditions, the correlation between exhaust temperature dispersion and thermal end deterioration is stronger. The model will increase the weights of semantic features and thermal engineering variable features; during the load increase or fuel switching stage, the control quantity and dynamic response are more likely to dominate the fusion result, and the fixed weight method is often not reliable in this case.

To verify the effectiveness of the fusion strategy, this paper simultaneously examines two benchmark methods in model design: series fusion and static weighted fusion. Series fusion directly concatenates $[H^{(s)}; H^{(t)}; H^{(f)}; H^{(c)}]$ and inputs it into the fully connected layer to complete the mapping. Although it retains most of the information, the dimension increases significantly, and it is prone to introducing redundant features; static weighted fusion completes the linear combination with preset coefficients, and the calculation is simple, but it is difficult to adapt to the rapid switching of heavy-duty gas turbine operating conditions, causing distribution drift. In contrast, the dynamic gated fusion adopted in this paper can automatically learn the contribution ratio of different information sources during training, enabling the model to have better adaptive ability when facing weak faults, small samples, and composite anomalies. From the perspective of model implementation, this design takes into account both expression ability and computational efficiency. On the one hand, it retains the learning advantage of sequential large models for context relationships; on the other hand, through structured fusion, it reduces the interference of invalid features and provides more discriminative input for the subsequent fault classification head. Thus, the multi-source operating feature extraction and fusion module not only performs the function of information integration but also becomes an important intermediate layer connecting semantic modeling and final diagnostic output.

3.5 Construction of the Fault Diagnosis Large Model

After completing the fault semantic representation, LoRA low-rank modulation adaptation, and multi-source operation feature extraction and fusion, it is necessary to further map the fused features into distinguishable fault state outputs. This paper adopts the construction method of "pre-trained temporal backbone network + lightweight diagnostic task head" to combine the context modeling ability of the large model with the fault classification task. Let the output features of the fusion module be $H^{(F)} \in \mathbb{R}^{T \times d}$, and after global pooling, the sample-level representation vector z is obtained, and its calculation form is

$$z = \text{Pool}(H^{(F)}) \quad (12)$$

Among them, $\text{Pool}(\cdot)$ represents the time-dimensional aggregation operation, which is used to compress the key information in the long sequence and retain the overall fault characteristics. This vector is then input into the diagnostic head composed of fully connected layers to complete the mapping of fault categories:

$$\hat{y} = \text{Softmax}(W_o \text{GELU}(W_h z + b_h) + b_o) \quad (13)$$

In the formula, W_h and W_o are weight matrices, b_h and b_o are bias terms, and \hat{y} represents the probability distribution of each fault category. Compared with directly using a single-layer linear classifier, this shallow task head can re-organize the fused features while keeping the computational cost controllable, thereby improving the ability to distinguish patterns such as compressor degradation, combustion anomaly, hot-end deterioration, and vibration faults.

Considering that there is a class imbalance phenomenon in the fault samples of heavy-duty gas turbines, the number of normal operating condition samples is usually significantly greater than that of abnormal samples. In the training stage, this paper adopts weighted cross-entropy as the objective function:

$$\mathcal{L} = - \sum_{i=1}^C \omega_i y_i \log \hat{y}_i \quad (14)$$

Among them, C represents the number of fault categories, y_i represents the true label, and ω_i represents the category weight. This processing method can appropriately reduce the contribution of small sample fault categories in the training process, alleviating the problem of the model being biased towards the majority class. At the same time, adding Dropout and layer normalization in the task head helps to alleviate the overfitting risk under the condition of small samples and enhances the output stability.

From the perspective of computer implementation, the fault diagnosis large model constructed in this paper is not simply stacking complex networks, but on the basis of maintaining the expression ability of the main model, it achieves efficient parameter transfer through LoRA low-rank modulation, and then uses the lightweight task head to complete the final classification. Such a structure takes into account training efficiency, inference speed and engineering deployability, and is more suitable for online monitoring and intelligent diagnosis scenarios of heavy-duty gas turbines. Overall, this model realizes a complete computational closed loop from input of multi-source operation data to output of fault categories, providing a unified model basis for subsequent experimental simulation and performance verification.

4 Experimental Simulation and Result Analysis

4.1 Experimental Environment and Dataset

To verify the applicability of the large model integrating LoRA low-rank modulation for heavy-duty gas turbine fault diagnosis under complex operating conditions, this paper completed model training and simulation analysis on a high-performance computing platform. The related experimental environment and dataset are presented in Table 1. The hardware environment is configured with multi-core server-level CPU, NVIDIA A100 GPU, and

large-capacity memory to support tasks such as long-time series sample loading, attention calculation, and low-rank modulation parameter update. The software environment uses Ubuntu 20.04 operating system, with Python 3.10 as the programming language, and PyTorch 2.1 as the deep learning framework, combined with CUDA and cuDNN for parallel acceleration. Considering that the model in this paper includes multiple computing steps such as time series encoding, LoRA adaptation, and multi-source feature fusion, during the training process, the AdamW optimizer is adopted, with a batch size of 64, an initial learning rate of 1×10^{-4} , and the learning rate is dynamically adjusted using the cosine annealing strategy to ensure a relatively stable convergence process.

The experimental data comes from a long-term operation monitoring platform and historical maintenance database of a certain type of heavy-duty gas turbine, with a time span from January 2021 to December 2024. The data content covers key operating parameters of the compressor, combustion chamber, turbine, and auxiliary control system, including intake and exhaust temperatures, pressure levels at each stage, rotational speed, gas flow rate, temperature dispersion of exhaust, vibration amplitude, lubricating oil state quantity, and control command feedback. The original data is cleaned, aligned, standardized, and segmented into sliding windows before being constructed into a unified time series sample set. To enhance the adaptability of the computer model to real industrial scenarios, multiple types of operating conditions such as steady-state operation, load variation, start-stop transition, and disturbance response are retained in the samples. The dataset is divided into training set, validation set, and test set in a ratio of 7:2:1. The fault categories include compressor degradation, combustion anomaly, heat-end component deterioration, vibration anomaly, and sensor anomaly. As shown in Table 2, the proportion of normal operating condition samples is relatively high, while the distribution of various fault samples is not balanced. To balance the characteristics of industrial data and the requirements of model training, this paper moderately retains fault samples during sample construction, ensuring that the dataset reflects the class imbalance characteristics in on-site operation and also meets the requirements for conducting comparative experiments.

Table 2: Overview of Experimental Environment and Dataset

Item	Configuration / Description
CPU	Intel Xeon Gold 6330, Multi-core Processor
GPU	NVIDIA A100 80GB
Memory	256GB
Operating System	Ubuntu 20.04
Programming Language	Python 3.10
Deep Learning Framework	PyTorch 2.1
CUDA/cuDNN	CUDA 12.1, cuDNN 8.9
Optimizer	AdamW
Batch Size	64
Initial Learning Rate	(1×10^{-4})
Data Time Range	January 2021 – December 2024
Total Number of Samples	96,000
Normal Samples	57,600
Compressor Degradation Samples	11,520
Combustion Anomaly Samples	9,600
Hot-End Component Degradation Samples	8,640
Vibration Anomaly Samples	5,760
Sensor Anomaly Samples	2,880
Dataset Split	Training Set: 67,200; Validation Set: 19,200; Test Set: 9,600

4.2 Evaluation Indicators Setting

To comprehensively assess the performance of the large model for fault diagnosis of heavy-duty gas turbines that incorporates LoRA low-rank modulation, this paper sets evaluation indicators from two aspects: diagnostic accuracy and computational cost. In the evaluation of fault diagnosis results, accuracy, recall rate, and macro-average F1 value are selected as the core indicators. Accuracy is used to measure the overall discriminative ability of the model for all samples, precision reflects the proportion of samples correctly identified as a certain type of fault among those actually classified as such, recall rate indicates the degree of model's adequate recognition of real fault samples, and F1 value comprehensively considers precision and recall, making it more suitable for evaluating classification performance under conditions of unbalanced class distribution. Considering that normal samples of heavy-duty gas turbines are usually significantly more than fault samples, this paper also calculates macro-average F1 to weaken the masking effect of the majority class on the results.

In addition to classification indicators, this paper also focuses on the computational efficiency of the model. In view of the characteristics of LoRA low-rank modulation, the number of trainable parameters and inference latency are additionally recorded to compare the differences in computational costs of different models beyond diagnostic accuracy. The reason for this setting is that online fault diagnosis of heavy-duty gas turbines not only requires reliable identification results but also requires the model to have good training efficiency and deployment feasibility. Through these indicators, this method can be evaluated more objectively from both the result performance and engineering application perspectives.

4.3 Model Selection for Comparison

To verify the effectiveness of the method proposed in this paper in the fault diagnosis task of heavy-duty gas turbines, the experiments selected comparison objects from three levels: traditional machine learning models, typical deep learning models, and parameter-efficient transfer models. The traditional models used were Support Vector Machine and Decision Tree. The Support Vector Machine is suitable for handling nonlinear classification problems with small to medium-sized samples and can learn the mapping relationship between temperature, pressure, vibration and fault categories; the Decision Tree has a relatively intuitive structure and can be used to observe the basic discriminative ability under multi-variable conditions.

In the deep learning and time series modeling models, CNN, LSTM and standard Transformer were selected as the comparison methods. CNN focuses on extracting local patterns and is suitable for identifying abnormal fluctuations within short windows; LSTM can handle the dependencies in time series and has a certain ability to depict the evolution of operating conditions; the standard Transformer emphasizes long-term sequence correlation modeling and can be used as a direct reference for the large model framework of this paper. In addition, this paper also set up a "full parameter fine-tuning model without introducing LoRA" as an ablation comparison to test the actual effect of low-rank modulation on parameter efficiency, training cost and diagnostic performance. Such a comparison setting can comprehensively reflect the differences of different computer models in the complex fault identification of heavy-duty gas turbines.

4.4 Experimental Results Presentation and Analysis

The fault diagnosis results of different models on the test set are shown in Table 3. Overall, the large model for heavy-duty gas turbine fault diagnosis proposed in this paper, which integrates LoRA low-rank modulation, achieved the best results in all main evaluation indicators. Specifically, the model accuracy reached 96.8%, which was 2.6 percentage points higher than

that of the standard Transformer (94.2%) and 5.5 percentage points higher than that of LSTM (91.3%). The recall rate reached 95.1%, which was 0.9 percentage points higher than that of the full-parameter fine-tuning large model (94.2%). This indicates that the method proposed in this paper has a more comprehensive coverage of real fault samples and a reduced rate of missed diagnoses. The macro-average F1 value reached 95.7%, which was 7.6, 5.7, and 2.6 percentage points higher than CNN, LSTM, and Transformer respectively. This shows that the model can maintain a relatively balanced classification performance under the condition of imbalanced class distribution. In terms of computational cost, the trainable parameters of this model are 9.6M, which is significantly lower than the 88.4M of the full-parameter fine-tuning large model. The parameter scale is approximately compressed by 89.1%. During inference, the latency is 11.2 ms/sample, which is slightly higher than traditional SVM and decision trees, but lower than that of the standard Transformer (14.1 ms/sample) and the full-parameter fine-tuning large model (13.8 ms/sample). This indicates that LoRA low-rank modulation does not sacrifice efficiency for accuracy; instead, it achieves a better balance between parameter efficiency and diagnostic performance.

Table 3: Comparison of Fault Diagnosis Results of Different Models

Model	Accuracy / %	Recall / %	Macro-F1 / %	Trainable Parameters / M	Inference Latency / ms·sample ⁻¹
SVM	83.6	79.8	80.9	—	5.4
DT	81.9	78.6	79.7	—	3.8
CNN	89.7	87.4	88.1	6.2	8.7
LSTM	91.3	89.1	90.0	8.5	10.6
Transformer	94.2	92.7	93.1	31.4	14.1
Fully Fine-Tuned Large Model	96.1	94.2	94.8	88.4	13.8
Proposed Model	96.8	95.1	95.7	9.6	11.2

From the experimental results of different sample sizes, it can be seen that the model accuracy increases overall as the training data increases. The specific changes are shown in Figure 4. When the sample size is 12,000, the accuracy of the model in this paper has reached 86.1%, which is higher than 83.5% of Transformer and 79.8% of LSTM. When the sample size increases to 36,000, the accuracy of the model in this paper rises to 93.1%, which is 7.0 percentage points higher than the initial stage, while Transformer and LSTM increase to 90.1% and 87.0% respectively. When the sample size reaches 48,000, the model in this paper has reached 95.0%, and then rises to 96.2% at 60,000 and 72,000 sample conditions, with the increase rate gradually slowing down. In contrast, Transformer is 94.2% at 72,000 samples, and LSTM is 91.3%. Both still show a continuous dependence on larger sample sizes. Thus, the model in this paper has already demonstrated stable fault representation capabilities at the medium-sized data stage. This is because the fault semantic representation module improves the efficiency of feature organization, LoRA low-rank modulation reduces the parameter update cost required for domain adaptation, and multi-source feature fusion enhances the density of effective information, enabling the model to form clearer classification boundaries even with relatively limited samples.

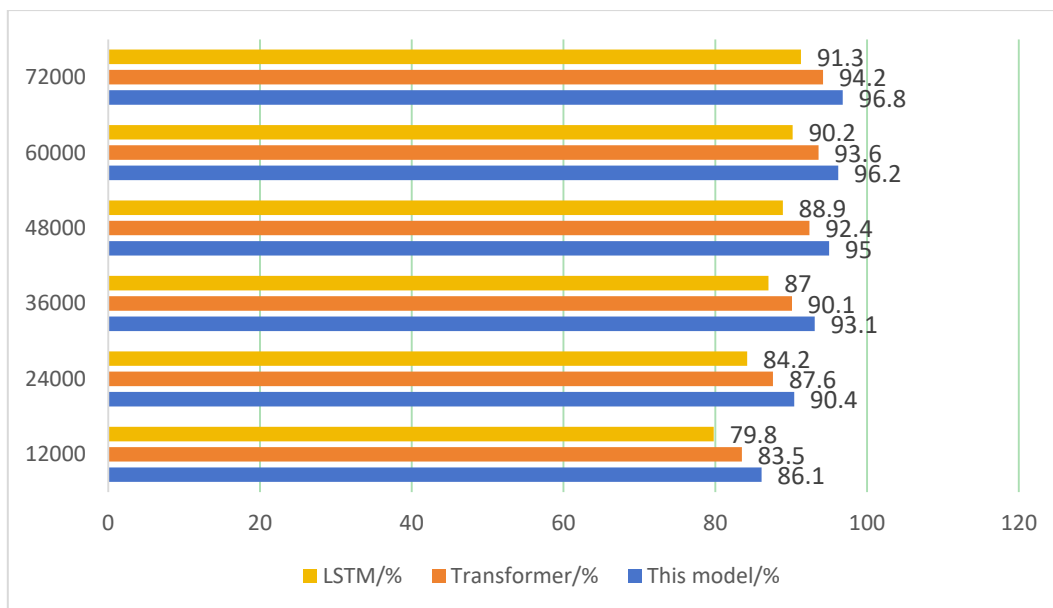


Figure 4: Changes in model accuracy under different sample sizes

The variations of macro average F1 under different operating conditions are shown in Figure 5. Under the steady-state full-load condition, the macro average F1 of the model in this paper is 96.0%, which is 2.5 percentage points higher than that of Transformer (93.5%) and 5.2 percentage points higher than that of LSTM (90.8%). During the transition stages of load increase and decrease, the model in this paper reaches 95.2% and 94.8% respectively, remaining at a relatively high level, while Transformer drops to 92.6% and 91.9%, and LSTM further decreases to 88.9% and 88.1%. In the more complex dynamic scenario of start-stop transient, the macro average F1 of this model is 93.9%, which is 2.1 percentage points lower than the steady-state condition, but still significantly higher than that of Transformer (89.8%) and LSTM (85.4%). When sensor noise interference and missing data perturbations are introduced, the macro average F1 of this model is 92.8% and 91.7% respectively, which is 3.2 and 4.3 percentage points lower than the steady-state full-load condition; corresponding to Transformer, it drops to 87.1% and 85.6%, and LSTM falls to 82.6% and 80.9%. This indicates that the model still has good robustness under conditions of noise pollution, data drift, and local missing data. The reason is that the fault semantic representation module can retain the context relationship of abnormal evolution, LoRA modulation enables the model to form parameter responses closer to the characteristic distribution of the gas turbine under different operating conditions, and the multi-source fusion mechanism reduces the misjudgment risk caused by the distortion of a single variable.

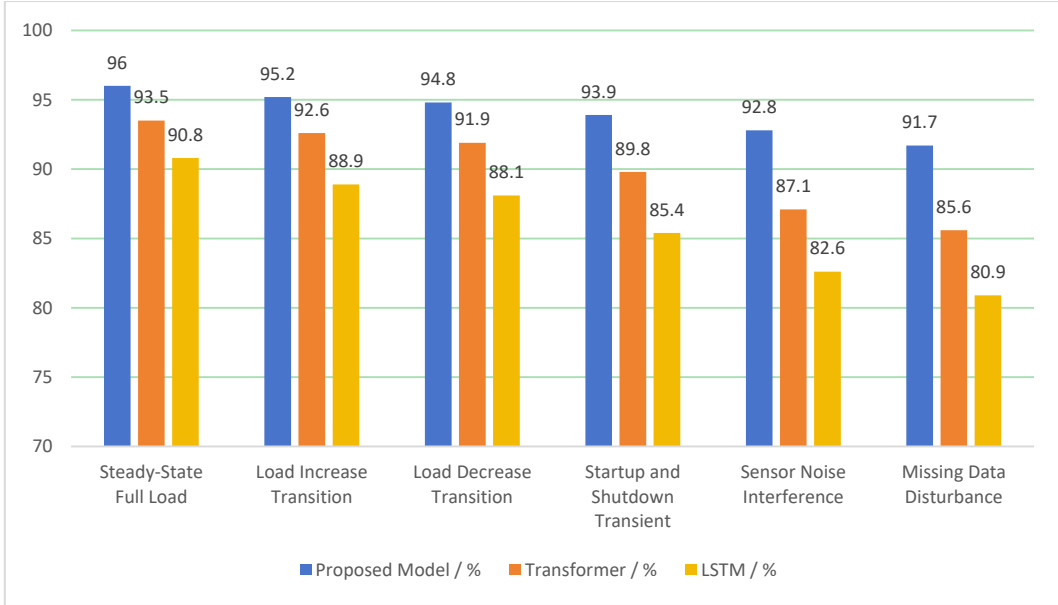


Figure 5: Changes in the macro average F1 value of the model under different operating conditions

From Table 3, Figure 4 and Figure 5, it can be seen that the advantages of the method proposed in this paper are not only reflected in higher single-test accuracy, but also in stronger adaptability to changes in sample size and disturbances in operating conditions. For industrial systems such as heavy-duty gas turbines with complex operating boundaries and long abnormal propagation chains, this model structure that takes into account both accuracy, parameter efficiency and robustness is closer to the actual requirements of online diagnosis scenarios.

5 Conclusion

This paper addresses the practical issues in the fault diagnosis of heavy-duty gas turbines, such as strong coupling of multi-source data, rapid changes in operating conditions, scarcity of fault samples, and high training costs of traditional deep models. A fault diagnosis large model integrating LoRA low-rank modulation was constructed. The research was conducted from aspects such as data collection during operation, fault semantic representation, multi-source feature extraction and fusion, and low-rank modulation adaptation, forming a complete computational framework for complex time-series data of heavy-duty gas turbines. Experimental results show that the model in the test set achieves an accuracy of 96.8%, a recall rate of 95.1%, and a macro-average F1 value of 95.7%, outperforming the comparison models such as SVM, decision tree, CNN, LSTM, and standard Transformer. At the same time, while improving the diagnostic performance, the trainable parameter quantity has been reduced from 88.4M in the full-parameter fine-tuning model to 9.6M, significantly compressing the parameter update scale, indicating that LoRA low-rank modulation can improve the domain adaptation efficiency and engineering deployment feasibility of heavy-duty gas turbine scenarios while retaining the time-series representation ability of the large model. The research also shows that this method maintains good stability under conditions such as start-stop transient, noise interference, and missing data perturbation, but still has issues such as high computational cost and the need for further enhancement of generalization ability in extreme operating condition switching, cross-unit migration, and long-term continuous monitoring scenarios. Future work can combine model pruning, quantization compression, and

cross-domain transfer learning methods to optimize the diagnostic large model for lightweighting, and introduce more abundant mechanism constraints and online incremental update mechanisms to further enhance the generalization ability and real-time application level of the model in different models and different operating boundaries.

About the Author

Ruan Hang(ID:230804197810010539), male, ethnic Han, born in Qiqihar City, Heilongjiang province, holds a bachelor's degree and the title of Senior Engineer, currently employed at CNOOC Gas and Power Group Co., Ltd., specializing in petrochemical production informatization, digital intelligence of gas power plants, and other intelligent construction.

Sun nailiang(ID:321322198402167214), male, ethnic han, born in Huaiyin City, Jiangsu province, holds a bachelor's degree and the title of Senior engineer, currently employed at CNOOC Gas and Power Group Co., Ltd., specializing in the digital and intelligent transformation of management and operations within the petrochemical industry.

He hong(ID:230722199010100823), female, ethnic Han, born in Yichun City, Heilongjiang province, holds a master's degree and the title of engineer, currently employed at CNOOC Gas and Power Group Co., Ltd., specializing in the construction of smart power plants and data governance.

References

- [1] Xiaofeng L I U, Yingjie C, Jianhua W, et al. Intelligent fault diagnosis methods toward gas turbine: A review[J]. Chinese Journal of Aeronautics, 2024, 37(4): 93-120. <https://doi.org/10.1016/j.cja.2023.09.024>
- [2] Tahan M, Tsoutsanis E, Muhammad M, et al. Performance-based health monitoring, diagnostics and prognostics for condition-based maintenance of gas turbines: A review[J]. Applied energy, 2017, 198: 122-144. <https://doi.org/10.1016/j.apenergy.2017.04.048>
- [3] Fentaye A D, Baheta A T, Gilani S I, et al. A review on gas turbine gas-path diagnostics: State-of-the-art methods, challenges and opportunities[J]. Aerospace, 2019, 6(7): 83. <https://doi.org/10.3390/aerospace6070083>
- [4] Zaccaria V, Rahman M, Aslanidou I, et al. A review of information fusion methods for gas turbine diagnostics[J]. Sustainability, 2019, 11(22): 6202. <https://doi.org/10.3390/su11226202>
- [5] Kim T S. Model-based performance diagnostics of heavy-duty gas turbines using compressor map adaptation[J]. Applied energy, 2018, 212: 1345-1359. <https://doi.org/10.1016/j.apenergy.2017.12.126>
- [6] Liu Y. Design of fault detection system for a heavy duty gas turbine with state observer and tracking filter[C]//Turbo Expo: Power for Land, Sea, and Air. American Society of Mechanical Engineers, 2017, 50916: V006T05A017. <https://doi.org/10.1115/GT2017-64089>

- [7] Mirhosseini A M, Adib Nazari S, Maghsoud Pour A, et al. Probabilistic failure analysis of hot gas path in a heavy-duty gas turbine using Bayesian networks[J]. *International Journal of System Assurance Engineering and Management*, 2019, 10(5): 1173-1185. <https://doi.org/10.1007/s13198-019-00848-z>
- [8] Liu J F, Zhu L H, Ma Y J, et al. Anomaly detection of hot components in gas turbine based on frequent pattern extraction[J]. *Science China Technological Sciences*, 2018, 61(4): 567-586. <https://doi.org/10.1007/s11431-017-9165-7>
- [9] Fentaye A D, Ul-Haq Gilani S I, Baheta A T, et al. Performance-based fault diagnosis of a gas turbine engine using an integrated support vector machine and artificial neural network method[J]. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 2019, 233(6): 786-802. <https://doi.org/10.1177/0957650918812510>
- [10] Zhong S, Fu S, Lin L. A novel gas turbine fault diagnosis method based on transfer learning with CNN[J]. *Measurement*, 2019, 137: 435-453. <https://doi.org/10.1016/j.measurement.2019.01.022>
- [11] Zhou D, Yao Q, Wu H, et al. Fault diagnosis of gas turbine based on partly interpretable convolutional neural networks[J]. *Energy*, 2020, 200: 117467. <https://doi.org/10.1016/j.energy.2020.117467>
- [12] Yang X, Bai M, Liu J, et al. Gas path fault diagnosis for gas turbine group based on deep transfer learning[J]. *Measurement*, 2021, 181: 109631. <https://doi.org/10.1016/j.measurement.2021.109631>
- [13] Bai M, Liu J, Long Z, et al. A comparative study on class-imbalanced gas turbine fault diagnosis[J]. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 2023, 237(3): 672-700. <https://doi.org/10.1177/09544100221107252>
- [14] Sarwar U, Muhammad M, Mokhtar A A, et al. Hybrid intelligence for enhanced fault detection and diagnosis for industrial gas turbine engine[J]. *Results in Engineering*, 2024, 21: 101841. <https://doi.org/10.1016/j.rineng.2024.101841>
- [15] Xu J, Wu H, Wang J, et al. Anomaly transformer: Time series anomaly detection with association discrepancy[J]. *arXiv preprint arXiv:2110.02642*, 2021. <https://doi.org/10.48550/arXiv.2110.02642>
- [16] Nie Y, Nguyen N H, Sinthong P, et al. A time series is worth 64 words: Long-term forecasting with transformers[J]. *arXiv preprint arXiv:2211.14730*, 2022.
- [17] Liang Y, Wen H, Nie Y, et al. Foundation models for time series analysis: A tutorial and survey[C]//*Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 2024: 6555-6565. <https://doi.org/10.1145/3637528.3671451>
- [18] Zhong G, Liu F, Jiang J, et al. Refining one-class representation: A unified transformer for unsupervised time-series anomaly detection[J]. *Information Sciences*, 2024, 656: 119914. <https://doi.org/10.1016/j.ins.2023.119914>

- [19] Han Z, Gao C, Liu J, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey[J]. arXiv preprint arXiv:2403.14608, 2024. <https://doi.org/10.48550/arXiv.2403.14608>
- [20] Zhao Y, Zhou M, Zhang N, et al. Fault diagnosis of gas turbine based on matrix capsules with EM routing[J]. Systems Science & Control Engineering, 2021, 9(sup1): 96-102. <https://doi.org/10.1080/21642583.2020.1833783>