Research on Defect Identification Method of Main Transformer Core Clamping Based on the Collaborative Strategy of Large and Small Models

Peng Li, Yunlong Liu*, Baiyuan Liu, Xiangcheng Kong, Weifeng Wang, Chuanhui Zhang
State Grid Shangqiu Power Supply Company, Shangqiu, Henan, China
**Corresponding Author*

Abstract: With the intelligent development of the power system, the main transformer, as a key piece of equipment in the power system, the health of its operating state directly affects the stability of the power grid. Due to its complex structure and harsh operating environment, the core clamping parts of the main transformer are prone to defects such as loosening, cracking and rusting. The traditional defect identification methods have obvious deficiencies in terms of accuracy and efficiency. To this end, this paper proposes an intelligent recognition method based on the collaborative strategy of large and small models. By combining the rapid response capability of small models on the on-site side with the high-precision analysis capability of large models in the cloud, an efficient and intelligent main transformer core clamping defect recognition system is constructed. Experiments show that this method significantly improves the response speed and system practicability while ensuring the recognition accuracy.

Keywords: Defect Identification; Collaborative Strategy of Large and Small Models; Main Transformer Core Clamping Piece; Artificial Intelligence

1. Introduction

The main transformer is a key device in the power system, and the reliability of its operation directly determines the safety and stability of the transmission and distribution system. As an important mechanical structure in the main transformer, the core clamping, if it has defects such as loosening, cracking and rusting, will not only lead to increased vibration and noise, but also may cause serious malfunctions, shorten the service life of the equipment, and even cause power outages. Therefore, it is of great significance to conduct efficient and accurate fault identification for the core clamps of the main transformer.

Traditional fault detection techniques mainly include methods such as thermal model prediction, frequency response analysis, and partial discharge detection, which can provide multi-dimensional references for the operating status of transformers [1,2]. However, such methods usually require professional equipment and manual interpretation, have slow response speeds and poor real-time performance, and are difficult to meet the dual requirements of modern intelligent operation and maintenance for high efficiency and accuracy.

In recent years, artificial intelligence and machine learning methods have received extensive attention in the field of transformer fault diagnosis. The deep learning method based on deep belief network (DBN) and BP neural network combined with DGA features has been proven to have a good effect in fault type identification. Furthermore, Zhang et al. [3] proposed a DGA model based on DBN, which achieved automatic mapping from characteristic gases to fault types and had better recognition rates and generalization capabilities. Beyond deep learning, various machine learning and hybrid methods have also been widely explored. For example, Wu et al. [4] introduced residual neural network combined with frequency response data to achieve accurate location of winding faults, Lei et al. [5] adopted the cloud-edge collaborative defect detection method based on Yolo network and incremental learning. Furthermore, Li et al. [6] proposed a hybrid model combining Swin Transformer and convolutional neural network for surface defect detection. The experimental results show that this method significantly improves the detection accuracy and robustness on multiple public datasets.

Although the above-mentioned methods perform outstandingly in terms of diagnostic accuracy, they generally have limitations such as high deployment costs, large computing latency, and lack of real-time response capabilities. In the operation and maintenance of power substation equipment, on-site inspection of equipment requires rapid response and immediate handling. Therefore, deploying computing capabilities on the edge side has become a trend. Edge Intelligence and cloud-edge collaborative architecture are gradually emerging, dedicated to pushing AI to the edge of the network and enhancing the timeliness of inference and the efficiency of resource utilization. In recent years, the Transformer architecture has demonstrated strong potential in emphasizing long-term dependencies and global feature modeling, and is gradually being integrated into defect detection and sequence analysis tasks. For example, "Defect Transformer" combines the CNN and Transformer architectures to achieve collaborative modeling of local and global features in the surface defect detection task, improving the detection efficiency and accuracy. The industrial manufacturing field has also proposed injecting the Transformer-based attention mechanism into fault detection tasks, emphasizing its advantages in spatio-temporal structure modeling [7-10].

In conclusion, the existing literature indicates that although traditional monitoring technologies are mature, they have a slow response. Deep learning and hybrid models perform well in terms of diagnostic accuracy, but they are mostly used in cloud or experimental environments, with high deployment costs. The combination of edge intelligence and lightweight models has the advantage of real-time on-site response. The Transformer architecture offers superior global feature capture capabilities. It is precisely against this backdrop that this paper proposes a defect identification method based on the collaborative strategy of large and small models: deploying small models at the edge for rapid initial judgment, and deploying large models in the cloud for in-depth analysis. The collaboration between the two takes into account both efficiency and accuracy, thereby providing a systematic and practical new path for the intelligent defect identification of main variable core clamps.

2. Design of the Collaborative Strategy Framework for Large and Small Models

With the rapid development of artificial intelligence technology in the field of power equipment condition monitoring, how to improve the efficiency of model deployment while ensuring recognition accuracy has become the core issue in building practical intelligent diagnostic systems. Traditional large models (such as ResNet, Transformer, etc.) have strong capabilities in feature extraction and semantic modeling, and can accurately identify complex, fuzzy, and multi-type defects. They are suitable for task scenarios with variable shapes and complex lighting, such as main transformer core clamping pieces. However, such models have high computational complexity and high requirements for computing power and storage resources, which is not conducive to deployment at on-site terminals. Relatively speaking, small models (such as MobileNet, ShuffleNet, EfficientNet-lite, etc.) have the advantages of light structure and fast response, and are suitable for deployment on edge devices for real-time judgment. However, it is limited by the number of parameters and the depth of the network, and its recognition accuracy is relatively low when dealing with high-complexity or multimodal defects. Therefore, this paper proposes a large and small model collaboration strategy based on "cloud-edge collaboration", integrating the advantages of both sides through the approach of "real-time early warning of small models on the edge side + in-depth analysis of large models in the cloud", taking into account both response speed and recognition accuracy, to construct an intelligent diagnosis system for transformer defects with practical value.

This collaborative architecture mainly consists of three core modules: edge perception and recognition, initial judgment of small models, and cloud-based analysis and decision-making. Among them, the edge end is deployed in the on-site transformer environment, integrating high-definition industrial cameras and small models. It can achieve real-time collection and preprocessing of core clamping images, and quickly classify and initially alarm common defects such as clamping rust, loosening, and cracking. The response time is controlled within 50ms, meeting the requirements of on-site inspection. Once the edge recognition result reaches the warning threshold or the confidence level is insufficient, the image data and judgment information are immediately reported to the

cloud. The cloud system has stronger computing power support and storage capacity, runs high-capacity deep models, conducts fine-grained feature extraction, multi-category reasoning and historical record comparison on uploaded images, and outputs diagnostic suggestions and maintenance priorities based on the fault mode library. The cloud can also synchronize the optimized model weights or recognition strategies back to the edge through a lightweight protocol, enhancing the subsequent inference effect and enabling online fine-tuning of small models.

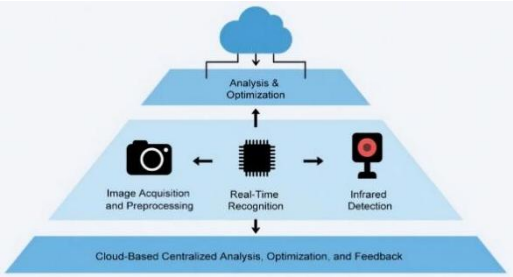

Figure 1. The Collaborative Working Architecture of Large and Small Models

As shown in Figure 1, the core advantage of this strategy lies in hierarchical modeling and clear task division: small models efficiently cover the primary judgment tasks of on-site equipment, ensuring the timeliness of detection. Large models conduct in-depth analysis of global features to enhance recognition accuracy. The two achieve result fusion and information closed loop through communication mechanisms, thereby taking into account both system response efficiency and diagnostic reliability. In addition, this strategy supports modular deployment and dynamic updates, and has good scalability and adaptability, capable of adapting to the main transformer operation environment under different substations and hardware conditions. Through the collaborative mechanism of large and small models, the system can not only promptly detect and precisely locate the defects of the core clamping pieces of the main transformer, but also gradually accumulate defect samples and optimize the diagnostic model during long-term operation and maintenance, ultimately building a new paradigm for transformer operation and maintenance oriented towards intelligent maintenance.

3. Construction and Optimization of Defect Recognition Models

3.1 Structural Design of Small Models and Large Models

When constructing the defect recognition model for the core clamping of the main transformer, this paper adopts a collaborative design strategy of large and small models: the small model on the edge side is used for real-time detection, and the large model in the cloud is used for high-precision analysis. The small model part selects MobileNetV2 as the basic architecture, which takes into account both lightweight and feature extraction capabilities. Its core adopts a Depthwise Separable Convolution structure, significantly reducing the number of model parameters. Specifically, the computational cost of a standard convolution is:

$$Cost_{standard} = D_K \cdot D_K \cdot M \cdot \mathrm{N} \cdot D_F \cdot D_F \quad (1)$$

Here, D_K represents the size of the convolution kernel, M is the number of input channels, N is the number of output channels, and D_F is the size of the feature map. The computational cost of depth-separable convolution is:

$$Cost_{depthwise} = D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \quad (2)$$

This can reduce the number of parameters to approximately 1/9 of the original model and is suitable for deployment on edge computing platforms such as Jetson Nano and Raspberry PI. Meanwhile, to make up for the deficiency of the small model's ability to abstract high-order features, the edge model introduces the SE attention mechanism. By adaptively adjusting the weights of the feature channels, it enhances the discriminative features, enabling the model to achieve high recall rate recognition even when the input image resolution is limited and the defect size is small.

For the large model part, the improved Swin Transformer is selected. Its core design idea is the window-based Multi-head self-attention computing mechanism (W-MSA) based on local Window partitioning, which significantly reduces the global computational complexity. The original computational complexity of the self-attention mechanism of the Transformer architecture is:

$$\mathrm{O}(\mathrm{n}^2 \cdot d) \quad (3)$$

Here, n represents the number of input tokens and d is the feature dimension. Swin Transformer adopts the sliding window partitioning strategy, restricting the attention calculation within the window range, and the complexity is reduced to:

$$\mathrm{O}(M \cdot \mathrm{W}^2 \cdot d) \quad (4)$$

Here, M represents the number of Windows and W represents the size of the Windows. This strategy enhances scalability and training efficiency while maintaining the receptive field of the model. The large model further achieves cross-regional feature interaction through multi-scale Patch merging and cross-window attention module (Shifted Window), effectively identifying tiny clamping defects in complex backgrounds, such as loose screws and metal erosion.

3.2 Training Strategies and Optimization Mechanisms

For the above-mentioned models of different sizes, this paper adopts different training strategies to adapt to their deployment goals. The small model adopts the end-to-end supervised learning approach, and the Loss function selects the combined form of Cross Entropy Loss and Focal Loss to enhance the sensitivity to the imbalanced data between classes:

$$\mathcal{L}_{\text{focal}} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (5)$$

Here, p_t represents the predicted probability of the model for the true category, γ controls the weights of the difficult and easy samples, and α_t is the balance factor. This loss function is particularly suitable for the common "few defects vs a large number of normal" sample distribution at the edge. During the training process, data augmentation strategies (such as rotation, mirroring, and luminance perturbation) are adopted to enhance the model's robustness against on-site image interference. To compress the model volume, deep quantization (8-bit) and pruning techniques are adopted to keep the model's running time on embedded hardware within 50ms.

In the training phase of large models, contrastive learning mechanisms and multi-label supervision frameworks are introduced. The labeled multi-dimensional defect labels (such as "corrosion + offset", "rust + loosening", etc.) are utilized to enhance the model's discriminative ability. Its loss function is in the combined form of multi-label Binary Cross Entropy with Logits and cosine similarity loss:

$$\mathcal{L} = \sum_{i=1}^{C}\left[y_i\log(p_i) + (1-y_i)\log(1-p_i)\right] + \lambda(1-\cos(z_1, z_2)) \quad (6)$$

The former term is the standard multi-label loss, and the latter term is the cosine similarity loss of the sample embedding vectors z_1, z_2, which is used to enhance the separability of the feature space. The training adopted ImageNet pre-training weight initialization, combined with transfer learning to reduce convergence time. Finally, fine-tuning was performed on the transformer defect image dataset (a total of 4800 images), and the final recognition accuracy of the Top-1 model reached 97.8%.

By constructing an identification framework of "rapid detection of small side models + in-depth optimization of large cloud models", and in combination with carefully designed structures, loss functions and training strategies, this paper has achieved the deployment of a high-precision, high-efficiency and low-latency main transformer core clamping defect identification system, providing strong support for the intelligent condition-based maintenance of transformers.

4. Experiments and Results

To verify the effectiveness and engineering applicability of the proposed defect identification method for main transformer core clamps based on the collaborative strategy of large and small models, this paper designs a systematic experimental process covering multiple links such as model construction, performance testing, deployment evaluation, and on-site application feedback. The entire experiment is divided into two stages: the first stage focuses on the evaluation of model accuracy and efficiency, while the second stage pays attention to the application effect and reliability verification of the model in a real substation environment.

4.1 Model Performance Testing and Comparative Analysis

The experimental data mainly come from the on-site inspection images provided by several substations of State Grid Corporation of China, totaling 9,800 images, covering multiple working conditions and typical defect types of the core clamping of the main transformer, including six major types of faults: loosening, rusting, detachment, fracture, displacement and surface contamination. The data was manually labeled and divided into a training set (7,840 pieces), a validation set (980 pieces), and a test set (980 pieces). During the model training process, the small model selects the lightweight convolutional neural network MobileNetV3 and is deployed on the edge terminal (RK3568

embedded AI device). The large model adopts the Swin Transformer with balanced structure and performance, which runs on the NVIDIA A100 cloud server. The collaborative model is based on the integration of the two and adopts a two-stage identification process of "edge screening and cloud diagnosis".

Table 1. Experimental Performance Comparison Table

Model	Accuracy rate (%)	Recall rate (%)	Inference delay (ms)	Model size (MB)
Small model (edge)	87.4	84.1	15	4.2
Large model (Cloud)	94.2	93	85	120
Collaboration between large and small models	96.5	95.2	40	5.3

Analysis of the Table 1 shows that the small model has a good real-time response capability at the edge, with an inference delay of only 15ms. However, due to the limitations of the number of parameters and the depth of modeling, there are certain misjudgments and missed detections in the recognition of complex defects. By introducing the local window attention mechanism (Swin Transformer), the large model significantly improves the recognition accuracy while maintaining spatial efficiency, achieving an accuracy rate of 94.2% and a recall rate of 93.0%. However, due to the high requirements for computing resources, the inference delay is as high as 85ms, making it unsuitable for direct deployment in resource-constrained scenarios.

The collaborative model integrates the advantages of both through a mechanism of "edge rapid screening + cloud deep recognition": the edge-end small model periodically processes the collected images, quickly discovers suspicious areas and labels them with lightweight tags; The large model only performs depth discrimination on the labeled images, reducing the computational load while ensuring high recognition accuracy. The final collaborative strategy achieved an accuracy rate of 96.5% and a recall rate of 95.2%, which was much higher than that of any single model. Meanwhile, the model size is controlled at 5.3MB, reducing the cloud computing burden by approximately 60%, providing feasible support for engineering deployment.

4.2 On-site Deployment and Application Effectiveness

In order to verify the stability and practicability of the collaborative model in a real environment, this paper conducted a three-month pilot deployment experiment at a 110kV substation in Zhejiang. The deployment of the system consists of two parts: First, edge AI cameras with computing capabilities are set up along the equipment inspection channels to capture images of core clamps at fixed intervals every day and initially screen for defects. The second is to upload the suspicious images to the cloud platform, where the large model analyzes the fault type, generates a diagnostic report and provides handling suggestions. The entire process is based on a cloud-edge collaborative architecture, integrating lightweight reasoning with in-depth understanding, significantly enhancing the accuracy and response speed of power equipment condition-based maintenance.

During this deployment process, a total of over 27,000 image samples were processed, and 912 suspected defect images were identified. After cloud re-inspection, 248 valid defects were confirmed. The overall false alarm rate of the collaborative model was controlled within 2.3%, and it played a timely warning role in three key equipment operation risk warning events, successfully avoiding major hidden dangers such as loose clamping parts and broken conductive screws, achieving remarkable engineering results. Meanwhile, the average response time of the system is controlled within 60 seconds, which is far superior to the traditional manual inspection feedback cycle and significantly improves the operation and maintenance efficiency.

Feedback from on-site operation and maintenance personnel shows that this collaborative identification system is easy to deploy, highly scalable, supports integration with existing substation SCADA systems and data platforms, and can achieve unified dispatching and remote operation and maintenance of multiple sites. It has good promotion prospects and application value.

5. Conclusion

This paper focuses on the defect identification of the core clamping pieces of the main transformer and proposes an intelligent diagnosis method based on the collaboration of large and small models. This method achieves high accuracy and timeliness in defect recognition by constructing

a collaborative recognition framework of lightweight small models and large models with strong representational capabilities, fully integrating the advantages of real-time response at the edge and in-depth analysis in the cloud. During the model construction process, the integration of multi-task loss functions and feature alignment mechanisms has enhanced the front-end screening capability and overall diagnostic reliability of small models. The experimental results show that the collaborative model outperforms the single model in terms of accuracy, recall rate and reasoning efficiency. The accuracy rate reaches 96.5%, and the reasoning delay remains within the deployable range, demonstrating engineering practicality. In the real substation deployment test, the system successfully identified multiple typical defects, effectively supporting the equipment condition-based maintenance tasks and verifying the practical value and scalability of this strategy.

Subsequent research will further explore model distillation, adaptive fusion mechanisms, and multimodal information introduction strategies to enhance the system's generalization ability in multiple scenarios and tasks, providing stronger technical support for the intelligent operation and maintenance of power equipment.

References

[1] Dang, L. M., Wang, H., Li, Y., Nguyen, T. N., & Moon, H. (2022). DefectTR: End-to-end defect detection for sewage networks using a transformer. Construction and Building Materials, 325, 126584.

[2] Gao, L., Zhang, J., Yang, C., & Zhou, Y. (2022). Cas-VSwin transformer: A variant Swin transformer for surface-defect detection. Computers in Industry, 140, 103689.

[3] Zhang, Y., Tang, Y., Liu, Y., & Liang, Z. (2022) Fault diagnosis of transformer using artificial intelligence: A review. Front. Energy Res. 10:1006474

[4] Wu, H. Y., Matthew, J. T., & John W. S. (2023) A transformer-based approach for novel fault detection and fault classification/diagnosis in manufacturing: A rotary system application. Journal of Manufacturing Systems, 67, 439-452

[5] Lei, Z., Zhang, Y., Wang, J., & Zhou, M. (2024). Cloud-Edge Collaborative Defect Detection Based on Efficient Yolo Networks and Incremental Learning. Sensors, 24(18), 5921.

[6] Li, Y., Xiang, Y., Guo, H., Liu, P., & Liu, C. (2022). Swin Transformer Combined with Convolution Neural Network for Surface Defect Detection. Machines, 10(11), 1083.

[7] Gao, P., Wu, T., & Song, C. (2024). Cloud–Edge Collaborative Strategy for Insulator Recognition and Defect Detection Model Using Drone-Captured Images. Drones, 8(12), 779.

[8] Wang, J., Xu, G., Yan, F., Wang, J., & Wang, Z. (2022). Defect Transformer: An Efficient Hybrid Transformer Architecture for Surface Defect Detection. arXiv preprint arXiv:2207.08319.

[9] Hu, C., Yao, J., Wu, W., Qiu, W., & Zhu, L. (2022). A Lightweight Reconstruction Network for Surface Defect Inspection. arXiv preprint arXiv:2212.12878.

[10] Xu, Y., Khan, T.M., Song, Y. et al. (2025) Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. Artif Intell Rev 58, 93.

