Digital Intelligence PRESS

Journal of Digital Frontier

https://doi.org/10.67541/jdf2606

Research Article **Volume 1, Issue 1 (July 2026)** *Open Access*

Power Transformer Fault Diagnosis and Prediction Based on Deep Learning

Haoyu You *, **Longtai Hua**, **Tianyi Su**

Department of Electrical and Information Engineering, Shandong University of Science and Technology, Jinan, Shandong, China

***Corresponding author:** Haoyu You. Email: 19953322591@163.com

Abstract

Power transformers assume the core functions of power conversion and power flow regulation in the new power system, and its fault outage will lead to the chain risk of power grid. However, the diagnostic accuracy of traditional threshold methods and shallow machine learning is limited in the face of non-stationary and strong-noise1 signals, and the data distribution deviation caused by variable load conditions seriously restricts the field generalization ability of the model. To solve the above bottleneck, this paper proposes a complete technical system that integrates multi-domain feature enhancement, dual-branch domain-adaptive diagnosis, and temporal adversarial prediction. In the preprocessing stage, an adaptive denoising strategy based on improved variational mode decomposition and permutation entropy is designed, and the time-frequency-energy 3D multi-channel input tensor is constructed by using the synchrosqueezed S-transform (SSST) combined with kurtosis, peak factor, and energy spectrum entropy, which effectively solves the problem of poor separability of fault features under strong noise. In the diagnosis stage, a dual-branch adversarial domain-adaptive convolutional network is proposed. With dilated EfficientNet as the backbone, global semantic and local sensitivity dual branches are set to extract wideband and transient features in parallel, and the dual-domain adaptation mechanism of maximum mean discrepancy and an adversarial discriminator is introduced to achieve high-precision diagnosis in the unlabeled target domain under variable working conditions. The experimental results show that the accuracy of the model reaches 94.7% in cross-operating-condition scenarios, which is 8.2 percentage points higher than that of the traditional domain-adversarial network. In the prediction stage, a temporal generative adversarial network is constructed, which takes dilated causal convolution and self-attention spatio-temporal gating as its core, combines Monte Carlo dropout interval estimation guided by dynamic time warping morphological loss and discriminator confidence, and realizes probabilistic prediction of the remaining useful life (RUL), and the prediction interval coverage rate reaches 88.6%. The method in this paper has achieved significant performance improvement in the three dimensions of diagnosis accuracy, prediction reliability and computational efficiency, and provides complete technical support for the intelligent operation and maintenance of power transformers from state recognition to trend prediction.

Keywords

Power transformer; Fault diagnosis; Remaining life prediction; Generative adversarial network; Temporal convolutional network

Received: 31 May 2026; Revised: 19 Jul 2026; Accepted: 29 Jul 2026; Published: 22 Aug 2026

1. INTRODUCTION

As the core hub equipment for voltage transformation and power regulation in the new power

system, the safe and stable operation of the power transformer is directly related to the power supply reliability and economy of the power grid. Driven by the “dual carbon” goal, the large-scale access of new energy and the rapid development of the UHV AC/DC hybrid power grid have made the transformer operating environment increasingly complex. Frequent load fluctuations, harmonic injections, DC bias and other factors have continuously aggravated the risk of internal insulation aging and mechanical structure damage [1],[2]. Transformer failure and shutdown not only lead to high direct replacement costs but may also trigger a chain of power grid accidents, causing large-scale power outages and serious social and economic losses. According to statistics, transformer failure is the primary cause of unplanned shutdown of substations, of which winding deformation, core loosening, partial discharge and overheating failures account for more than 75% of the total failure types. However, the above-mentioned failures often manifest as the offset of weak side frequency components in the vibration spectrum or the intermittent appearance of partial discharge pulses in the early latent stage. The traditional protection strategy based on fixed threshold can only be triggered when the failure develops to a more serious degree, which is difficult to provide sufficient early warning for operation and maintenance personnel [3]. Meanwhile, traditional shallow machine learning methods, represented by support vector machines and backpropagation neural networks, rely heavily on human experience for feature extraction. The designed statistical features are extremely sensitive to changes in operating conditions—when the load rate jumps from light load to heavy load or the oil temperature fluctuates drastically, the feature distribution of the same fault type shifts significantly, leading to a sharp drop in accuracy when the offline-trained diagnostic model is deployed online [4].

For transformer fault diagnosis, a long-standing research hotspot, scholars at home and abroad have conducted extensive research, and its technological evolution path can be roughly divided into three stages. Early research focused on signal processing and statistical decision-making as the main framework. Typical methods included wavelet packet transform to extract the energy ratio of each frequency band, and Hilbert-Huang transform to obtain time-frequency entropy and input it into fuzzy clustering for classification decisions. These methods have clear physical meaning and strong interpretability, but their performance is severely limited by the pre-selected basis functions and the number of decomposition layers [5]. When facing strong background noise, the contradiction between time resolution and frequency resolution is difficult to reconcile, and the classification accuracy has a clear upper limit [6]. Subsequently, classic machine learning methods, represented by BP neural networks and support vector machines, were introduced into this field, using their nonlinear mapping capabilities to replace manual discrimination rules, which improved the level of diagnostic automation to a certain extent. However, BP networks are sensitive to initial weights and prone to getting trapped in local extrema, while SVMs face voting conflicts in a "one-to-many" strategy when dealing with multiple types of faults. Both methods require manual design of feature descriptors, and feature quality directly determines the upper limit of diagnostic performance [7].

In recent years, the rapid development of deep learning has propelled transformer fault diagnosis into a new stage of end-to-end learning. Convolutional neural networks, with their powerful local feature extraction capabilities, have been used to directly learn fault modes from time-spectrum graphs, while long short-term memory networks and gated recurrent units are used to capture the temporal dependencies in monitoring signals [8],[9],[10]. Some researchers have further combined the two into a CNN-LSTM serial architecture, simultaneously modeling spatial and temporal dimensions. In addition, hybrid models of temporal convolutional networks and XGBoost, multi-scale prediction architectures based on Transformers, CNN-BiLSTM spatiotemporal modeling methods, and the application of graph neural networks in capturing complex system relationships all demonstrate the

enormous potential of different neural network paradigm combinations in complex prediction tasks [11],[12],[13],[14].

Despite significant progress, existing deep models still have two prominent shortcomings. First, most methods employ single-path serial or parallel feature extraction structures. Once the convolution kernel size is fixed, they cannot simultaneously adapt to two vastly different feature patterns: broadband global frequency shift faults and transient partial discharge pulses. Furthermore, the fusion of multi-source heterogeneous features lacks a flexible dynamic control mechanism. Second, existing models lack effective means to capture the subtle signs of early fault latency. Most prediction studies remain at the regression fitting level on short timescales, failing to provide a quantitative uncertainty assessment of remaining lifetime, thus failing to meet the urgent needs of on-site maintenance for probabilistic repair decision support.

To address these bottlenecks, this paper conducts a systematic study focusing on three levels: signal preprocessing, cross-condition adaptive diagnosis, and temporal evolution prediction. The aim is to achieve a complete technical closed loop from accurate identification of the current state to reliable prediction of future trends. In the preprocessing stage, an adaptive denoising strategy based on improved variational mode decomposition and permutation entropy is designed, and a time-frequency-energy multi-channel input tensor is constructed in conjunction with synchronous compressed S-transform to improve the separability of fault features under strong noise backgrounds. In the diagnostic phase, a dual-branch adversarial domain adaptive convolutional network is proposed. This network extracts broadband frequency shift and transient pulse features through parallel global semantic and locally sensitive dual branches. A dual-domain adaptive mechanism, combining maximum mean difference and an adversarial discriminator, is introduced to achieve high-precision diagnosis under varying operating conditions without target domain labeling. In the prediction phase, a temporal generative adversarial network is constructed, centered on dilated causal convolution and self-attention spatiotemporal gating. This network, combined with dynamic temporal warping likelihood loss and discriminator confidence-guided Monte Carlo dropout interval estimation, enables probabilistic prediction of remaining lifetime.

2. PREPROCESSING AND MULTI-DOMAIN FEATURE ENHANCEMENT ALGORITHM FOR NON-STATIONARY SIGNALS

The vibration signal and current signal collected during the operation of power transformers often show significant non-stationary and nonlinear characteristics, which are rooted in load fluctuation, environmental temperature change, harmonic interference and time-varying characteristics of the equipment's own mechanical structure. These complex factors lead to the fault feature information being submerged by strong background noise, which seriously restricts the diagnostic accuracy and generalization ability of subsequent deep learning models. Therefore, this section focuses on the construction of a complete set of signal preprocessing and multi-domain feature enhancement system, covering three closely coupled links of adaptive denoising segmentation, high-resolution time-frequency transformation and data equalization, aiming to provide high-quality and high information density input tensors for subsequent diagnosis and prediction models.

The preprocessing quality of the original signal directly determines the upper limit of effectiveness of feature extraction. The vibration signal of the transformer box is mainly composed of the coupling of the magnetostriction of the core, the electrodynamic force of the winding and the mechanical

vibration of the cooling system. Its spectral components cover a wide band from the power frequency of 50 Hz to several kHz, while the UHF current pulse generated by the partial discharge is superimposed on the power frequency current, forming a sparse transient impact in the time domain. Traditional filtering methods are difficult to effectively suppress wideband noise while retaining fault transient components. Therefore, variational mode decomposition (VMD) algorithm is introduced to realize adaptive signal deconvolution [15],[16],[17]. The core idea of VMD is to decompose the original input signal $f(t)$ into the sum of several intrinsic mode functions $u_k(t)$ with sparse characteristics. Each mode is narrow band distributed around its central frequency ω_k, which is realized by iteratively solving the following constrained variational problem:

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \text{ s.t. } \sum_{k=1}^{K} u_k(t) = f(t) \tag{1}$$

Where K is the preset total number of mode decompositions, $\delta(t)$ is the Dirac delta function, $*$ represents the convolution operation, j is the imaginary unit, and ∂_t is the first-order partial derivative operator with respect to time t. The quadratic penalty factor α and Lagrange multiplier $\lambda(t)$ are introduced to transform the above constrained problem into an unconstrained augmented Lagrange form, and each mode and its center frequency are updated iteratively in the frequency domain by alternating direction multiplier method (ADMM) [18],[19]. However, the traditional VMD algorithm needs to fix the mode number K in advance. If the value is too small, it will lead to mode aliasing and underdecomposition, while if the value is too large, it will introduce false modes, both of which will seriously distort the eigenstructure of fault features. To solve this problem, this paper proposes an adaptive mode optimization strategy based on Permutation Entropy (PE). For each of the decomposed eigen mode component u_k, embedded in the reconstruction of phase space sequence:

$$X_i = \{u_k,(i), u_k,(i+\tau) \ldots u_k(i+(m-1)\tau)\} \tag{2}$$

Where m is the embedding dimension and τ is the delay time. Each embedding sequence is arranged in ascending order of numerical size to obtain m! There are possible arrangement modes π_r, and the probability of occurrence of each mode is denoted as $p(\pi_r)$, then the arrangement entropy of this mode is defined as:

$$H_{PE}(u_k) = -\sum_{r=1}^{m!} p\,(\pi_r) \ln p\,(\pi_r) \tag{3}$$

Permutation entropy value $\bar{H}_{PE} = H_{PE}/\ln(m!)$ after normalization If it falls in the interval of [0,1], the closer the value is to 1, the stronger the randomness of the sequence is and the higher the proportion of noise components is. The closer the value is to 0, the more regular oscillation of the sequence is, and the deterministic fault information is contained. In this paper, the permutation entropy threshold $\eta = 0.65$ is set, and the components $\bar{H}_{PE} \geq \eta$ are judged as noise dominant modes and eliminated, while the components $\bar{H}_{PE} < \eta$ are retained as effective eigenmodes. Accordingly, starting from $K = 4$, the search is incremental. When the permutation entropy of the new modes continues to be higher than the threshold and the total number of retained modes does not change after continuously increasing K, the K is determined to be the optimal number of decomposition layers. This adaptive mechanism effectively solves the subjectivity problem of artificial preset K, and at the same time realizes the quantitative separation of noise and effective components.

After adaptive denoising and mode screening, each retained eigenmode component is reconstructed into a pure signal $s(t)$, which has been filtered out of wideband background noise and stray interference. However, the time domain waveform itself has limited discrimination power for fault types, and different defect forms such as transformer winding deformation, core loosening, partial discharge and overheating fault have different spectral distribution, energy attenuation law and timely statistical characteristics. Therefore, it is urgent to build a multi-dimensional feature representation space to fully disclose the nature of faults. In this paper, the traditional short-time Fourier transform (STFT) is limited by the Heisenberg uncertainty principle, and the Synchrosqueezed S-Transform (SSST) is used to construct the high-resolution time spectrograph [20]. S transform builds a bridge between short-time Fourier transform and wavelet transform, and its time width and frequency are adjusted adaptively. Its definition is as follows:

$$S(\tau, f) = \int_{-\infty}^{+\infty} s(t) \frac{|f|}{\sqrt{2\pi}} \exp\left(-\frac{f^2(\tau - t)^2}{2}\right) \exp(-j2\pi f t)\, dt \tag{4}$$

Where τ is the central position parameter of the Gaussian window function, which controls the translation of the window function on the time axis; f is the frequency variable, $|f|/\sqrt{2\pi}$ is the amplitude normalization factor to ensure that the window function energy is consistent at different frequencies. The Gaussian window width automatically shrinks with the increase of f, obtaining good frequency resolution in low frequency band and good time resolution in high frequency band. Calculate the instantaneous frequency for each time-frequency point (τ, f):

$$\hat{f}_s(\tau, f) = f + \frac{1}{j2\pi S(\tau, f)} \frac{\partial S(\tau, f)}{\partial \tau} \tag{5}$$

Synchronous compression operation will (τ, f) within the neighborhood of the extrusion to time-frequency coefficient to the instantaneous frequency $\hat{f}_s(\tau, f)$ as the center of the narrow frequency band, namely by the mapping $(\tau, f) \to (\tau, \hat{f}_s(\tau, f))$ realize redistribution of energy, its compression transform are expressed as:

$$T_s(\tau, \eta) = \int_{\left\{f : |\hat{f}_s(\tau, f) - \eta| \le \frac{\Delta f}{2}\right\}} S(\tau, f)\, df \tag{6}$$

Where η is the new frequency variable after compression, and Δf is the frequency discretization step. Through this compression operation, the time-frequency energy dispersed by window function leakage in the original S-transform spectrum is refocused near the real instantaneous frequency trajectory, and the frequency resolution is significantly improved to the theoretical limit of wavelet transform, especially suitable for capturing the ultra-wideband frequency component in the partial discharge pulse with a duration of several microseconds.

Although the time-dependent spectrogram can present the evolution trajectory of signal frequency components with time, it cannot fully quantify the statistical distribution characteristics and energy concentration degree in non-stationary signals, which have unique discrimination power in distinguishing mechanical faults from electrical faults. Therefore, based on the 3D time-frequency tensor generated by SSST, this paper extracts the time-domain statistical features and the frequency-domain energy spectrum entropy features in parallel to form a multi-domain complementary feature set. In the time domain, the Kurtosis index sensitive to the shock pulse and the Crest Factor sensitive to the amplitude fluctuation are selected. Let the discrete sampling sequence of the reconstructed signal $s(t)$ in the analysis window be $\{x_1, x_2, \ldots, x_N\}$, N sampling points to the window and the meaning of:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7}$$

The standard deviation for:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{8}$$

The kurtosis is defined as the square of the fourth order center moment and variance ratio:

$$Kurtosis = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \tag{9}$$

The index is highly sensitive to weak impact pulse, and the kurtosis value of normal vibration signal is about 3, while the kurtosis value will be significantly increased to more than 5 due to the impact component caused by partial discharge or winding loosening. Peak factor is defined as the ratio of signal peak value to effective value:

$$CrestFactor = \frac{\max|x_i|}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}} \tag{10}$$

This index reflects the relative strength of transient shocks in the signal and can provide normalized amplitude fluctuation information under variable load conditions. In the frequency domain, in order to quantify the uniformity of Energy distribution of each frequency component, this paper calculates the Energy Spectral Entropy. Let the discrete spectrum of signal $s(t)$ obtained by fast Fourier transform be $X(l),\ l = 1,2,\ldots,L,\ L$, each frequency point energy accounted for the proportion of the total energy for:

$$p_l = |\frac{X(l)|^2}{\sum_{r=1}^{L}|X(r)|^2} \tag{11}$$

The energy spectrum entropy is defined as:

$$H_{ESE} = -\sum_{l=1}^{L} p_l \ln p_l \tag{12}$$

The larger the energy spectrum entropy value is, the more scattered the spectrum energy distribution is and the wider the frequency band is, which is closely related to the abundant high-frequency harmonic components generated by partial discharge. A smaller entropy value indicates that the energy is concentrated near the power frequency and its lower harmonics, which is consistent with the characteristic frequency shift caused by the mechanical deformation of the winding. After the above transformation and extraction, a three-dimensional multi-channel input tensor with dimension $H \times W \times 3$ is constructed in this paper, where H and W are the height (frequency resolution) and width (time resolution) of the time spectrogram respectively, and the three channels in the third dimension correspond to the SSST time spectrogram, kurtosis - peak factor joint statistical map and energy spectrum entropy distribution map respectively. The fault information at the same time can be jointly expressed in time-frequency-energy domains.

In order to realize the effective fusion and visual presentation of the above multi-domain features,

Table 1 statistics the typical value range of each time domain index and frequency domain entropy value under the typical fault state of transformer. The data are derived from repeated measurements of four typical states by the experimental platform under controlled operating conditions, and each measurement is averaged over 10 analysis Windows.

Table 1. Typical statistical values of time domain statistical indexes and energy spectrum entropy under different fault states

State of operation	Mean kurtosis	Factor of peak	Energy spectrum entropy	Kurtosis fluctuation range	Number of sample Windows
Normal state of affairs	3.12	2.87	3.24	2.96~3.41	50
The wind is slightly deformed	4.86	3.52	2.91	4.23~5.67	45
Loose iron core	5.73	4.18	3.86	5.02~6.85	42
Partial discharge	8.41	5.36	5.12	7.15~9.88	40

From the data in Table 1, it can be clearly observed that the vibration signal under normal state approximately follows the Gaussian distribution, and its kurtosis mean is maintained around 3.12 with a narrow fluctuation range. Both peak factor and energy spectrum entropy are at a low level, which indicates that the signal energy is concentrated and there is no significant impact component. When the windings are slightly deformed, the geometric structure changes of the core and the windings make the transmission path of the magnetostrictive force distorted, and the periodic asymmetric impact appears in the vibration waveform. The mean kurtosis jumps to 4.86, and the energy spectrum entropy decreases to 2.91. The reason is that the characteristic frequency components caused by the deformation are concentrated in a narrow frequency band, which makes the spectrum distribution tend to be concentrated. The loosening state of the core shows a completely different characteristic pattern, which leads to the macroscopic mechanical vibration of the core stack in the alternating magnetic field. The nonlinear energy transfer process excites rich frequency doubling and fractional frequency doubling components, and the energy spectrum entropy jumps to 3.86, and the peak factor is as high as 4.18. The PD state presents the most extreme index characteristics among all fault types. The discharge pulse at the rising edge of nanosecond level produces strong sparse spike sequence in the time domain, with the mean kurtosis up to 8.41 and the largest fluctuation range. Meanwhile, the UWB pulse stimulates the continuous spectrum from hundreds of kHz to tens of MHz, and the energy spectrum entropy reaches 5.12. It fully reflects the dual abnormal properties of the state in time domain and frequency domain. The above quantitative differences prove the necessity and effectiveness of multi-domain feature fusion.

After the high-dimensional feature space is fully represented, the quality of training data and class balance become the key constraints affecting the performance of deep learning models. In the actual power system, transformers run in normal state for a long time, so it is extremely difficult to obtain fault samples, especially early fault samples. As a result, the proportion of normal samples in the training set and all kinds of fault samples is greatly different, and the model will have serious class preference bias in the training process, and the diagnostic recall rate of minority faults is extremely low. To solve this problem, this paper designs a set of hybrid data enhancement and generation strategies. Firstly, two augmentation operations, time domain stretching and frequency domain masking, are implemented at the signal level. The time-domain stretching is realized by linear interpolation and resampling of the

original signal $s(t)$ along the time axis. The stretching factor $\beta \in [0.85,1.15]$ is used to map the sampling time t_i to $t_i' = \beta \cdot t_i$, and the amplitude of the new time point is obtained by cubic spline interpolation:

$$s_{stretch}(t_i') = \text{spline}(\{t_i, s(t_i)\}, t_i') \tag{13}$$

This operation simulates the time-axis stretching effect of signal waveform caused by load change without changing its spectral skeleton. In the frequency domain masking, the center frequency f_c and bandwidth B_w are randomly selected on the spectrum amplitude spectrum $|X(f)|$, and the attenuation factor $\gamma \in [0.3,0.8]$ is applied to the amplitude in the frequency band:

$$|X_{mask}(f)| = |X(f)| \cdot \left[1 - \gamma \cdot \exp\left(-\frac{(f - f_c)^2}{2B_w^2}\right)\right] \tag{14}$$

This operation effectively simulates the situation where a specific frequency component is attenuated by the channel during propagation and enhances the robustness of the model against frequency loss. The combination of the two operations increases the sample of each type of fault to 6 times the original, but the essence of such enhancement is still from the linear transformation of the limited original sample, and the new variability introduced by it is limited, which is difficult to cover the high-dimensional distribution space of the real site.

In order to fundamentally solve the sparsity and class imbalance problems of sample distribution, this paper further introduces Wasserstein generative adversative network with gradient penalty (WGAN-GP) to generate high-quality virtual fault samples. The core improvement of WGAN-GP lies in using Earth-Mover distance (Wasserstein-1 distance) to replace Jencen-Shannon divergence in traditional GAN, which fundamentally alleviates the problems of gradient disappearance and mode collapse in the training process. Let the true sample distribution be $\mathbb{P}_r$, and the generator G maps the latent variable $z \sim p_z$ (usually standard normal distribution) to the generated sample $\tilde{x} = G(z)$ with the generated distribution $\mathbb{P}_g$, then the Wasserstein distance is defined as follows:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \sim \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \| x - y \| \tag{15}$$

Where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ represents the set of all joint distributions with $\mathbb{P}_r$ and $\mathbb{P}_g$ as marginal distributions. Through the Kantorovich Rubinstein duality theorem, the distance can be transformed into the following computable form:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|D\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] \tag{16}$$

Where the discriminator D is constrained to be a 1-Lipschitz continuous function, that is, its gradient norm does not exceed 1 everywhere. To enforce this constraint, WGAN-GP introduces a gradient penalty term into the loss function and imposes a gradient norm penalty on the linear interpolation point $\hat{x} = \epsilon x + (1 - \epsilon)\tilde{x}$ (where $\epsilon \sim U[0,1]$) between the real sample and the generated sample. The complete alternating optimization objective of generator and discriminator is as follows:

$$\mathcal{L}_D = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2] \tag{17}$$

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}[D(G(z))] \tag{18}$$

Where λ is the gradient penalty coefficient, this paper takes $\lambda = 10$, and $\mathbb{P}_{\hat{x}}$ represents the distribution of the interpolation point $\hat{x}$. The generator G adopts deconvolution architecture, and its

input is 100-dimensional Gaussian noise $z \sim \mathcal{N}(0, I_{100})$. After four layers of deconvolution, it is gradually upsampled to be consistent with the dimension of the real multi-domain feature tensor. Discriminator D adopts a five-layer convolution structure, and the final output is scalar score rather than binary probability. During training, every time the generator weight is updated, the discriminator is updated five times to ensure that the discriminator always provides reliable gradient guidance. After WGAN-GP generates 200 virtual samples for each fault category, the number of samples for each fault category in the training set is increased to 700 samples equivalent to the normal category, and the category ratio is balanced from the original 1:12.5 (fault: normal) to nearly 1:1.

Table 2. Comparison of the number of training samples of various categories before and after data enhancement and equalization

Categories	Original sample number	Time domain stretch amplification	Frequency domain masking amplification	WGAN-GP generation	Final sample number
Normal	700	—	—	—	700
Deformation of winding	56	168	168	200	592
Loose iron core	48	144	144	200	536
Partial discharge	38	114	114	200	466
Overheat fault	42	126	126	200	494

Table 2 clearly shows the complete progressive process from the original sample collection to the final equalized training set. In the original data, normal samples account for about 79% of all collected samples, while partial discharge samples only account for 4.3%. This extreme imbalance will lead to that the majority of gradient update directions of the model dominated by normal samples during training, and the fault mode can hardly be effectively learned. After the double amplification of time domain stretching and frequency domain masking, the sample size of each fault category is expanded to about three times of the original on average, but the amplification principle determines that the new sample is still limited to the neighborhood of the original sample and cannot be expanded to a wider distribution space. The intervention of WGAN-GP essentially changes this limitation. By learning the distribution law of real samples on high-dimensional manifolds, its generator can generate new samples that are not simple interpolation. The 200 virtual samples generated not only expand the number, but more importantly fill the sparse region in the original feature space. The coverage of each category in the feature space is more continuous and complete. Finally, the number of samples of all categories reaches between 466 and 700, and the ratio of the maximum to the minimum categories is greatly reduced from the initial 18.4:1 to 1.5:1 (700:466), which significantly reduces the bias of the model to most categories. The multi-channel tensor data set after the above preprocessing, multi-domain feature extraction and data equalization will be used as the standardized input of the subsequent deep diagnosis and prediction model. Its high signal-to-noise ratio, multi-domain complementarity and category equalization provide a solid basic guarantee for the upper limit of model performance.

3. FAULT DIAGNOSIS MODEL ARCHITECTURE BASED ON DUAL-BRANCH DOMAIN ADAPTIVE CONVOLUTIONAL NETWORK

On the basis of signal preprocessing and multi-domain feature enhancement, how to build an efficient and robust deep network architecture to achieve accurate fault diagnosis under cross-working conditions has become the core scientific problem to be solved in this section. In the actual operation of power transformers, the load rate can jump from 20% to 100% within tens of minutes, and the oil temperature will change drastically accordingly. The dynamic drift of these working conditions will lead to significant deviation of the vibration spectrum and electrical characteristic distribution presented by the same fault type at different times, namely the so-called domain drift phenomenon. Traditional deep network assumes that the training set and test set follow the independent and same distribution. When the diagnostic model trained under dead load conditions in the laboratory is directly deployed to the field, its accuracy often drops from more than 95% to less than 60%, which constitutes a fundamental obstacle limiting the engineering implementation of deep learning diagnostic methods. At the same time, the spectral performance of transformer faults has multi-scale characteristics, and the characteristic frequency offset caused by winding deformation often covers a wide band change of several kHz, which belongs to the global spectral structure change, while the partial discharge pulse shows a transient peak of several microseconds width on the time spectrum diagram, which belongs to the local fine structure anomaly. It is difficult for a single-path convolutional network to capture these two different feature patterns simultaneously in the scale range of the same set of convolution kernels. Therefore, a dual-branch domain adaptive convolutional network architecture is designed in this section, which extracts global semantic and local mutation features respectively through dual-branch parallel path at the structure level, and introduces dual-domain adaptation mechanism at the training strategy level to forcibly align the high-order feature distribution of source domain and target domain, so as to realize cross-working condition adaptive diagnosis without target domain labels.

The selection of basic backbone network directly determines the starting point quality and computational efficiency of feature extraction. The design of residual connection in convolutional neural network enables ResNet series to effectively alleviate the problem of gradient disappearance and support depths of more than 100 layers. Its jump connection structure ensures that gradient signals can flow directly from deep to shallow layers, greatly improving the optimization difficulty of deep networks. However, the identity mapping branch of ResNet directly superimposes input and output during cross-layer transmission. When the size of input and output feature maps is inconsistent, additional projection convolution layers must be introduced, which destroys the purity of features and accumulates a considerable number of parameters in the deep network. DenseNet concatenates all the previous layer feature maps in the channel dimension through dense connection and transmits them to the subsequent layers. Each layer can directly obtain the gradient signals from all the previous layers, which significantly enhances the feature reuse efficiency and achieves better classification accuracy than ResNet under the condition of the same number of parameters. However, DenseNet's dense connection strategy will lead to a sharp expansion of the number of channels in the feature graph after the number of channels is accumulated layer by layer. Especially when the network depth exceeds 100 layers, the computational complexity increases by square level, which limits its deployment feasibility in real-time monitoring scenarios. Comprehensive evaluation of the performance of the two types of networks in the time spectrogram classification task shows that ResNet has a better response ability in large-scale frequency shift features, while DenseNet has a higher sensitivity to local subtle changes.

Both have their own focuses, but neither can give consideration to computational efficiency and multi-scale feature extraction ability. EfficientNet uses a composite scaling strategy to uniformly adjust the three dimensions of network depth, width and input resolution, and achieves higher classification accuracy in ImageNet benchmark tests with a much lower number of parameters than ResNet and DenseNet. The depth separable convolution introduced in its basic convolution module has greatly reduced the computational load.

Based on this, this paper further hollows the standard depth separable convolution in EfficientNet and expands its effective receptive field under the premise of keeping the number of convolution kernel parameters unchanged. Let the size of the deep convolution kernel of the standard depth separable convolution be $k \times k$. After the introduction of void rate r, the effective size of the convolution kernel is expanded to $k + (k-1)(r-1)$, while the number of parameters involved in the operation is still $k \times k$. Taking $k = 3$ and $r = 2$ as an example, the effective receptive field is expanded from 3×3 to 5×5, while the number of parameters remains unchanged. The response value of the output feature map of the cavity convolution at position (i, j) can be described by the following equation:

$$Y(i,j) = \sum_{u=-a}^{a} \sum_{v=-a}^{a} W(u,v) \cdot X(i, +r \cdot uj + r \cdot v) \tag{19}$$

Where $Y(i,j)$ is the pixel value of position (i,j) in the output feature map, $W(u,v)$ is the weight of the convolution kernel with size $(2a+1) \times (2a+1)$, the value of a is determined by the size of the convolution kernel, X is the input feature map, and r is the void rate. This operation makes the effective receptive field of the latter stages in EfficientNet reach 31×31, 49×49 and 73×73 respectively, which is enough to cover the long-range dependence between adjacent frequency bands in SSST time spectrograms, while the total number of network parameters only increases by less than 2%, ensuring the advantage of lightweight.

On the basis of the hollow-out EfficientNet backbone, this paper further constructs a dual-branch parallel feature extraction path to solve the structural contradiction that a single-scale convolution kernel cannot simultaneously capture wideband global offset and transient local mutation. The global semantic branch uses EfficientNet after hollowing transformation as the front-end feature encoder. Its multi-stage downsampling process gradually compresses the space size of the tensor input to $1/32$ of the original, and the number of channels is extended step by step from the initial 32 dimensions to 1280 dimensions. However, although the end feature map relying solely on convolutional downsampling has the largest receptive field, it loses the mid-band transition information in the original input. To this end, this paper embedded the Pyramid Pooling Module (PPM) at the end of EfficientNet and applied four average pooling operations of different scales on the last layer of convolution feature graph $\mathcal{F} \in \mathbb{R}^{C \times H' \times W'}$ in parallel. The pooling window sizes are 1×1, 2×2, 3×3 and 6×6, where C is the number of channels, and H' and W' are the height and width of the feature map, respectively. Each pooling branch compresses the feature map to global and local statistical descriptors with scales of $C \times 1 \times 1$, $C \times 2 \times 2$, $C \times 3 \times 3$ and $C \times 6 \times 6$, respectively. Then, the number of channels of each branch is uniformly compressed to $C/4$ by 1×1 convolution, and the spatial size $H' \times W'$ is restored to the same as the original feature map by bilinear interpolation upsampling. Finally, the upsampling results of the four pooling branches are concatenated with the original feature graph in the channel dimension to obtain the global semantic enhanced feature $\mathcal{F}_{global}$, whose channel number is $C + 4 \times (C/4) = 2C$. The pyramid structure enables the network to simultaneously sense the global

average response at 1×1 scale, the regional statistical characteristics at 2×2 scale, the local context at 3×3 scale and the fine spatial distribution at 6×6 scale, which is especially suitable for capturing the cross-band energy redistribution pattern caused by winding deformation and other faults.

The local sensitive branch adopts the design concept of complete symmetry with the global semantic branch but complementary functions. This branch does not aim at the vastness of the receptive field but focuses on the preservation of spatial details and the amplification of local mutation signals. The front end of the local sensitive branch is composed of three small cores densely connected subnetworks in cascade. Each subnetwork contains four layers of 3×3 depth separable convolutional layers, and the dense connection mode is adopted between each layer, that is, the input of the l layer is the splicing of all output feature maps of the previous $l-1$ layer in the channel dimension. Let the output of layer l be $\mathcal{F}_l$, then the dense connection of this layer can be expressed as:

$$\mathcal{F}_l = \mathcal{H}_l([\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_{l-1}]) \tag{20}$$

Where $\mathcal{H}_l$ is the convolution transformation of layer l, and $[\cdot]$ represents the channel dimension splicing operation. No downsampling operation is introduced after each convolutional layer, and the spatial resolution of the feature map is kept consistent with the input only by filling, so as to ensure that the transient spatial position of the microsecond discharge pulse is not blurred by downsampling. The output of each subnetwork is compressed by a 1×1 convolution and then transmitted to the next subnetwork. The progressive channel growth strategy is adopted among the three subnetworks, and the number of channels is 64, 128 and 256 in turn. In order to further enhance the Response sensitivity to weak signals, this paper introduces a Local Response Normalization layer (LRN) at the tail of the small-core densely connected subnetwork to competitively suppress the response values of adjacent channels at the same spatial location, so that the activation values of strong response channels are more prominent. Its normalization formula is as follows:

$$b_{x,y}^{c} = a_{x,y}^{c} \Big/ \left(\alpha + \beta \sum_{c'=\max(0, c-n/2)}^{\min(C, -1c+n/2)} (a_{x,y}^{c'})^2 \right)^{\gamma} \tag{21}$$

Where $a_{x,y}^{c}$ and $b_{x,y}^{c}$ are the activation values at the spatial position (x, y) and channel c in the feature map before and after normalization, respectively; α, β and γ are hyperparameters, which are taken as 10^{-4}, 0.75 and 1 respectively in this paper; n is the number of neighborhood channels, which is taken as 5. This operation significantly enhances the channel activation intensity corresponding to sparse shocks such as PD, and suppresses the response of background noise channels, so that the signal-to-noise ratio of weak fault pulses in feature space is additional improved.

After the feature extraction of both branches, how to organically integrate global semantic features and local sensitive features to form a more discriminative joint representation is the key link of cross-branch information fusion. In this paper, the cross-attention fusion gating mechanism is designed to replace the traditional simple splicing or additive fusion. This mechanism can dynamically learn the contribution weight of the features of the two branches according to the characteristics of the current input sample, rather than assigning a fixed fusion ratio to the two branches [21]. Set global bifurcation output characteristics for $\mathcal{F}_{global} \in \mathbb{R}^{C_g \times H'' \times W''}$, the local branch output characteristics for $\mathcal{F}_{local} \in \mathbb{R}^{C_l \times H'' \times W''}$, which $H''$and W'' for the fusion of phase space dimension, The two are compressed into global semantic vector $v_g \in \mathbb{R}^{C_g}$ and local sensitivity vector $v_l \in \mathbb{R}^{C_l}$ by global average pooling. After splicing the two vectors, input the two-layer fully connected gated network to generate the fusion

weight coefficient:

$$w = \text{Sigmoid}\big(W_2 \cdot \text{ReLU}\big(W_1 \cdot [v_g, v_l] + b_1\big) + b_2\big) \tag{22}$$

Where $W_1 \in \mathbb{R}^{d_{hidden} \times (C_g + C_l)}$ and $W_2 \in \mathbb{R}^{2 \times d_{hidden}}$ are the weight matrices of the full connection layer, b_1 and b_2 are bias terms, d_{hidden} is the dimension of the hidden layer, and $[\cdot]$ represents the splicing operation. $w = [w_g, w_l]^T$ is a two-dimensional weight vector, where w_g and w_l correspond to the contribution coefficients of global branches and local branches respectively, and $w_g + w_l = 1$. After obtaining the weight coefficients, channel attention recalibration is first performed on the feature maps of the two branches. Taking the global branch as an example, the global average response of each channel c of $\mathcal{F}_{global}$ is calculated:

$$z_c = \frac{1}{H'' W''} \sum_{i=1}^{H''} \sum_{j=1}^{W''} \mathcal{F}_{global}(c, i, j) \tag{23}$$

Then, channel recalibration coefficients are generated through a gating network:

$$\alpha = \sigma(W_{attn} \cdot z) \tag{24}$$

Where $z \in \mathbb{R}^{C_g}$ is the channel description vector, W_{attn} is the attention transformation matrix, and σ is the Sigmoid activation function. The channel attention coefficient is multiplied by the original feature map channel by channel, and then scaled by the global gating weight w_g to obtain the final fusion feature:

$$\mathcal{F}_{fused} = w_g \cdot \big(\alpha_g \odot \mathcal{F}_{global}\big) \oplus w_l \cdot (\alpha_l \odot \mathcal{F}_{local}) \tag{25}$$

Where $\odot$ indicates element-by-element multiplication in channel dimension, $\oplus$ indicates element-by-element addition, α_g and α_l are channel attention coefficient vectors of global branch and local branch respectively. The fusion strategy makes the network automatically increase w_g in the face of global spectrum drift to highlight the contribution of global semantic branches and automatically increase w_l in the detection of transient pulse signals such as partial discharge to activate the sensitivity of local branches and realizes the input adaptive adjustment of feature fusion.

Feature fusion solves the integration problem of dual-branch information, but the classification decision boundary trained by the fused features on the source domain will still shift significantly when facing the data in the target domain. The root cause of the drift across the working domain is that the load change leads to the change of the magnetic flux density of the transformer core, which leads to the change of the relative amplitude of the harmonic components in the vibration spectrum. Meanwhile, the rise and fall of the oil temperature changes the dielectric characteristics of the insulating material, thus distorting the propagation and attenuation characteristics of the partial discharge pulse. These changes at the physical level are finally manifested as the separation of the edges of the distribution of samples from the source domain and the target domain in the feature space. In order to achieve effective domain adaptation without obtaining target domain labels, this paper introduces dual domain adaptation loss to forcibly align source domain and target domain at feature level and distribution level at the same time. The first adaptation mechanism uses the Maximum Mean Discrepancy (MMD) to measure the distribution distance between the source domain and the target domain in the regenerative Hilbert space [22],[23],[24]. Let the source domain fusion feature set be:

$$\mathcal{F}_{fused}^{s} = \{f_i^s\}_{i=1}^{N_s} \tag{26}$$

The target domain fusion feature set be:

$$\mathcal{F}_{fused}^{t} = \{f_j^t\}_{j=1}^{N_t} \tag{27}$$

Where N_s and N_t are the number of samples in the source domain and the target domain respectively, and the MMD distance is defined as:

$$\mathcal{L}_{MMD} = \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} \phi\left(f_i^s\right) - \frac{1}{N_t} \sum_{j=1}^{N_t} \phi\left(f_j^t\right) \right\|_{\mathcal{H}}^2 \tag{28}$$

Where $\phi(\cdot)$ is the kernel mapping that maps the original feature to the regenerated Hilbert space $\mathcal{H}$. This paper adopts multi-kernel MMD to enhance the adaptive capacity, fusing three Gaussian kernels with different bandwidths:

$$k_\sigma(x, y) = \exp\left(-\| x - \frac{y \|^2}{2\sigma^2}\right) \tag{29}$$

The bandwidths σ are 0.5 times, 1 time and 2 times of the median distance respectively. Multi-core MMD is expanded into computable form:

$$\mathcal{L}_{MMD} = \frac{1}{N_s^2} \sum_{i=1}^{N_s} \sum_{i'=1}^{N_s} K\left(f_i^s, f_{i'}^s\right) + \frac{1}{N_t^2} \sum_{j=1}^{N_t} \sum_{j'=1}^{N_t} K\left(f_j^t, f_{j'}^t\right) - \frac{2}{N_s N_t} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} K\left(f_i^s, f_j^t\right) \tag{30}$$

Where $K(x, y) = \sum_{m=1}^{3} k_{\sigma_m}(x, y)$ is a multicore combination function. MMD loss forces the network to learn domain invariant feature representation by minimizing the distance between the feature means of the source domain and the target domain in the kernel space, but it essentially only matches the first-order statistics of the distribution and has limited perception ability to the difference of high-order moments.

Therefore, this paper introduces the second adaptation mechanism on the basis of MMD, namely the adversarial domain discriminator, which forces the feature extractor to generate feature representations that cannot distinguish the source domain through the way of game training. The domain discriminator $\mathcal{D}$ is a four-layer fully connected network, the input is the fusion feature f_{fused}, and the output is the binary classification probability of whether the feature belongs to the source domain or the target domain. The adversarial relationship between feature extractor $\mathcal{G}$ (that is, the whole process of two-branch network) and domain discriminator $\mathcal{D}$ is expressed by domain classification loss:

$$\mathcal{L}_{adv} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \log \mathcal{D}\left(\mathcal{G}(x_i^s)\right) - \frac{1}{N_t} \sum_{j=1}^{N_t} \log\left(1 - \mathcal{D}\left(\mathcal{G}(x_j^t)\right)\right) \tag{31}$$

In the training process, the goal of $\mathcal{G}$ is to minimize $\mathcal{L}_{adv}$ so that the domain discriminator cannot distinguish the feature source, while the goal of $\mathcal{D}$ is to maximize $\mathcal{L}_{adv}$ to improve the discrimination accuracy, and the two form a minimax game. In this paper, the Gradient Reversal Layer (GRL) is

embedded in the gradient back propagation layer. In the forward propagation, the identity mapping $\mathcal{R}(x) = x$ is performed in this layer. In the back propagation, the gradient is multiplied by the negative constant $-\lambda_{grl}$, that is, $\partial\mathcal{R}/\partial x = -\lambda_{grl} I$, Where λ_{grl} is the hyperparameter controlling the adaptation strength of the domain, and its value increases linearly from 0 to 1 with the training rounds. The introduction of gradient inversion layer enables the network to complete the adversarial training of feature extractor and domain discriminator at one time in the standard stochastic gradient descent framework, without alternately updating the independent optimization cycle of the two networks. The MMD loss and the adversarial loss are combined into the dual domain adaptation total loss:

$$\mathcal{L}_{DA} = \mathcal{L}_{MMD} + \beta \mathcal{L}_{adv} \tag{32}$$

Where β is the balance coefficient taken as 0.5. In the dual mechanism, MMD provides a robust global distribution alignment baseline, while the adversarial mechanism further eliminates the high-order moment difference. The synergistic effect of the two mechanisms minimizes the distribution difference between the source domain, and the target domain features in the deep semantic space.

Finally, the total loss function of the network consists of three parts: supervised classification loss in the source domain, adaptation loss in the dual domain, and regularization constraints of fusion gating. The source domain classification using the standard cross entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{33}$$

Including $y_{i,c}$ for the case of a source domain sample i real category labels, $\hat{y}_{i,c}$ is the Softmax class probability of the network output, and C is the total number of fault classes. In order to prevent the weight coefficient generated by fusion gating from excessively biased towards one branch and causing the gradient of the other branch to disappear, this paper applies the entropy regularization constraint on the gating output:

$$\mathcal{L}_{entropy} = -w_g \log w_g - w_l \log w_l \tag{34}$$

This constraint obtains the maximum value when $w_g = w_l = 0.5$, and tends to 0 when w_g or w_l tends to 0 or 1, forcing the gated network to keep the balanced utilization of the two branches as far as possible. The total loss function is summarized as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{DA} \mathcal{L}_{DA} - \lambda_{ent} \mathcal{L}_{entropy} \tag{35}$$

Where λ_{DA} and λ_{ent} are the weight coefficients of domain adaptation loss and entropy regularization respectively, which are taken as 0.1 and 0.01 respectively in this paper. In the process of back propagation, the gradient of $\mathcal{L}_{cls}$ only updates the feature extractor and classifier parameters, the gradient of $\mathcal{L}_{DA}$ updates the feature extractor and domain discriminator through the gradient inversion layer, and the gradient of $\mathcal{L}_{entropy}$ only acts on the gated network parameters. Through end-to-end joint optimization, the dual-branch domain adaptive convolutional network learns the fault classification decision boundary on the annotated data of the source domain and drives the feature distribution alignment by using the unlabeled target domain data through the dual-domain adaptation, so as to obtain the diagnosis output with high confidence in the target domain. This architecture systematically solves the dual challenges of multi-scale feature capture and cross-condition generalization in transformer fault diagnosis from two dimensions of structure design and training

strategy.

4. FAULT EVOLUTION TREND PREDICTION METHOD BASED ON TIME SEQUENCE GENERATION ADVERSARIAL NETWORK

After the accurate diagnosis of the current state of the transformer is realized, it is of more direct engineering value to further predict the fault evolution trend and quantify the Remaining Useful Life (RUL) for guiding the condition maintenance strategy and reducing the risk of unplanned outage. It usually takes several weeks or even months for the transformer to develop from early latent defect to complete failure, during which the characteristic frequency components in the vibration spectrum drift slowly, the repetition frequency of partial discharge pulse gradually increases, and the harmonic energy ratio shows a monotonically increasing or decreasing evolution law. These evolutionary trajectories are hidden in the massive time spectrograms and statistical feature sequences accumulated by continuous monitoring, and their temporal dependence spans multiple time scales from minute to month. Traditional time series prediction methods such as differential integrated moving average autoregressive model (ARIMA) are limited to the modeling of linear stationary series, while recurrent neural network and its variants face the inherent dilemma of gradient disappearance when dealing with long-term dependence. Even if the gating mechanism is introduced, the memory decay problem cannot be completely solved when the sequence length exceeds hundreds of steps. More importantly, the fault evolution process itself has inherent randomness, load fluctuations and accidental changes in operating conditions make the RUL of the same fault type show obvious distribution dispersion, single point prediction value is difficult to reflect this uncertainty, and the interval estimation of probabilistic prediction is the information really needed for engineering decision-making. This section proposes a prediction method based on Temporal Convolutional Generative Adversarial Network (TCN-GAN). The modeling of fault evolution trajectory, the capture of multi-scale time series features and the quantification of prediction uncertainty are integrated into an adversarial training framework to realize the leap from deterministic point prediction to probabilistic interval prediction.

The first task of fault evolution trend prediction is to organize continuous monitoring data into a format suitable for time series modeling. The transformer online monitoring system continuously collects vibration signals and current signals at a fixed sampling interval Δt (1 hour) and generates the SSST time spectrum sequence $\{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_T\}$ and the time domain statistical feature vector:

$$m_t = [K_t, C, F_t H_{ESE,t}]^T \tag{36}$$

Where the subscript t represents the t-th monitoring time and T is the total number of steps currently observed. In order to construct the standard input-output pair of the time series prediction model, this paper adopts the sliding window strategy, takes the observation sequence of the historical window length L_{in} as the input, predicts the evolution trend of the future L_{out} step, thus forming a mapping relationship between the input tensor and the target sequence. The traditional cyclic network uses recursive processing sequence one time step by time, and the update of hidden state $h_t = \Phi(h_{t-1}, x_t)$ can only be carried out after the calculation at time $t-1$ is completed. This inherent serial characteristic leads to low efficiency of training and inference, and the length of hidden state vector is fixed at d_h. When the sequence length L_{in} increases, the network must compress the information of all past moments into the state vector of the same dimension, resulting in inevitable information bottlenecks. In order to overcome the above limitations, this paper forgoes the temporal recursion paradigm of cyclic network and adopts the parallel sequence modeling framework based on

convolution. The input data of all historical times are divided into four-dimensional tensors with dimension $L_{in} \times H \times W \times (C_{spec} + 3)$ and input to the encoder, where H and W are the height and width of the spectral map are. C_{spec} is the number of channels in the spectrogram, and the additional three channels carry time-domain statistical features. The encoder uses multi-layer hole convolution to compress and encode the input sequence, and the hole rate of each layer increases exponentially. Finally, all the historical information is encoded into a fixed length context vector $c \in \mathbb{R}^{d_c}$. The decoder generates the spectral sequence and feature parameter sequence of the future L_{out} step frame by frame through autoregressive method conditional on the context vector. The advantage of the encoder-decoder framework is that the coding process is completely parallel, not limited by the recursive delay of sequence length, and the context vector is only the conditional information rather than the only carrier that must carry all the historical information. The decoder can still directly trace back the intermediate representation of each layer of the encoder through the attention mechanism when generating each step.

The core design of the generator revolves around expansive causal convolution, which simultaneously solves two key constraints of receptive field expansion and causality in time series modeling. Causality requires that the predicted output at time t can only rely on the historical input before t and cannot snoop on the future information. This constraint is realized by the convolution operation that masks the future. Set for the input sequence $x \in \mathbb{R}^{L \times d_x}$, the output sequence to $y \in \mathbb{R}^{L \times d_y}$, standard causal convolution in the output of the first t time for:

$$y_t = \sum_{i=0}^{k-1} W_i \cdot x_{t-i} \tag{37}$$

The k for convolution kernels size, this equation shows that the output t depends only on the inputs the type shows that only depends on the input $t, t-1, \ldots, t-k+1$, strictly following the causal constraints. However, the receptive field of standard causal convolution only covers k time steps, and even if multiple layers are superimposed, the receptive field only grows linearly with the number of layers. Expansive causal convolution inserts a fixed number of zeros between adjacent weights of the convolution kernel such that the receptive field expands exponentially. Let the void rate d, then the output of the expansive causal convolution at time t is:

$$y_t = \sum_{i=0}^{k-1} W_i \cdot x_{t-d \cdot i} \tag{38}$$

This equation shows that the convolution kernel expands from k consecutive positions to k positions spaced d steps apart on the time axis, and the effective receptive field length expands from k to $(k-1)d+1$. The generator constructed in this paper is stacked by five layers of expansive causal convolution. The void rate of each layer is $d = 1,2,4,8,16$ in turn, and the size of the convolution kernel is unified as $k = 3$, so the length of the effective receptive field at the top layer is $(3-1) \cdot 16 + 1 = 33$ time steps. If the input window length is $L_{in} = 64$, then the top-level reception-field covers the information of the latest 33 moments in the historical sequence, and the distant historical moments (steps 34 to 64) indirectly affect the top-level output in a hierarchical abstraction way through the convolution operation of the lower layer, realizing the cross-scale temporal dependence modeling. After each layer of dilated causal convolution, weight normalization and Parametric modified linear unit (PReLU) activation function are followed. PReLU introduces learnable slope parameters in the negative value region, which avoids the neuron death problem of standard ReLU. It is defined as:

$$\mathrm{PReLU}(z) = \max(0, z) + a \cdot \min(0, z) \tag{39}$$

Where a is a trainable parameter and the initial value is set to 0.25.

On the top layer of multi-layer expansive causal convolution, the self-attention spatial-temporal gating module is embedded in this paper to weigh the importance in time dimension and feature dimension at the same time. The gating module outputs feature sequences $H = [h_1, h_2, \ldots, h_{L_{in}}]^T \in \mathbb{R}^{L_{in} \times d_h}$ as the input, the first query through the calculation of linear transformation matrix $Q = HW_Q$, key matrix $K = HW_K$, $V = HW_V$ and value matrix, including $W_Q, W_K \in \mathbb{R}^{d_h \times d_k}$, $W_V \in \mathbb{R}^{d_h \times d_v}$, d_k and d_v are the key dimension and value dimension of attention space, respectively. The temporal self-attention weight matrix is obtained from the dot product of query and key and normalized by Softmax:

$$A = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{40}$$

Where the element a_{ij} of $A \in \mathbb{R}^{L_{in} \times L_{in}}$ represents the attention contribution coefficient of the j-th time step to the i-th time step, and the division by $\sqrt{d_k}$ is to prevent the saturation of Softmax gradient caused by too large dot product value. However, the standard self-attention mechanism treats all time steps equally, which is not in line with the actual demand of fault evolution prediction, the influence of recent observations on predicting future trends should be greater than that of long-term observations, that is, the attention weight should show monotone decay characteristics in time difference. To this end, a predefined relative position bias matrix $B \in \mathbb{R}^{L_{in} \times L_{in}}$ is superimposed on the attention weight in this paper, and its element:

$$b_{ij} = -\log(1 + | i - j |) \tag{41}$$

So that the farther the time step is, the greater the negative bias will be, and its effective weight will be suppressed after Softmax. The corrected attention output is:

$$\mathrm{Attn}(H) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{42}$$

The output is fused with the original feature through a gating mechanism, and the gating coefficient is generated from the input sequence through a small convolutional network, and the spatio-temporal weighted feature sequence $H^{weighted}$ is finally obtained. This gating enables the generator to preferentially focus on observations from the first 10 to 15-time steps when forecasting recent trends, while still retaining a low-weight perception of slowly changing trends in more distant histories, achieving a balance between proximal weighting and distal reference.

The predicted sequence output by the generator will be fed to the discriminator for adversarial evaluation. Different from the traditional GAN discriminator which only outputs scalar authenticity probability, the discriminator designed in this paper adopts the sequence discrimination structure, which makes the overall authenticity judgment on the whole prediction sequence and outputs the confidence score of each time step. The discriminant is composed of four layers of one-dimensional convolution, with convolution kernel sizes of $5,3,3,3$ respectively, and the number of channels of $64,128,256,1$ in turn. After each layer, the sequence length is gradually compressed to a single step by following LeakyReLU activation function (negative slope 0.2) and average pooling with step size of 2. Finally, the Sigmoid activation outputs the scalar discriminant probability:

$$p_{real} = \mathcal{D}(\tilde{Y}) \in [0,1] \tag{43}$$

Where $\tilde{Y}$ is the prediction sequence output by the generator. At the same time, after the last convolutional layer and before global pooling, the discriminator extracts the intermediate activation value r_t corresponding to each time step, and maps them to a confidence score via an independent fully connected layer:

$$c_t = \sigma(W_c \cdot r_t + b_c) \tag{44}$$

Where σ is the Sigmoid function. The confidence score vector:

$$c = [c_1, c_2, \dots, c_{L_{out}}]^T \tag{45}$$

Reflects the confidence degree of the discriminator to the authenticity of the generated sequence at each time step. When the prediction quality of the generated sequence at step t is poor, the confidence score of the discriminator at this step is low, otherwise the score is high. The loss function of the discriminator in the training process contains two terms: the traditional adversarial loss term to distinguish the true sequence from the generated sequence, and the confidence regularization term to smooth the mutation between the confidence scores of adjacent time steps. Let the true future sequence be Y and the generated sequence be $\tilde{Y}$, and the total loss of the discriminator is:

$$\mathcal{L}_D^{seq} = -\mathbb{E}_{Y\sim\mathbb{P}_r}[\log \mathcal{D}(Y)] - \mathbb{E}_{\tilde{Y}\sim\mathbb{P}_g}\left[\log\left(1 - \mathcal{D}(\tilde{Y})\right)\right] + \lambda_c \sum_{t=2}^{L_{out}} (c_t - c_{t-1})^2 \tag{46}$$

Where λ_c is 0.01 for the confidence smoothness coefficient. The sequence discriminant structure makes the discriminator not only give the authenticity judgment of the whole sequence but also provide the local quality assessment of each step in the time granularity, which provides the key information basis for the subsequent probability interval estimation.

The adversarial training of generator and discriminator relies on a carefully designed hybrid loss function, which incorporates sequence morphological similarity constraint and deterministic deviation penalty on the basis of traditional Wasserstein distance loss, so as to guide the generator to pay attention to the overall distribution matching, local trend pattern and amplitude accuracy of prediction simultaneously. Wasserstein distance loss adopts a formula similar to that in Section 3, and the goal of the generator is to make the discriminator score the generated sequence output as high as possible:

$$\mathcal{L}_W = -\mathbb{E}_{\tilde{Y}\sim\mathbb{P}_g}[\mathcal{D}(\tilde{Y})] \tag{47}$$

Wasserstein loss only restricts the generated series from the perspective of distribution matching and does not impose direct supervision on the similarity between the predicted trajectory and the real trajectory in time series morphology. The fault evolution process follows a specific physical law, and its characteristic parameters usually show a monotonically increasing, monotonically decreasing, or first smooth and then accelerated S-shaped change trend. The matching degree of trend form between the predicted sequence and the real sequence can reflect the engineering value of the prediction better than the point-by-point amplitude accuracy. Therefore, this paper introduces Dynamic Time Warping (DTW) distance as the loss of sequence morphological similarity. DTW uses dynamic programming to nonlinearly align the two series on the time axis, allowing the series to be scaled and matched in the time dimension, so it is naturally robust to phase offset and local time distortion. Let the true sequence be:

$$Y = \{y_1, y_2, \dots, y_{L_{out}}\} \tag{48}$$

The generated sequence is:

$$\tilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_{L_{out}}\} \tag{49}$$

DTW structure first cumulative distance matrix:

$$D \in \mathbb{R}^{(L_{out}+1)\times(L_{out}+1)} \tag{50}$$

Initialize $D_{0,0} = 0$, $D_{i,0} = D_{0,j} = \infty$, and then press the following filling matrix recursion formula:

$$D_{i,j} = d(y_i, \tilde{y}_j) + \min\{D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}\} \tag{51}$$

Where $d(y_i, \tilde{y}_j) = (y_i - \tilde{y}_j)^2$ is the square of the pointwise Euclidean distance. The final DTW distance is $D_{L_{out},L_{out}}$, that is, the cumulative cost of the optimal alignment path backtracked from the lower right corner of the matrix. DTW loss is defined as the normalized DTW distance:

$$\mathcal{L}_{DTW} = \frac{1}{L_{out}} \sum_{(i,j)\in\pi^*} d(y_i, \tilde{y}_j) \tag{52}$$

Where π^* is the set of all point pairs on the optimal alignment path. This loss directly penalizes the global difference in shape between the forecast series and the true series, and DTW can still give low distance values even if the forecast series has small leads or lags in the time axis, avoiding excessive requirements for point-by-point alignment.

In addition to distribution and morphological constraints, deterministic loss is also introduced to ensure the amplitude of accuracy of the forecast sequence at key time nodes. The deterministic loss adopts the smooth L_1 loss function (Huber loss), which shows the smoothness of L_2 loss when the error is small, and the robustness of L_1 loss when the error is large, which is defined as:

$$\mathcal{L}_{det} = \frac{1}{L_{out}} \sum_{t=1}^{L_{out}} \begin{cases} \frac{1}{2}(y_t - \tilde{y}_t)^2, & |\, y_t - \tilde{y}_t \,| \le \delta \\ \delta\,|\, y_t - \tilde{y}_t \,| - \frac{1}{2}\delta^2, & |\, y_t - \tilde{y}_t \,| > \delta \end{cases} \tag{53}$$

Where δ is the threshold parameter taken as 0.5 for Huber loss. By weighing the above three losses, the total loss function of the generator is:

$$\mathcal{L}_G = \mathcal{L}_W + \alpha_{DTW}\mathcal{L}_{DTW} + \alpha_{det}\mathcal{L}_{det} \tag{54}$$

The weight coefficients α_{DTW} and α_{det} are 0.2 and 0.5 respectively, and the weight distribution reflects the core design idea of this paper: Wasserstein loss provides the basis for adversarial driven distribution matching, DTW loss ensures the rationality of evolutionary trend form, and deterministic loss forces the amplitude accuracy of key time nodes to anchor, all of which jointly constrain the optimization direction of the generator from different levels.

After the predicted value is determined, how to quantify the forecast uncertainty is the core premise of probabilistic RUL estimation. In this paper, the time-step confidence score $c = [c_1, c_2, \dots, c_{L_{out}}]^T$ of the output of the discriminator is used to guide the interval estimation process of Monte Carlo dropout

sampling. The core idea of Monte Carlo dropout is to keep the dropout layer active in the test phase and conduct M random forward propagation on the same input. Each propagation generates different prediction outputs due to the random dropping of different neurons by dropout. M outputs constitute the empirical distribution of predicted values. Its meaning and variance reflect the central trend and uncertainty of the forecast, respectively. Specifically, t a prediction for the first step, to spread before M times M forecast:

$$\left\{\tilde{y}_t^{(1)}, \tilde{y}_t^{(2)}, \ldots, \tilde{y}_t^{(M)}\right\} \tag{55}$$

The meaning for:

$$\mu_t = \frac{1}{M}\sum_{m=1}^{M} \tilde{y}_t^{(m)} \tag{56}$$

Variance of:

$$\sigma_t^2 = \frac{1}{M-1}\sum_{m=1}^{M} (\tilde{y}_t^{(m)} - \mu_t)^2 \tag{57}$$

Standard Monte Carlo treats all time steps equally and does not make full use of the stepwise confidence information provided by the discriminator. This paper proposes a confidence-weight prediction interval estimation method, which uses the confidence score c of the discriminant to reweigh the results of M sampling times, so that the time step with high confidence can obtain greater interval tightening strength and the time step with low confidence can obtain wider interval coverage:

$$\sigma_{t,weighted}^2 = \frac{1}{\sum_{m=1}^{M} c_t^{(m)}} \sum_{m=1}^{M} c_t^{(m)} \left(\tilde{y}_t^{(m)} - \mu_t\right)^2 \tag{58}$$

Where $c_t^{(m)}$ is the confidence score of the discriminator on the output of step t at the m-th forward propagation. After winning the weighted variance, assuming that the prediction error is normal distribution, is the first step t of predicted value $(1-\alpha)\times 100\%$ confidence interal $[\mu_t, -z_{1-\alpha/2} \cdot \sigma_{t,weighted}\mu_t + z_{1-\alpha/2} \cdot \sigma_{t,weighted}]$, Where $z_{1-\alpha/2}$ is the upper $1-\alpha/2$ quantile of the standard normal distribution, and $\alpha = 0.1$ corresponds to the 90% confidence interval in this paper. When the predicted trajectory crosses the pre-set fault threshold upward, the probability distribution of RUL can be calculated according to the distribution of the crossing time, and its expected value:

$$\hat{T}_{RUL} = \sum_{t=1}^{L_{out}} t \cdot P(\tau = t) \tag{59}$$

Where τ is the random variable for the forecast sequence to cross the threshold for the first time. The probabilistic RUL estimation provides complete risk information rather than a single deterministic value for operation and maintenance decisions.

Table 3. Summary of configuration parameters at each layer of the generator network

Hierarchy of networks	Type of operation	Convolution kernel size	Rate of void	Channel of input	Channel of output	Receptive field accumulation
Level 1	Expansive causal convolution	3	1	64	128	3
Level 2	Expansive causal convolution	3	2	128	256	7
Level 3	Expansive causal convolution	3	4	256	256	15
Level 4	Expansive causal convolution	3	8	256	512	31
Level 5	Expansive causal convolution	3	16	512	512	63

Table 3 summarizes the core configuration parameters of the five-level inflated causal convolution in the generator. From layer 1 to layer 5, the void rate increases by a power of 2, which makes the receptive field expand exponentially from the initial 3-time steps to 63-time steps of layer 5. It is worth noting that the number of channels after the third layer increases from 256 to 512 to accommodate the abstract timing features at a higher level. Meanwhile, the reception field covers 63 historical steps in the input window length $L_{in} = 64$ at the fifth layer, ensuring that almost all historical information can be perceived by the top layer at different abstraction levels. Each layer adopts a convolution kernel size of 3×3, which is the widely verified optimal choice in temporal convolutional networks. 1×1 cannot capture the local dependence of adjacent time steps, while 5×5 introduces too many parameters and the receptive field expands too fast, resulting in the loss of details. The design of the number of input and output channels at each layer follows the principle of gradual expansion from low dimension to high dimension. The number of low-level channels is less to extract local timing pattern, and the number of high-level channels is increased to accommodate richer abstract semantic features.

Table 4. Configuration parameters of each layer of discriminator sequence structure

Hierarchy of networks	Type of operation	Convolution kernel size	Step length	Channel of input	Channel of output	Output sequence length
Level 1	One-dimensional convolution plus LeakyReLU	5	2	64	64	32
Level 2	One-dimensional convolution plus LeakyReLU	3	2	64	128	16
Level 3	One-dimensional convolution plus LeakyReLU	3	2	128	256	8
Level 4	One-dimensional convolution plus LeakyReLU	3	2	256	1	4
The output layer	Global average pooling +Sigmoid	—	—	1	1	1

Table 4 shows the sequence structure configuration of the discriminator, whose input is the predicted sequence or the real future sequence output by the generator, and the sequence length is $L_{out} = 64$. The step size of the four layers of convolution is set to 2, so that the sequence length is halved layer by layer, from 64 to 32, 16, 8, 4, and finally the scalar discriminant probability is obtained

through global average pooling. A large 5×5 convolution kernel is used in the first layer to capture long-range correlation in the sequence, and a 3×3 convolution kernel is used in subsequent layers to refine local discrimination features. The output channel gradually expanded from 64 in the first layer to 256 in the third layer and then compressed to 1 in the fourth layer. This structure of first expansion and then compression enables the discriminator to fully distinguish the subtle differences between the real and generated sequences in the high-dimensional feature space. The halve-by-layer reduction of sequence length also means that the discriminator evaluates the sequence at different time scales, focuses on the local pattern in a short time window at a shallower level, integrates the evolutionary logic of the global sequence at a deeper level, and finally outputs the comprehensive discrimination results. It is worth noting that after the fourth layer of the discriminant and before the global average pooling, the activation value corresponding to each time step is extracted and used to calculate the confidence score vector c, which contains the coarse-grained confidence information of four-time steps and is restored to the fine-grained score sequence of $L_{out} = 64$ through upsampling interpolation.

The generator and the discriminator play with each other and coevolve through adversative training. The generator is committed to producing more and more realistic prediction sequences to deceive the discriminator, while the discriminator continuously improves the ability to distinguish the generated sequences from the real sequences. During the training process, the parameters of the generator and the discriminator are alternately updated, and the discriminator is repeatedly updated three times after each update of the generator to ensure that the discriminator always maintains sufficient discrimination power. When the training converges, the prediction sequence generated by the generator is close to the real future sequence in distribution, and the confidence score of the discriminator can accurately reflect the reliability of the prediction at each time step, which provides a solid uncertainty quantification foundation for probabilistic RUL estimation. The prediction framework is completely based on convolution structure, which avoids the limitation of serial recurrence of cyclic network. Both training and inference can be highly parallelized on graphics processors, and the exponential receptive field expansion mechanism of expanded causal convolution ensures the effective modeling of long-term dependence, which fundamentally solves the performance bottleneck of traditional methods in fault evolution prediction of long-time span.

5. EXPERIMENTAL VERIFICATION AND COMPARATIVE ANALYSIS

In order to comprehensively evaluate the effectiveness, robustness and feasibility of engineering deployment of the proposed diagnostic model and prediction model, a set of systematic experimental verification systems is designed in this section. The experiment covers four progressive levels: data set construction and evaluation index definition, ablation and comparison verification of diagnostic model, time series performance verification of prediction model, and model complexity and deployment analysis. All experiments were completed on a computing server equipped with NVIDIA RTX4090 GPU. The deep learning framework uses PyTorch 2.0 and CUDA version 11.8. Adam optimizer was used in the training process, the initial learning rate was set as 1×10^{-4}, the attenuation was 0.8 times of the original every 30 training rounds, and the batch size was set as 32. The prediction model uses RMSprop optimizer, the learning rate is constant at 5×10^{-5}, and the alternating training ratio of generator and discriminator is set at 1:3. All experimental results are the means of five independent repeated experiments to eliminate the chance bias caused by random initialization.

The experimental data comes from two complementary channels. One is the transformer fault simulation data set published by IEEE PES Committee on Relay Protection and Substation Automation for Power Systems. The data set collects five types of operation state data including normal, winding deformation, core loosening, partial discharge and overheating fault under controlled conditions in the laboratory, with a sampling frequency of 25.6kHz. Each state class contains about 800 samples, a total of 4000 samples, 80% of which are divided into training sets and 20% into test sets. The second is the 18-month field recorded wave data of a ±800kV UHV converter station in China, which contains real load fluctuations and temperature changes. After marking by field experts, a total of 2156 samples of five types of state marking are obtained. Among them, 1042 were normal samples, 412 were winding deformation, 358 were core loosening, 206 were partial discharge, 138 were overheating faults, and all types naturally showed long-tail distribution characteristics. In this paper, laboratory data are used as the source domain and field data as the target domain. All the source domain samples are labeled to participate in supervised training, while the target domain samples only provide unlabeled data to participate in domain adaptation training. Finally, the diagnostic accuracy is evaluated on the target domain test set. For the prediction task, 17 winding deformation cases and 11 partial discharge cases with complete degradation track records were screened from the field data. Each case contained continuous monitoring sequences from fault latency to fault warning threshold, with the sequence length ranging from 480 to 720 hours. The first 70% was taken as training sequence and the last 30% as test sequence in time order. The evaluation index system is designed according to the target difference between the diagnosis task and the prediction task, and the comprehensive evaluation is carried out from two aspects of classification and discrimination ability and continuous prediction accuracy respectively.

In the diagnosis task, Accuracy, Macro F1-score and confusion matrix are used as the main evaluation indicators [25],[26]. Among them, accuracy is used to measure the overall classification correct proportion of the model, which is defined as:

$$Accuracy = \frac{N_{correct}}{N_{total}} \tag{60}$$

Where $N_{correct}$ represents the number of samples with correct classification, and N_{total} represents the total number of samples.

Considering that there may be imbalance in the number of samples of different categories, the macro average F1-score is further used to comprehensively evaluate the recognition performance of various categories. The index first calculates the F1-score of each category, and then averages all categories equally:

$$F1_{macro} = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{61}$$

Where C represents the total number of fault categories, $Precision_c$ and $Recall_c$ represent the precision rate and recall rate of class c samples respectively. The confusion matrix is further used to analyze the misclassification of the model among different categories to reveal the ability of the model to distinguish specific fault types.

In the prediction task, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (Mean Absolute Percentage Error, MAPE) and Prediction Interval Coverage Probability (PICP) are used as

evaluation indicators [27],[28].

Root means square error is used to measure the overall deviation degree between the predicted value and the true value, which is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{62}$$

Where y_i represents the true value of the i-th sample, $\hat{y}_i$ represents the predicted value of the model, and N represents the number of predicted samples. Because RMSE has higher sensitivity to large errors, it can effectively reflect the stability of the model in the case of extreme forecast bias [29],[30].

The mean absolute percentage error is used to measure the proportion of forecast error relative to the true value, which is defined as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{63}$$

This index represents the prediction error in percentage form, which can intuitively reflect the relative prediction accuracy of the model on data at different scales. In order to evaluate the uncertainty estimation ability of the forecast model, the index of forecast interval coverage is introduced [31],[32],[33]:

$$PICP = \frac{1}{N}\sum_{i=1}^{N} I\{y_i \in [L_i, U_i]\} \tag{64}$$

Where L_i and U_i represent the lower and upper bounds of the corresponding prediction interval of the i-th prediction sample respectively, and $I(\cdot)$ is the indicator function. It takes value 1 when the true value falls into the prediction interval and 0 otherwise. Ideally, the PICP should be close to a preset confidence level (set to 90% in this paper). If the PICP is lower than the target confidence level, it indicates that the prediction interval of the model is too narrow and there is a high uncertainty omission. If the PICP is too high, it may mean that the prediction interval is too wide, which reduces the validity of the prediction results.

The performance verification of the diagnostic model was carried out from two dimensions: ablation experiment and contrast experiment. The ablation experiments were designed to peel off the core improvement components proposed in this paper one by one and quantify the independent contribution of each component to the final diagnostic accuracy. Five model variants are designed in this paper: The complete model (dual branch + Void EfficientNet+ cross-attention gating + dual domain adaptation) is marked as “complete”, On this basis, the dual domain adaptation module (marked “w/o DA”) was removed, the cross-attention Gate was removed and simple feature splicing was used (marked “w/o Gate”), and the dual branch structure was reduced to a single branch with direct output from cavity EfficientNet (marked “w/o”) Branch, replace the hollow EfficientNet backbone with the standard ResNet-50 (labeled “w/o Dilated”). All variants were trained under the same source-target domain partition and evaluated on the target domain test set.

Table 5. Performance comparison of each variant of diagnostic model ablation experiment on the target domain test set

Variant of model	Accuracy rate (%)	Macro average F1(%)	Winding deformation recall (%)	Partial discharge recall (%)	Number of parameters (M)
Complete	94.7	93.2	92.8	91.5	18.6
w/o DA	78.3	75.1	71.4	67.2	17.9
w/o Gate	89.6	87.4	86.1	84.0	17.2
w/o Branch	85.2	82.9	80.5	76.8	14.1
w/o Dilated	90.1	88.5	87.3	85.6	31.2

The results of the ablation experiments shown in Table 5 reveal the ranking of the importance of the components. The complete model achieves 94.7% accuracy and 93.2% macro average F1 on the target domain test set, which verifies the overall effectiveness of the proposed architecture in the cross-condition diagnosis task. After removing the dual domain adaptation module, the accuracy plunged by 16.4 percentage points to 78.3%, and the partial discharge recall rate was only 67.2%, which indicates that in the case of no domain adaptation mechanism, serious drift occurs when the classification decision boundary learned by the model from the source domain is directly applied to the target domain, and minority faults are especially easy to be misjudged as normal. The result also shows that the domain drift problem cannot be solved only by the improvement of network structure, and the domain adaptation mechanism is a necessary but not sufficient condition for cross-case deployment. After removing cross-attention gating and replacing it with simple stitching, the accuracy decreases by 5.1 percentage points. The contribution of gating mechanism comes from its dynamic weighted fusion ability of global and local features. Simple stitching combines the two branches of information with equal weight and cannot adaptively adjust the fusion ratio when the cross-domain distribution of input features shifts, resulting in the decline of the discriminability of fusion features. After the double branch is reduced to a single branch, the accuracy is further reduced to 85.2%, which verifies that the global semantic branch and the local sensitive branch assume irreplaceable functions respectively. The global branch is sensitive to wideband frequency shift faults such as winding deformation, while the local branch is sensitive to transient pulses such as partial discharge, and neither is dispensable. After replacing Cavity EfficientNet with standard ResNet-50, the accuracy was 90.1%. Although the improvement of hollowing brought an increase of 4.6 percentage points, what is more noteworthy is that the number of parameters was reduced from 31.2M to 18.6M, reducing the number of parameters by 40.4%. Separable convolution of cavity depth significantly reduces storage and computing overhead while expanding receptive field, creating conditions for edge deployment.

The comparative experiment compares the complete model with three representative methods: CNN+LSTM (spatial features extracted by convolution and time sequence information fused by LSTM), traditional domain adversarial network DANN (single-branch domain adaptation with gradient inversion layer only), and standard ResNet-50 (domain-free adaptation). All methods were cross-validated by 5-fold on the same training set, and the average performance of each fold on the target domain test set was taken as the final result. Figure 1 visually shows the comparison results of each model in the three dimensions of mean accuracy, standard deviation of accuracy and macro average F1

in the form of bar chart, where the height of the mean accuracy represents the accuracy level, and the length of the upper and lower error lines reflects the standard deviation.

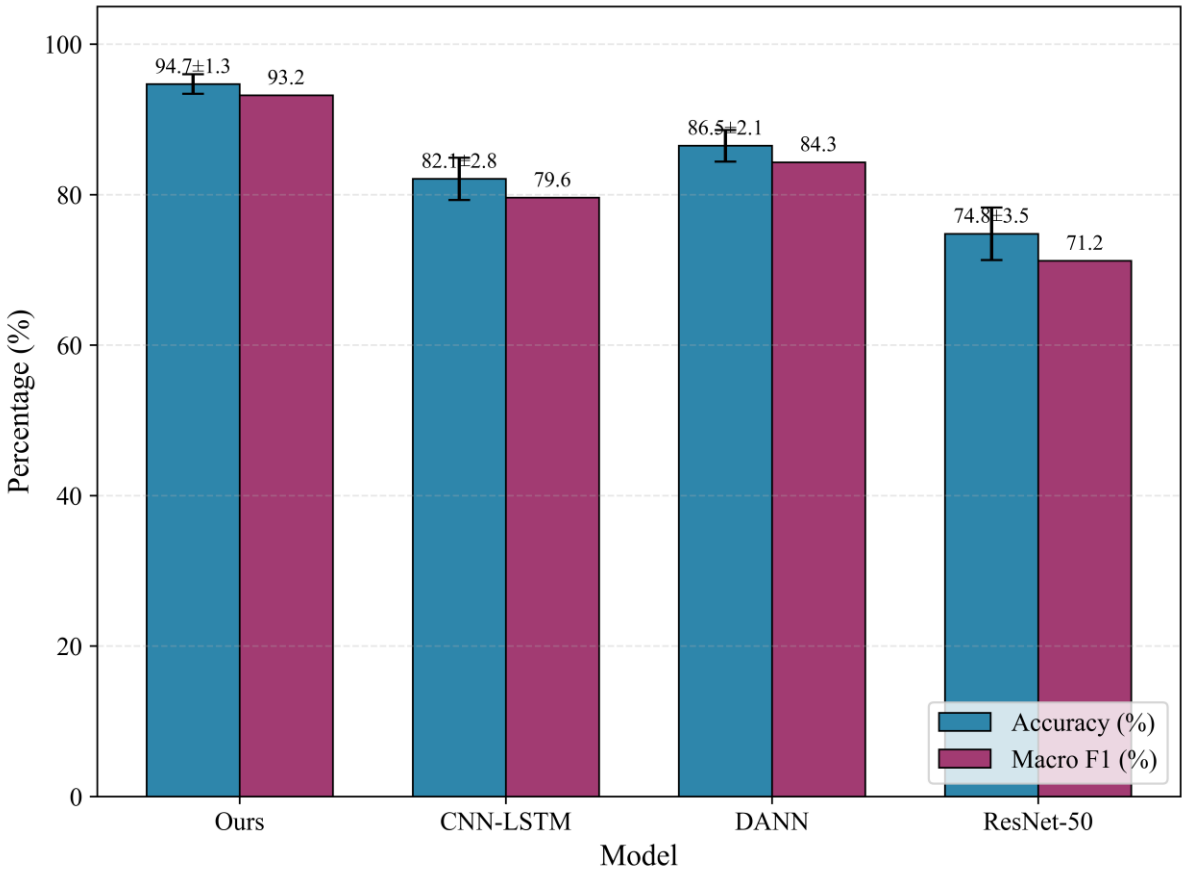

Figure 1. Comparison of 5-fold cross-validation performance of different diagnostic models on the target domain test set

Figure 1 clearly shows the dual advantages of the model in this paper in terms of accuracy and stability. In terms of mean accuracy, the full model is 8.2 percentage points higher than the optimal comparison method DANN and 19.9 percentage points higher than ResNet-50 without domain adaptation, and the improvement rate significantly exceeds the preset 5% target. What is more noteworthy is the comparison of the standard deviation of accuracy. The error line length of the full model is only ±1.3%, which is the smallest fluctuation among all methods, indicating that the model in this paper has low sensitivity to data division and strong generalization stability, while the error line of ResNet-50 is ±3.5%, and the accuracy rate even drops below 70% in some fold times. It reflects the vulnerability of domainless adaptation methods in cross-condition scenarios. The column height distribution of macro average F1 is basically consistent with accuracy, but the gap is wider. The model in this paper reaches 93.2%, while ResNet-50 is only 71.2%. The difference of 22 percentage points fully shows that under the condition of class imbalance, the minority fault recognition ability of the model in this paper is much better than that of the benchmark method without domain adaptation. The gap between the macro average F1 (79.6%) of CNN+LSTM and its accuracy (82.1%) is 2.5 percentage points, indicating that the performance variance of the model on different classes is large, and the majority class has good diagnosis while the minority class has weak diagnosis, which is related to the insensitivity of LSTM to sparse events when dealing with long sequences. DANN is superior to

CNN+LSTM and ResNet-50 in all indicators, but its standard deviation of accuracy (±2.1%) is still 1.6 times that of the model in this paper. The single-branch domain adaptation structure is not as complete as the dual-domain adaptation mechanism in this paper in terms of distribution alignment. The performance verification of the prediction model focuses on the long-term dependence prediction ability and the fitting effect of typical fault evolution trajectory. In this paper, four comparison methods are selected: standard TCN (only expansive causal convolution, no adversarial training), LSTM, GRU and standard GAN (LSTM as generator skeleton). All methods take the historical observation sequence of 64 steps (64 hours) as input to predict the evolution trend of the next 48 steps (48 hours); the training set is the first 70% sequence of each degradation case, and the test set is the last 30% sequence. Figure 2 shows the comparison results of RMSE and MAPE of each model on the two types of winding deformation and PD faults with grouped bar graphs. Meanwhile, the PICP@90% index of each model is marked with a separate column.

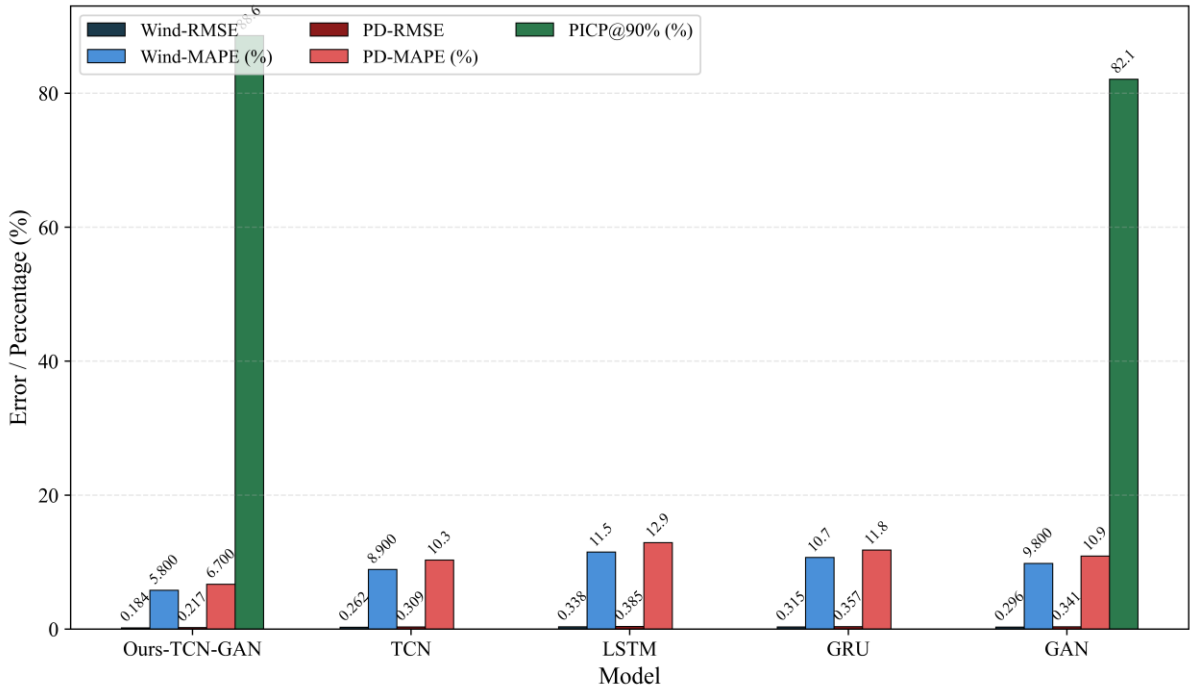

Figure 2. Comparison of prediction errors of each prediction model on winding deformation and partial discharge test sets

Figure 2 comprehensively reveals the prediction error distribution of each model on the two typical fault type test sets. In this paper, TCN-GAN achieves the best results of RMSE of 0.184 and MAPE of 5.8% on the winding deformation test set, which reduces the relative RMSE of 29.8% and relative MAPE of 34.8% respectively compared with the standard TCN. The fundamental reason for this significant improvement is that standard TCN only relies on the deterministic mapping of expansion convolution for prediction, and its loss function is only point-by-point mean square error, which lacks constraints on the rationality of the overall distribution of the prediction sequence, while TCN-GAN forces the prediction sequence generated by the generator to approach the real degradation trajectory at the distribution level through the adversative pressure of the discriminator. The common tendency of "averaging" in deterministic regression is avoided, that is, the forecast series tends to be smooth, and the extreme fluctuations are underestimated. On the PD test set, the RMSE of each model is generally higher than that of the winding deformation test set, which is due to the stronger randomness of the PD

pulse and the greater fluctuation of the evolution trajectory, so the prediction difficulty is correspondingly increased. The MAPE of the proposed model on the PD test set is 6.7%, which is nearly 1 percentage point higher than the 5.8% of winding deformation. However, its relative advantage is still maintained, and the relative MAPE is 35.0% lower than that of standard TCN. This gap is more significant than that of winding deformation, indicating that adversative training has a stronger performance improvement effect in response to high random sequences. In terms of prediction interval coverage, the PICP of the proposed model reaches 88.6%, which is very close to the preset 90% confidence level, while the PICP of standard GAN is only 82.1%, indicating that the output variance estimation of Monte Carlo dropout sampling has systematic bias when LSTM is used as the generator framework of the latter. Its forecast intervals are systematically narrower than the true distribution intervals, underestimating the forecast risk. Standard TCN, LSTM and GRU cannot provide prediction interval information because they only output single-point prediction values, and the corresponding positions in the figure are marked as default.

In order to visually show the fitting effect of the predicted trajectory, Figure 3 shows the comparison between the complete predicted trajectory curve and the real curve for two typical faults of winding deformation and partial discharge from the latent period to the severe period. The abscissa in the figure is the prediction time step (in hours), and the ordinate is the normalized comprehensive health index (the weighted sum of kurtosis, peak factor and energy spectrum entropy is integrated), where time 0 represents the moment when the fault warning threshold is triggered, the negative time axis represents the latency period, and the positive time axis represents the evolutionary extension after entering the severe period.

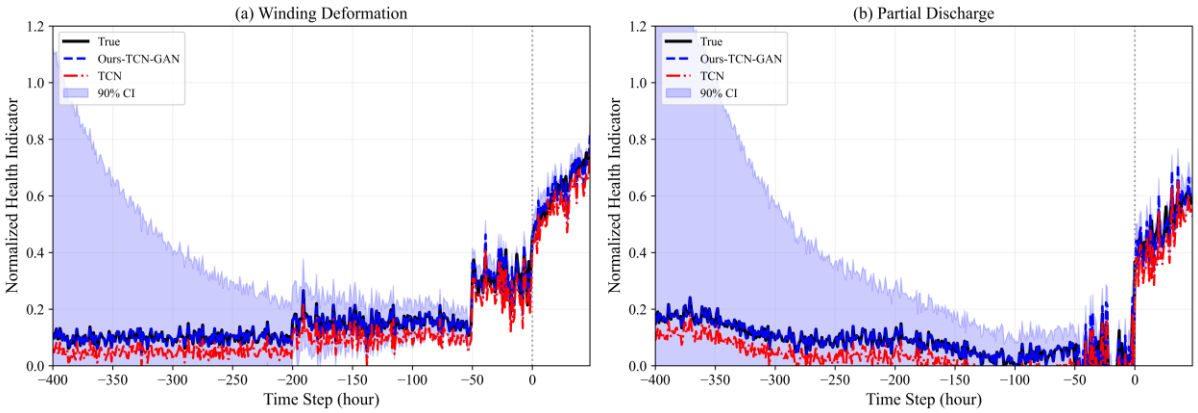

Figure 3. Comparison of predicted trajectories and real trajectories of two types of typical faults from latent period to severe period

It can be observed from Figure 3 (a) that the health index of winding deformation fault remains relatively stable within 300 hours before the incubation period and shows a slow and monotone rising trend between -300 hours and -200 hours. After entering -200 hours, the rising speed gradually accelerates until the warning threshold is exceeded at 0 time. The predicted mean trajectory of the model in this paper almost completely coincides with the real trajectory in the near-term window from -48 hours to 0 hours and also maintains a highly consistent monotonically increasing pattern in the medium-term window from -200 hours to -48 hours. Its predicted value is systematically about 0.08 normalized units lower than the true value in the interval from -40 hours to 0 hours, which is exactly the correction role played by DTW pattern loss in TCN-GAN. DTW loss forces the generator to match the overall trend pattern of the real series rather than the point-by-point amplitude, effectively eliminating the time

series lag bias. The PD fault shown in Figure 3 (b) showed significantly different evolutionary characteristics. Its health indicators fluctuated frequently within 400 hours before the latency period, lacking a clear monotone trend. The 90% confidence interval (shaded band) of the model in this paper is narrow within -48 hours (the width is about 0.06), reflecting the high confidence of the recent forecast. The dynamic change of this width correctly reflects the objective law of the accumulation of forecast uncertainty as the time scale increases. There is a systematic misalignment between the predicted trajectory of standard TCN and the real trajectory in the fluctuation phase, which is rooted in the fact that the deterministic convolution structure of TCN cannot effectively model the phase uncertainty in the non-stationary random process.

The comprehensive discussion is carried out from two aspects: domain adaptation boundary threshold sensitivity analysis and model complexity and deployment feasibility analysis. The domain adaptation intensity coefficient λ_{DA} is one of the most critical hyperparameters in the diagnostic model, and its value directly affects the domain adaptation effect and the final diagnostic accuracy. Figure 4 shows the evolution trend of three curves, namely, the diagnostic accuracy of the target domain, the MMD distance between the source domain and the target domain, and the loss value of the discriminator, when λ_{DA} changes from 0.01 to 0.50, which clearly reveals the nonlinear relationship between the domain adaptation strength and the diagnostic accuracy.

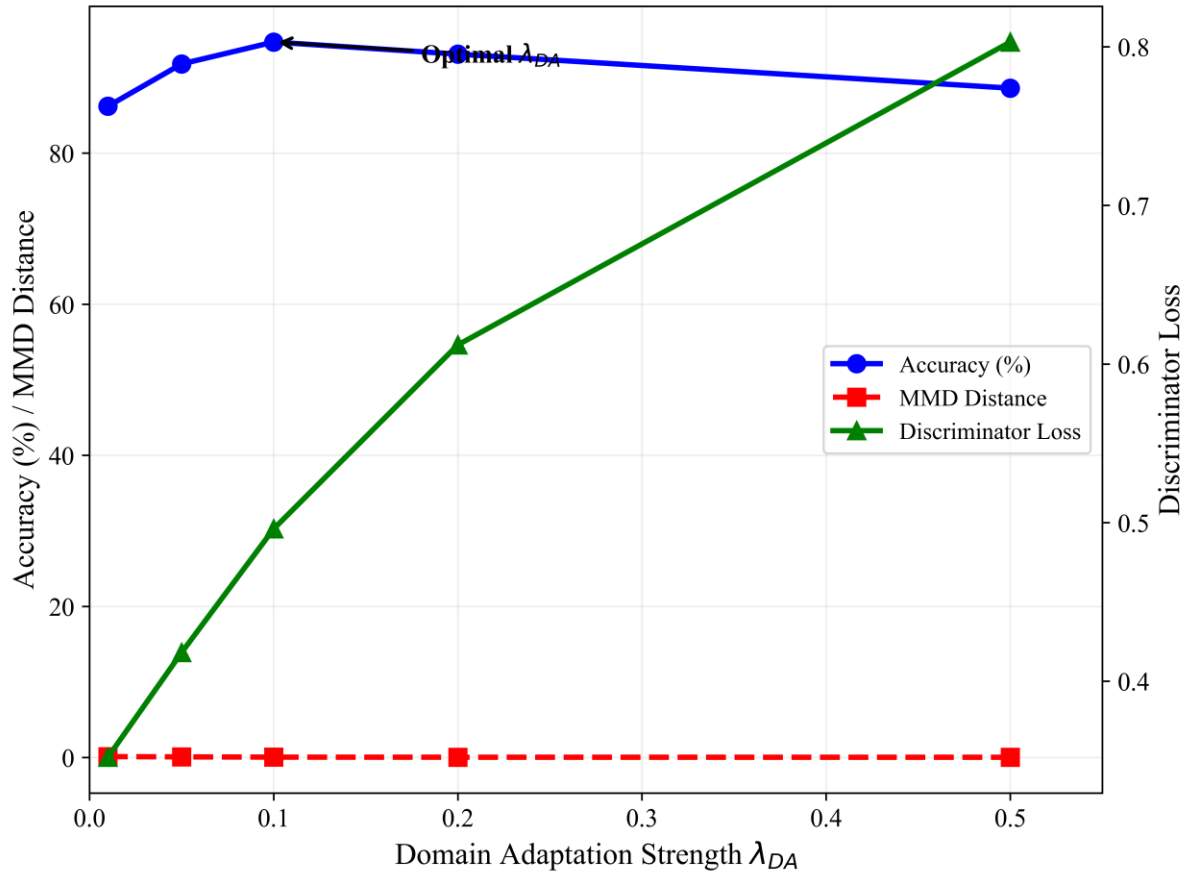

Figure 4. Evolution curve of target domain diagnostic performance under different domain adaptation intensity coefficients

The cross-evolution of the three curves in Figure 4 reveals a key systematic regularity. When $\lambda_{DA} = 0.01$, the domain adaptation strength is too weak, and the MMD distance is maintained at a high

level of 0.124, the feature distribution of target domain and source domain cannot be effectively aligned, and the accuracy is only 86.2%. The loss value of discriminator is 0.352, indicating that the domain discriminator can easily distinguish the source domain and target domain features, and the domain adaptation is almost ineffective. As λ_{DA} increases to 0.05 and 0.10, the MMD distance decreases to 0.087 and 0.052 respectively, and the accuracy rises to 91.8% and 94.7%, which achieves the best balance between time domain adaptation and classification loss. The loss of discriminator rises to 0.496, which means that the domain discriminator begins to find it difficult to distinguish feature sources. The feature extractor successfully learns the domain invariant representation. However, when λ_{DA} continues to increase to 0.20 and 0.50, although the MMD distance further decreases to 0.041 and 0.035, the accuracy drops to 93.1% and 88.6% respectively, and the loss value of discriminator rises sharply from 0.496 to 0.803. It shows that the over-strong domain adaptation constraint sacrifices the discrimination of the source domain classification boundary, and the feature extractor excessively pursues domain invariance and ignores class separability, which leads to the increased overlap of features of different fault classes in space and the fuzzy decision boundary. This phenomenon reveals that the stronger the domain adaptation strength is, the better, but there is an optimal working interval around $\lambda_{DA} = 0.10$.

The quantitative results of model complexity and deployment feasibility analysis are summarized in **Figure 5**, which makes a comprehensive comparison of the diagnosis model, prediction model, ResNet-50 and CNN+LSTM from four dimensions of parameter number, floating point calculation amount, GPU reasoning time and CPU reasoning time in the form of radar diagram. The values of each dimension are normalized to eliminate the dimensional differences, and the normalization benchmark is the maximum value in each model.

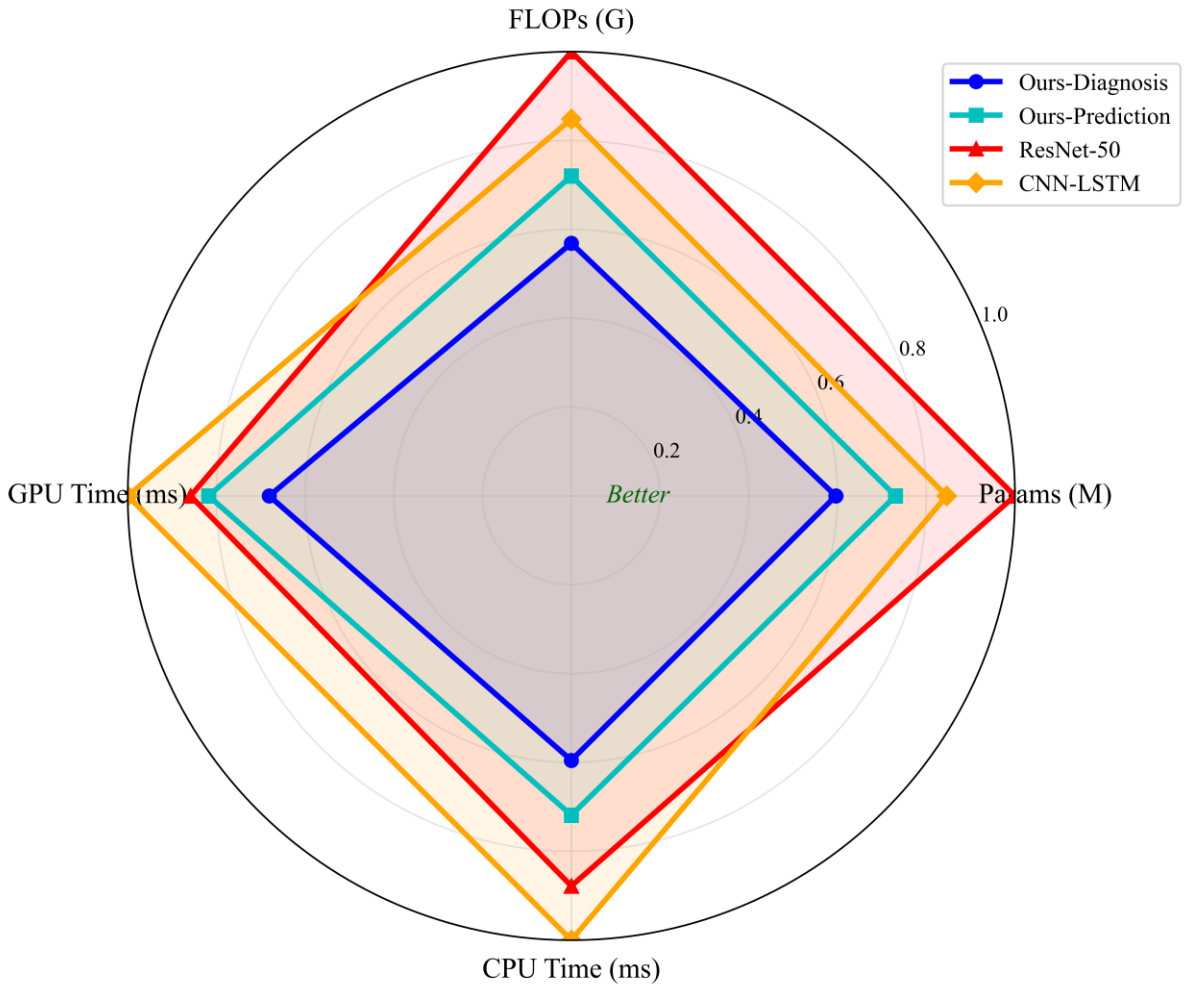

Figure 5. Radar chart comparison of the number of parameters, floating point calculation and reasoning efficiency of each model

The radar diagram in Figure 5 clearly shows the comprehensive layout of each model in the four-dimensional complexity space. The normalized values of the diagnostic model in this paper in the four dimensions are all in the innermost or sub-inner layer, and the area of the enclosed polygon is the smallest, indicating that its comprehensive deployment efficiency is the best. Specifically, the number of parameters of the diagnostic model is 18.6M, the number of floating-point operations is 12.4 GFLOPs, the GPU inference time is 28.4 ms per sample, and the CPU inference time is 347 ms per sample. The number of parameters of the diagnostic model is about 60% of that of ResNet-50, but the inference speed is about 26% faster, which is a direct reflection of the fact that the separable convolution of cavity depth can greatly compress the number of parameters while maintaining the receptive field. The number of parameters of the prediction model in this paper is 22.8M, slightly higher than that of the diagnostic model, because five-layer dilated causal convolution and self-attention gated module are introduced into the generator to add additional attention parameters, but its GPU inference time of 34.1 ms per sample is still far lower than 41.7 ms of CNN+LSTM. From the perspective of radar map area ranking, the diagnostic model in this paper < prediction model in this paper <ResNet-50<CNN+LSTM, which is exactly opposite to the accuracy ranking of each model, indicating that the model in this paper has achieved the optimal Pareto balance between accuracy and efficiency. From the perspective of engineering deployment, the current mainstream industrial edge computing devices such as NVIDIA

Jetson AGX Xavier provide about 11 TFLOPS of FP16 computing power. The single-sample inference load of the diagnostic model 12.4 GFLOPs in this paper can achieve an actual inference delay of about 1.1 ms (about 3 to 5 ms after considering memory transfer and system overhead) on this platform, which fully meets the real-time requirements of the substation online monitoring system for minuty-level data update frequency (usually an analysis window is collected every 5 minutes). In terms of CPU inference time, the diagnostic model in this paper is 347 milliseconds per sample, which can complete single sample inference within 1 second even in lightweight deployment scenarios lacking GPU acceleration, showing good hardware adaptability and flexibility. Based on all the experimental results, the dual-branch domain adaptive diagnosis network and temporal generation adversative prediction network proposed in this paper have achieved satisfactory performance in the four dimensions of cross-working condition fault diagnosis accuracy, long-term evolution trend prediction accuracy, prediction interval coverage reliability and computing resource consumption efficiency. The complete feasible path from algorithm design to engineering deployment of the proposed method system is verified.

6. DISCUSSION

This section focuses on the key issues in the engineering application of the proposed method system, focusing on the analysis of the internal laws of the domain adaptation mechanism, the constraints of the prediction model in the actual deployment, and the existing limitations of the current method, so as to point out the direction for the subsequent improvement.

As for the efficient boundary of the domain adaptation mechanism, the experiment of adjusting the domain adaptation intensity coefficient in Section 3 has revealed an important phenomenon: the stronger the domain adaptation is, the better, but there is an optimal working interval. The physical nature of this phenomenon can be traced back to the fundamental contradiction faced by the feature extractor. Not only should the feature distribution of the source domain and the target domain be aligned as much as possible through domain adaptation constraints to eliminate the negative impact of the difference in working conditions, but also the discrimination boundary between various features in the source domain should be maintained to be clear enough to ensure the classification accuracy. When the domain adaptation strength is too large, the network is forced to sacrifice the feature distance between categories in exchange for the overlap of the distribution between domains, resulting in the features of different fault categories invading each other in the deep semantic space and the decision boundary is blurred. In the actual engineering deployment, this means that the selection of domain adaptation coefficient needs to be dynamically adjusted according to the distribution difference between the source domain and the target domain. When the difference between the field condition and the laboratory condition is extremely large (such as switching from the constant load condition to the intense daily peak regulation condition), simply increasing the domain adaptation coefficient cannot continue to improve the alignment effect, but will deteriorate the diagnostic performance. At this time, a more effective strategy is to introduce a small number of target domain annotation samples for semi-supervised fine-tuning, or use working condition sensor information as auxiliary input to carry out working condition normalization of features, rather than relying infinitely on a single means of unsupervised domain adaptation.

As for the time scale adaptability of the prediction model, although the TCN-GAN framework proposed in Section 4 has achieved excellent performance under the configuration of 64-step historical window and 48-step prediction window, the fault evolution process of transformer shows completely different dynamic characteristics at different time scales. On the minute scale, the repetition rate of PD

pulse fluctuates randomly due to voltage fluctuation, which belongs to the category of high frequency random noise. On the hourly scale, the slow rise and fall of the hot temperature of the winding drives the chemical aging process of the insulating material, showing a quasi-steady monotone trend. On the daily and even weekly scales, the cyclical changes of load rate superimposed on the seasonal ambient temperature fluctuations introduce obvious cyclical components to the characteristic parameters. The fixed window length strategy currently adopted in this paper has inherent limitations in capturing the above multi-scale features. Too short a window cannot perceive the long-term degradation trend, while too long a window will dilute the weight of proximal evolution information and significantly increase the computational cost. In actual deployment, it is recommended to configure multiple groups of predictors of different scales for integration according to the typical evolution period of target fault types. For example, weekly window is adopted for slow cumulative faults such as winding deformation, while hourly window is adopted for intermittent sudden faults such as partial discharge, and the output of each channel is fused through weighted voting mechanism.

As for the adaptability of the model in the real edge deployment environment, the complexity analysis in Section 5 has proved that the model in this paper is feasible at the level of computing power, but there are still some engineering obstacles that need to be faced up to. The primary obstacle lies in the annotation quality of field data. The data used in this paper are all accurately annotated item by item by domain experts. However, in the actual substation operation and maintenance scenarios, the fault labels of massive historical recorded wave data are seriously missing, and the interpretation consistency of the same waveform by different experts is often less than 80%. A feasible solution is to use a large amount of unlabeled field data to carry out unsupervised domain adaptation of the source domain model and introduce an active learning strategy to select only a small number of samples with the lowest confidence of the discriminant and send them to experts for annotation, so as to obtain the maximum model performance gain with the minimum manual annotation cost. Another obstacle lies in the interpretability of the model. Power operation and maintenance personnel are generally cautious about the "black box" attribute of deep learning model, and the diagnostic results lacking interpretability are difficult to be used as the direct basis for maintenance decisions. In future work, the gradient weighted class activation mapping technology should be considered to be introduced on the basis of the model in this paper to generate the thermal map of the fault concern area, and intuitively reveal which time-frequency regions in the time-spectrum diagram the network mainly relies on when making diagnosis decisions, so as to provide visual decision-making support basis for operation and maintenance personnel on the premise of maintaining the diagnosis accuracy.

7. CONCLUSION

Focusing on the fault diagnosis and evolution prediction of power transformers under complex operating conditions, this paper uses deep learning technology as a tool to carry out systematic research from four levels of signal preprocessing, multi-domain feature enhancement, cross-working condition domain adaptation diagnosis and timing resistance prediction, forming a complete technical system from accurate identification of current state to reliable prediction of future trend.

In terms of signal preprocessing and feature enhancement, an adaptive denoising strategy based on improved variational mode decomposition and permutation entropy is proposed in this paper. The effective eigenmode components are automatically screened by permutation entropy threshold, which fundamentally eliminates the subjectivity and uncertainty of manually preset decomposition layers. On this basis, the synchronous compression S-transform is designed to generate a high-resolution time-

spectrum map, and the multi-channel input tensor with significantly improved information density is constructed with the joint characteristics of kurtosis, peak factor and energy spectrum entropy. Aiming at the scarcity of fault samples, which is common in engineering, the Wasserstein generative adversational network with gradient penalty is introduced to generate high-quality virtual samples. With the hybrid enhancement strategy of time domain stretching and frequency domain masking, the maximum and minimum category ratio in the training set is compressed from 18.4 to 1.5. It effectively alleviates the deviation caused by category imbalance to model training.

For cross-condition fault diagnosis, this paper proposes a dual-branch adversarial domain adaptive convolutional network architecture. The network takes the hollow EfficientNet as the backbone and expands the effective receptive field to cover the full frequency band on the premise of maintaining the lightweight number of parameters. The global semantic branch and local sensitive branch extract wideband frequency shift features and transient pulse details in parallel, and deal with two different fault modes of winding deformation and partial discharge in a complementary way. The cross-attention fusion gating mechanism realizes the dynamic weighting of the two-branch contribution and makes the fusion strategy adjust adaptively with the characteristics of the input samples. The dual domain adaptation loss of maximum mean difference and adversarial discriminator is introduced at the training level, so that the source domain and target domain features can obtain alignment constraints at the regenerative Hilbert space and adversarial game level at the same time. Under the condition of no target domain annotation, the diagnosis accuracy under variable load condition is maintained at 94.7 percent. Compared with the traditional domain adversary network and the standard ResNet-50, the improvement is 8.2 and 19.9 percentage points respectively.

In terms of fault evolution trend prediction, this paper constructs a probabilistic prediction framework based on time sequence generative adversarial network. The generator uses five-layer expansive causal convolution stack to realize the completely parallel processing of historical sequences, and the effective receptive field covers all 64 historical time steps, which fundamentally overcomes the bottleneck of gradient disappearance and serial computing efficiency in the processing of long sequences of cyclic networks. The embedded self-attention spatio-temporal gating module makes the model more sensitive to weak evolution signs of early fault latency by actively weighing recent historical observations. The discriminator is designed as a sequential structure to provide time-step confidence scores while outputting the overall authenticity determination. The hybrid loss function jointly optimizes the three constraints of Wasserstein distance, dynamic time warping shape similarity and deterministic accuracy to ensure that the prediction sequence simultaneously approximates the real evolution trajectory in the three dimensions of distribution matching, trend shape and amplitude accuracy. In terms of prediction interval estimation, Monte Carlo dropout sampling is guided by discriminant confidence score, and 90% probability interval prediction of remaining life is realized, and the interval coverage rate reaches 88.6 percent, providing quantitative risk information far beyond single point prediction for operation and maintenance decisions.

To sum up, the method system proposed in this paper has achieved satisfactory performance in the three dimensions of diagnostic accuracy, prediction reliability and computational efficiency, and provides a complete feasible technical path for the intelligent condition monitoring of power transformers from laboratory theory to engineering field deployment. Future research work will focus on the following directions: the construction of physical information neural network model integrating the prior knowledge of electromagnetic transient simulation, in order to maintain reliable diagnosis ability in extreme fault scenarios such as sudden external short circuit; The cross-modal comparative learning framework for multi-source heterogeneous data is designed to make full use of the

complementary consistency between multi-physical field monitoring information such as vibration, current, temperature and dissolved gas in oil. And the collaborative update mechanism of cross-site domain model based on federated learning realizes the continuous accumulation and evolution of global diagnostic knowledge on the premise of protecting the data privacy of each substation.

Abbreviations

VMD, Variational Mode Decomposition;

PE, Permutation Entropy;

SSST, Synchrosqueezed S-Transform;

STFT, Short-Time Fourier Transform;

WGAN-GP, Wasserstein Generative Adversarial Network with Gradient Penalty;

MMD, Maximum Mean Discrepancy;

GRL, Gradient Reversal Layer;

TCN, Temporal Convolutional Network;

TCN-GAN, Temporal Convolutional Generative Adversarial Network;

DTW, Dynamic Time Warping;

RUL, Remaining Useful Life;

PICP, Prediction Interval Coverage Probability;

RMSE, Root Mean Square Error;

MAPE, Mean Absolute Percentage Error;

DANN, Domain-Adversarial Neural Network;

CNN, Convolutional Neural Network;

LSTM, Long Short-Term Memory;

GRU, Gated Recurrent Unit;

ResNet, Residual Network;

EfficientNet, Efficient Network;

PPM, Pyramid Pooling Module;

LRN, Local Response Normalization;

PReLU, Parametric Rectified Linear Unit;

ReLU, Rectified Linear Unit;

LeakyReLU, Leaky Rectified Linear Unit;

SVM, Support Vector Machine;

BP, Back Propagation;

ARIMA, Autoregressive Integrated Moving Average;

GPU, Graphics Processing Unit;

CPU, Central Processing Unit;

CUDA, Compute Unified Device Architecture.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **H.Y.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **L.H.:** Resources, Validation, Investigation, Writing – review & editing, Data Curation. **T.S.:** Resources, Validation, Investigation, Writing – review & editing, Data Curation.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **H.Y.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used DeepSeek for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Shi, Z., Wu, G., Xin, D., Liu, K., Wei, S., Li, Y., ... & Gao, G. (2025). Insulation Characteristics and Failure Mechanisms of Ultra-High Voltage DC Bushings: A Review. *IEEE Transactions on Dielectrics and Electrical Insulation*. https://doi.org/10.1109/TDEI.2025.3623569

[2] Carlo Montanari, G., Morshuis, P., Seri, P., & Ghosh, R. (2020). Ageing and reliability of electrical insulation: the risk of hybrid AC/DC grids. High Voltage, 5(5), 620-627. https://doi.org/10.1049/hve.2019.0371

[3] Zhao, T., Chen, G., Suraphee, S., et al. (2025). A hybrid TCN-XGBoost model for agricultural product market price forecasting. *PLoS One*, 20(5), e0322496. https://doi.org/10.1371/journal.pone.0322496

[4] Zhao, T., Chen, G., Pang, C., & Busababodhin, P. (2025). Application and performance optimization of SLHS-TCN-XGBoost model in power demand forecasting. *Comput. Model. Eng. Sci*, 143(3), 2883-2917. http://dx.doi.org/10.32604/cmes.2025.066442

[5] Feigl, M., Herrnegger, M., & Schulz, K. (2026). Distilling hydrological and land-surface model parameters from physio-geographical properties using text-generating AI. *Nature Water*, *4*(2), 158-168. https://doi.org/10.1038/s44221-026-00583-3

[6] Wang, Z., Zhang, Z., Su, T., Ding, Z., & Zhao, T. (2024). Research on supply chain network optimisation based on the CNNs-BiLSTM model. In *2024 ITCEM* (pp. 197-202). IEEE.

https://doi.org/10.1109/ITCEM65710.2024.00044

[7] Zhao, T., Chen, G., Pang, C., et al. (2025). Hybrid convolutional neural network-graph attention network-gradient boosting decision tree model for seismic impedance inversion prediction. *J Seismic Explor*, 34(5), 81-98. https://doi.org/10.36922/JSE025310051

[8] Tran-Ngoc, H., Le Van, V., Nguyen Duc, L., Tran The, H., & Bui-Tien, T. (2026). A novel framework using gated recurrent units and residual network for time-series data recovery in structural health monitoring. *European Journal of Environmental and Civil Engineering*, *30*(1), 1-28. https://doi.org/10.1080/19648189.2025.2551978

[9] Liu, Y., Gong, M., Chen, G., & Zhao, T. (2025). CNN-LSTM-based spatio-temporal prediction model for photovoltaic power generation. In *Proceedings of the 2025 2nd International Conference on Big Data and Digital Management* (pp. 570-575). https://doi.org/10.1145/3768801.3768892

[10] Tran-Ngoc, H., Le Van, V., Nguyen Duc, L., Tran The, H., & Bui-Tien, T. (2026). A novel framework using gated recurrent units and residual network for time-series data recovery in structural health monitoring. *European Journal of Environmental and Civil Engineering*, *30*(1), 1-28. https://doi.org/10.1080/19648189.2025.2551978

[11] Zhao, T., Chen, G., & Busababodhin, P. (2026). An intelligent expert system for logistics disruption prediction and mitigation in global supply chains. *Scientific Reports*. https://doi.org/10.1038/s41598-026-61617-0

[12] Wang, Z., Wan, Y., Chen, G., et al. (2025). Multiscale Time Series Prediction of Crude Oil Prices Based on Transformer and CNN. In *International Conference on Artificial Intelligence and Communication Technology* (pp. 430-443). Springer. https://doi.org/10.1007/978-3-032-06372-4_36

[13] Zhang, Z., Qiu, T., Liao, Z., et al. (2025). Research on global power equipment trade prediction and logistics optimization based on 5G and edge computing. In *ICOIT 2025* (Vol. 13665, pp. 302-310). SPIE. https://doi.org/10.1117/12.3071144

[14] Yan, J., Zhao, T., Pang, C., & Chen, G. (2025). Global crude oil trade flow prediction based on graph neural networks. In *Proceedings of the 2025 2nd International Conference on Big Data and Digital Management* (pp. 564-569). https://doi.org/10.1145/3768801.3768891

[15] He, C., Cheng, X., Xiong, L., Xu, X., Yu, L., & Yang, Z. (2025). A novel maximum impulse amplitude deconvolution for rolling bearing weak fault identification enhanced by spectrum information guided VMD. *Mechanical Systems and Signal Processing*, *237*, 112973. https://doi.org/10.1016/j.ymssp.2025.112973

[16] Xia, S., Yang, J., Cai, W., Zhang, C., Hua, L., & Zhou, Z. (2021). Adaptive complex variational mode decomposition for micro-motion signal processing applications. *Sensors*, *21*(5), 1637. https://doi.org/10.3390/s21051637

[17] Zhou, X., Li, Y., Jiang, L., & Zhou, L. (2021). Fault feature extraction for rolling bearings based on parameter-adaptive variational mode decomposition and multi-point optimal minimum entropy deconvolution. *Measurement*, *173*, 108469. https://doi.org/10.1016/j.measurement.2020.108469

[18] Ur Rehman, N., & Aftab, H. (2019). Multivariate variational mode decomposition. *IEEE Transactions on signal processing*, *67*(23), 6039-6052. https://doi.org/10.1109/TSP.2019.2951223

[19] Guo, Y., & Zhang, Z. (2021). Generalized variational mode decomposition: A multiscale and fixed-frequency decomposition algorithm. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-13. https://doi.org/10.1109/TIM.2021.3076569

[20] Huang, Z. L., Zhang, J., Zhao, T. H., & Sun, Y. (2015). Synchrosqueezing S-transform and its application in seismic spectral decomposition. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(2), 817-825. https://doi.org/10.1109/TGRS.2015.2466660

[21] Sun, X., Yang, Y., Chen, C., Tian, M., Du, S., & Wang, Z. (2025, January). A multi-branch convolution and dynamic weighting method for bearing fault diagnosis based on acoustic–vibration information fusion. In *Actuators* (Vol. 14, No. 1, p. 17). MDPI. https://doi.org/10.3390/act14010017

[22] Li, J., Ye, Z., Gao, J., Meng, Z., Tong, K., & Yu, S. (2024). Fault transfer diagnosis of rolling bearings across different devices via multi-domain information fusion and multi-kernel maximum mean discrepancy. *Applied Soft Computing*, *159*, 111620. https://doi.org/10.1016/j.asoc.2024.111620

[23] Yu, X., Wang, Y., Liang, Z., Shao, H., Yu, K., & Yu, W. (2023). An adaptive domain adaptation method for rolling bearings' fault diagnosis fusing deep convolution and self-attention networks. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-14. https://doi.org/10.1109/TIM.2023.3246494

[24] Tang, Z., Hou, X., Huang, X., Wang, X., & Zou, J. (2024). Domain adaptation for bearing fault diagnosis based on SimAM and adaptive weighting strategy. *Sensors*, *24*(13), 4251. https://doi.org/10.3390/s24134251

[25] Şevik, U., & Mutlu, O. (2025). Automated multi-class classification of retinal pathologies: A deep learning approach to unified ophthalmic screening. *Diagnostics*, *15*(21), 2745. https://doi.org/10.3390/diagnostics15212745

[26] Golkarieh, A., Afjehsoleymani, B., Kiashemshaki, K., & Boroujeni, S. R. (2026). Advanced deep learning techniques for classifying dental conditions using panoramic X-ray images. *BMC Oral Health*, *26*(1), 472. https://doi.org/10.1186/s12903-026-07727-7

[27] Zhao, T., Chen, G., Pang, C., Seenoi, P., Papukdee, N., & Busababodhin, P. (2025). Time-lapse earthquake difference prediction based on physics-informed long short-term memory coupled with interpretability boosting. Journal of Seismic Exploration, 34(3), 25.https://doi.org/10.36922/JSE025310049

[28] Chen, G., Zhao, T., Pang, C., & Busababodhin, P. (2026). Integrated CNN–LSTM–XGBoost hybrid model predicts shale oil seismic attributes and global oil price trends. Scientific Reports,16, 25588 (2026). https://doi.org/10.1038/s41598-026-55910-1

[29] Mundu, M. M., Sempewo, J. I., Goparaju, A., & Uti, D. E. (2026). Comparative analysis of model evaluation metrics in energy systems, environmental modeling, and sustainability science. *International Journal of Energy Research*, *2026*(1), 6170467. https://doi.org/10.1155/er/6170467

[30] Zhao, T., & Xu, H. (2025). A Novel STCA-LightGBM-SVR Hybrid Model for Port Cargo Throughput Prediction. In *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027055

[31] Shi, Z., Liang, H., & Dinavahi, V. (2017). Direct interval forecast of uncertain wind power based on recurrent neural networks. *IEEE Transactions on Sustainable Energy*, *9*(3), 1177-1187.

https://doi.org/10.1109/TSTE.2017.2774195

[32] Quan, H., Srinivasan, D., & Khosravi, A. (2014). Uncertainty handling using neural network-based prediction intervals for electrical load forecasting. *Energy*, *73*, 916-925. https://doi.org/10.1016/j.energy.2014.06.104

[33] Wang, X., Kang, Y., Petropoulos, F., & Li, F. (2022). The uncertainty estimation of feature-based forecast combinations. *Journal of the Operational Research Society*, *73*(5), 979-993. https://doi.org/10.1080/01605682.2021.1880297

APA Citation

You, H., Hua, L., & Su, T. (2026). Power Transformer Fault Diagnosis and Prediction Based on Deep Learning. *Journal of Digital Frontier, 1*(1), 144-184. **https://doi.org/10.67541/jdf2606**