Transient Stability Assessment of Power Systems Based on Transformer and Graph Attention Network

Shengyuan Wang, Fengzhang Luo*, Zhengyang Xu
State Key Laboratory of Smart Power Distribution Equipment and System
Tianjin University
Tianjin China
luofengzhang@tju.edu.cn

Lukun Ge, Chao Pang
State Grid Tianjin Electric Power Company
Tianjin, China

***Abstract*—Transient stability assessment (TSA) is critical for ensuring the secure operation of power systems. To address the limitations of traditional methods in feature extraction and spatial correlation modeling, this paper proposes a TSA model based on Transformer and Graph Attention Network (GAT). The model takes pre-fault, during-fault, and post-fault bus voltage magnitudes and phase angles as input. The Transformer module is employed to adaptively extract temporal features and capture long-range dependencies in transient data. Then, the GAT module is used to model the power grid topology, enabling adaptive aggregation of electrical correlation strengths among nodes and fully exploiting the spatial structure of the system. During training, the hybrid loss function composed of Binary Cross-Entropy Loss and Dice Loss is introduced to address class imbalance and enhance the model's ability to distinguish boundary samples. Simulation results on the IEEE 39-bus and IEEE 145-bus systems demonstrate that the proposed method achieves higher accuracy than existing mainstream models and maintains robust performance under topological changes and noise disturbances.**

***Index Terms*--Transient stability assessment; transformer; graph attention network; hybrid loss function.**

I. Introduction

Ensuring secure and stable power system operation is a core objective of grid dispatching. Transient stability, which reflects the system's ability to recover after major disturbances, is crucial for grid security [1]. With rising renewable integration and increasing system complexity, traditional transient stability assessment (TSA) methods face challenges in balancing accuracy and real-time performance[2]. Thus, developing TSA models that are both efficient and generalizable has become a key research focus.

Traditional TSA methods largely rely on manual feature extraction and shallow machine learning models, such as decision trees [3], support vector machines (SVM) [4], and random forests [5]. Although these methods offer relatively fast computation and acceptable prediction accuracy in specific scenarios, their limited model expressiveness makes it difficult to capture the nonlinear dynamic features of power system transients effectively. In recent years, advances in deep learning have provided new approaches for stability assessment. Reference [6] proposed an efficient TSA method based on temporal convolutional networks (TCN), where a grey wolf optimizer was used to enhance model accuracy. Reference [7] developed a TSA model named long short-term memory with spatial attention fusion, which incorporates a self-attention mechanism and focal loss function to achieve fast and reliable TSA. These data-driven methods learn temporal dependencies from time-series data, enabling stability identification. However, their neglect of grid topology and electrical coupling limits performance under topology changes or operational fluctuations, reducing generalization in complex systems.

To account for the spatial structural characteristics of power systems, graph neural networks (GNNs) [8] have been widely introduced into TSA research. These methods treat the power grid as graph-structured data, with buses represented as nodes and transmission lines as edges, enabling direct modeling of electrical coupling between nodes in the graph space. Studies such as [9], [10] apply graph convolutional networks (GCN) to incorporate the influence of neighboring nodes during feature aggregation, thereby achieving joint modeling of spatial and temporal characteristics of the system. This approach improves generalization under varying topologies by enabling models to adapt to dynamic responses. However, most GNN-based methods use fixed adjacency matrices, limiting their ability to capture nonlinear coupling and reducing assessment accuracy.

Transient stability datasets in power systems typically exhibit significant class imbalance [11], [12]. In large-scale simulations or real-world operational data, stable samples greatly outnumber unstable ones. This imbalance introduces bias during model training, causing the model to favor predictions of the stable class and reducing its accuracy in identifying critical unstable samples. To address this issue, previous studies have introduced generative adversarial networks (GAN) [13], [14] to generate minority class samples and balance the dataset. Other researchers have employed loss functions such as cross-entropy [15] or focal loss [16] by assigning higher weights to unstable samples or applying stronger gradient constraints to hard-to-classify instances, thereby enhancing the model's ability to learn from minority class samples. However, these methods struggle with samples near the stability boundary. Generated data may lack authenticity, and single loss functions often fail to balance overall accuracy with boundary discrimination, limiting performance under complex disturbances.

To this end, this paper proposes a transient stability assessment method for power systems based on the Transformer and graph attention network (GAT). The

979-8-3315-8138-1

proposed approach introduces improvements in both feature extraction and model optimization, aiming to achieve high accuracy and robust assessment of transient processes in power systems. The Transformer is employed to capture long-range temporal dependencies in voltage and phase angle sequences, while the GAT is used to adaptively model the electrical coupling strength between nodes. Together, they enable deep integration of spatiotemporal features. In addition, a hybrid loss function combining binary cross-entropy (BCE) loss and dice loss is introduced to enhance the model's performance under scenarios with imbalanced samples and near-boundary instance recognition. The main contributions of this paper are as follows:

1)The Transformer-GAT based spatiotemporal TSA framework is developed for power systems, where the Transformer captures global temporal dependencies from post-fault voltage and angle trajectories, and the GAT adaptively models electrical couplings over the grid topology, enabling joint learning of temporal dynamics and spatial interactions rather than treating them independently.

2)The attention-driven topology-aware aggregation mechanism is introduced, in which node influences are learned through adaptive attention weights instead of fixed adjacency or convolutional aggregation, allowing the model to flexibly characterize varying electrical coupling strengths and improving generalization under topology changes.

3)The hybrid BCE-Dice loss is designed for imbalanced TSA datasets, enhancing boundary-sample discrimination and improving stability identification performance compared with conventional single-loss training.

4)Comprehensive robustness and scalability validations are conducted, including large-scale systems, topology variations (N-1/N-2), and noisy measurements, together with millisecond-level inference latency, demonstrating practical applicability for online transient stability assessment in control centers.

II. Model Principle and Overall Framework

A. *Temporal Feature Extraction Based on Transformer*

During a transient event, bus voltages and phase angles show strong temporal dependencies and nonlinear dynamics. The Transformer captures global temporal correlations via self-attention. For an input feature sequence:

$$X = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T\times d_x} \tag{1}$$

where, T denotes the number of time steps, and d_x represents the feature dimension at each time step.

The Transformer model first applies linear projections to obtain the query matrix $Q\in \mathrm{R}^{T\times dk}$, key matrix $K\in \mathrm{R}^{T\times dk}$, and value matrix $V\in \mathrm{R}^{T\times dk}$, d_k denotes the feature dimension of the key vectors in the attention mechanism.

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{2}$$

where, W_Q, W_K, W_V are trainable weight matrices.

The self-attention layer computes the attention weight matrix by evaluating the correlations between time steps:

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

where, the scaling factor $\sqrt{d_k}$ is introduced to prevent gradient explosion. This mechanism enables the model to capture dependencies across all time steps in a single forward pass, achieving efficient global temporal modeling.

To further enhance the model's ability to represent multi-scale temporal features, the Transformer incorporates a Multi-Head Attention mechanism, which performs self-attention operations in parallel across h subspaces:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad i = 1,2,\ldots,h, \tag{4}$$

where, W_i^Q , W_i^K , W_i^V are the trainable weight matrices of the i-th attention head.

The outputs of all heads are concatenated and then passed through a linear transformation to obtain the final feature representation:

$$MHA(Q,K,V) = Concat\left(head_1, head_2, \ldots, head_h\right)W^O \tag{5}$$

where, $W^O \in \mathbb{R}^{hd_v \times d_x}$ is the output projection matrix used to integrate the multi-head attention results into a unified temporal feature representation, $d_v$represents the feature dimension of the value vectors.

To retain temporal order without recurrence, the Transformer uses a positional encoding function PE(t) based on sine and cosine mappings, allowing the model to capture temporal indices and long-range dependencies.

$$H_T = f_{\text{Trans}}(X) \tag{6}$$

where, $H_T \in \mathbb{R}^{T\times d_x}$ serves as the input for subsequent spatial feature modeling (GAT module), providing global temporal dependency representations for learning the topological characteristics of the power grid.

B. *Spatial Feature Extraction Based on GAT*

The topology of a power system consists of buses and the transmission lines between them, which can essentially be represented as a weighted undirected graph G=(V, E), where V denotes the set of nodes with $|V|=N$ being the number of nodes, and E represents the set of edges (i.e., transmission lines). The electrical coupling relationship between any two nodes i and j in the system can be described by the adjacency matrix $A\in R^{N\times N}$, where a_{ij}=1 if a direct connection exists between nodes i and j, and a_{ij}=0 otherwise. GAT [8] overcomes this limitation by introducing an attention mechanism that enables adaptive feature aggregation between nodes. For the input node feature matrix:

$$H = [h_1, h_2, \ldots, h_N]^{\mathrm{T}} \in \mathbb{R}^{N\times d_h} \tag{7}$$

where, $h_i \in \mathbb{R}^{d_h}$ denotes the feature vector of node i, and d_h is the feature dimension, d_h denotes the overall feature dimension of the hidden layer, and h represents the number of attention heads in the multi-head attention mechanism.

Linear transformation is applied to obtain the node's feature representation.

$$z_i = Wh_i \tag{8}$$

where, $W \in \mathbb{R}^{d_z \times d_h}$ is a trainable weight matrix.

Single-layer feedforward network is used to compute the attention coefficient between node i and its neighboring nodes $j \in N_i$:

$$e_{ij} = \text{LeakyReLU}\left(a^{\mathrm{T}}\left[z_i \| z_j\right]\right) \tag{9}$$

where, a is learnable parameter vector, $a \in \mathbb{R}^{2d_z}$ denotes the vector concatenation operation, and LeakyReLU$(\cdot)$ is the leaky rectified linear unit with a negative slope coefficient.

The attention coefficients of all neighboring nodes of node i are then normalized as follows:

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(e_{ik}\right)} \tag{10}$$

where $\sigma(\cdot)$ is a nonlinear activation function. This mechanism enables the model to adaptively weight neighboring nodes based on both topology and feature correlations, allowing differentiated modeling of electrical coupling strength.

To enhance feature expressiveness, GAT typically adopts a Multi-Head Attention mechanism, where K independent attention heads are computed in parallel and their outputs are fused as:

$$h_i' = \|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(k)} W^{(k)} h_j\right) \tag{11}$$

C. Spatiotemporal Feature Fusion Framework Based on Transformer-GAT

The overall structure of the model is illustrated in Fig. 1.

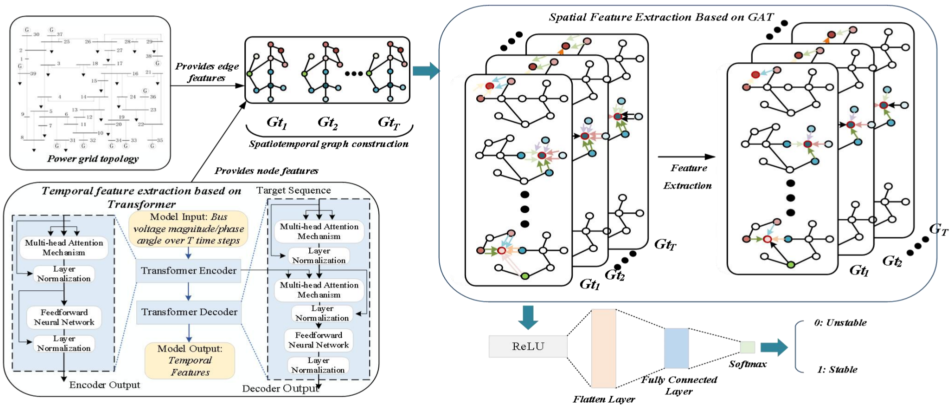

Fig. 1 Overall Structure of the Proposed Model

The input layer receives bus voltage magnitude and phase angle sequence data after normalization. Let N be the number of system nodes and T the sampling window length, then the input tensor can be represented as:

$$X \in \mathbb{R}^{N \times T \times d} \tag{12}$$

The Transformer module performs feature extraction along the temporal dimension. By leveraging the multi-head self-attention mechanism, it captures global dependencies across pre-fault, fault, and post-fault stages, generating a comprehensive temporal representation $H_T \in \mathbb{R}^{N \times d_T}$ for each node.

The GAT module takes H_T as the input node features and aggregates spatial features based on the system's topological adjacency. Through the adaptive learning of attention coefficients α_{ij}, the model distinguishes the electrical coupling strength between different nodes and generates spatial feature representations. In the feature fusion stage, residual connections and layer normalization are employed to concatenate temporal and spatial features, followed by a nonlinear transformation:

$$H_F = \text{LayerNorm}\left(H_T + H_S\right) \tag{13}$$

where, H_F is the fused spatiotemporal feature representation used for subsequent stability classification. The residual structure helps alleviate the vanishing gradient problem and preserves feature continuity, while layer normalization stabilizes model training and improves convergence speed.

In the classification stage, the fused feature H_F is passed through a fully connected layer and mapped to a binary classification space via the Softmax function to determine the system's post-disturbance stability state. The output vector is given by:

$$y = Softmax\left(W_o H_F + b_o\right) \tag{14}$$

where W_o and b_o are the weight and bias parameters of the output layer, and y=[1, 0] represents the predicted probabilities of the system being stable or unstable.

D. Hybrid Loss Function Design

To address both class imbalance and the challenge of boundary sample recognition, this paper designs a hybrid loss function composed of BCE loss and dice loss, aiming to balance global classification performance and local boundary discrimination ability.

1) Binary Cross-Entropy Loss

The BCE loss measures the difference between the predicted probability distribution and the true labels. It is defined as:

$$L_{BCE}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right] \quad (15)$$

where N is the total number of samples, $y_i \in \{0,1\}$ is the true label of the i-th sample, and $\hat{y}_i \in [0,1]$ is the predicted probability of being in the stable state. BCE loss effectively constrains overall classification accuracy.

2) Dice Loss

To improve the model's discriminative capability in boundary and minority class regions, Dice loss is introduced as it directly optimizes the overlap between predictions and true labels. By focusing on the matching area, it enhances sensitivity to misclassifications in unstable regions and mitigates the limitations of conventional loss functions under class imbalance.

$$L_{\text{Dice}}=1-\frac{2\sum_{i=1}^{N}y_i\hat{y}_i+\epsilon}{\sum_{i=1}^{N}(y_i+\hat{y}_i)+\epsilon} \quad (16)$$

where ε is a small constant to prevent division by zero.

3) Hybrid Loss Construction

$$L_{hyb}=\lambda L_{BCE}+(1-\lambda)L_{\text{Dice}} \quad (17)$$

where $\lambda \in [0,1]$ is a weighting coefficient used to balance global classification accuracy and boundary recognition.

E. Performance Evaluation Metrics

The transient stability assessment is formulated as a binary classification task, with accuracy (ACC) used as the evaluation metric. Let T_s and T_u represent the number of correctly predicted stable and unstable samples, respectively, and F_s and F_u represent the number of misclassified stable and unstable samples. The evaluation metric is defined as follows:

$$\text{Accuracy}=\frac{T_s+T_u}{T_s+T_u+F_s+F_u} \quad (18)$$

III. Simulation and Result Analysis

A. Simulation System and Dataset Construction

The IEEE 39-bus system and the IEEE 145-bus system are selected as case studies. Time-domain simulations are conducted on the PSD-BPA platform. The systems operate under rated conditions, and multiple disturbance scenarios are introduced, including varying load levels (75%-120% of the base load), fault locations (10%-90% along the transmission line), and fault durations (0.10 s, 0.14 s, 0.18 s, and 0.22 s). Power flow convergence is ensured by adjusting generator outputs, resulting in a diverse set of transient operation samples. A total of 15,000 samples are generated for each system, with a stable-to-unstable sample ratio of approximately 3:1.

B. Model Training Configuration

The model is implemented using Python 3.8 and the PyTorch framework, and all experiments are conducted on an NVIDIA RTX 4060Ti GPU. The Transformer module adopts a 3-layer encoder structure, with each layer containing 4 attention heads and a hidden dimension of d_h=128. The dropout rate is set to 0.3. The GAT module consists of 3 layers, each using 8-head attention with average aggregation. The weighting coefficient for the hybrid loss function is set to $\lambda = 0.6$. The training batch size is 128, the initial learning rate is 0.001, and the Adam optimizer is used. The model is trained for 120 epochs. The input of the proposed model consists of 60 ms of post-fault voltage and phase angle time-series data. The model's scalability is validated on the IEEE 145-bus system. It adopts an offline training and online inference framework, achieving inference times of approximately 30 ms (IEEE 39-bus) and 38 ms (IEEE 145-bus), making it suitable for real-time online assessment and early warning in control centers, while also supporting offline event screening and simulation-based analysis.

D. Analysis of Different Loss Functions

To validate the effectiveness of the hybrid loss design, comparative experiments are conducted between the proposed hybrid loss and focal loss. The test results under each loss function are presented in Table II.

TABLE II TEST RESULTS OF DIFFERENT LOSS FUNCTIONS

Model	Focal Loss	**Hybrid Loss (Proposed)**
ACC/%	98.63	**99.62**

The proposed hybrid loss achieves a higher ACC compared to focal loss, confirming its ability to improve overall model performance under class imbalance conditions. Furthermore, to further validate the effectiveness of the proposed loss function, Fig. 2 presents the accuracy convergence curves of three different loss functions, recorded every 10 training epochs. The results clearly demonstrate the superiority of the Hybrid loss.

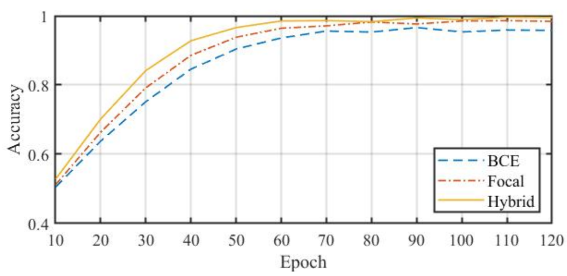

Fig. 2 Convergence comparison of different loss functions

E. Robustness Analysis Under Noise Disturbance and Topology Variations

1) Noise Robustness Test: Gaussian noise is added to the input voltage and phase angle measurements with error levels set to 0.2%, 0.4%, 0.6%, and 0.8%. The proposed model is compared against the well-performing GCN and Transformer-GCN models as reported in Table III. The results show that while the ACC of all models decreases with increasing noise levels, the proposed model consistently maintains the highest performance.

TABLE III PERFORMANCE COMPARISON UNDER DIFFERENT NOISE LEVELS

Method	0%	0.2%	0.4%	0.6%	0.8%
GCN	97.23	97.02	96.69	96.32	95.89
Transformer-GCN	98.52	98.36	97.86	97.23	96.89
Proposed	**99.62**	**99.31**	**98.93**	**98.65**	**98.26**

2) Topology Variation Test: In the IEEE 39-bus system, 1-2 lines are randomly disconnected to generate N−1 and N-2 topologies, resulting in 40 new topology scenarios. Experimental results show that the proposed method achieves an accuracy of 99.06% under these altered topologies, with performance degradation of less than 1%, significantly outperforming other methods. This demonstrates that the proposed framework possesses strong generalization and adaptability under topological changes.

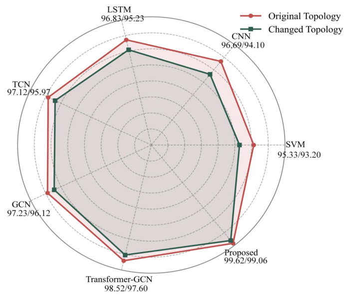

Fig. 3 Comparison of Different Methods Under Topology Variation

F. *Performance Analysis on a Larger-Scale Network*

To evaluate the adaptability of the proposed model in large-scale systems, simulation tests are conducted on the IEEE 145-bus system. The system includes 50 generators, 145 buses, and 453 branches, with a total of 8,000 samples generated (stable to unstable ratio of 3:1). All models are trained and tested under the same parameter settings, and the results are shown in Table IV.

TABLE IV COMPARISON OF DIFFERENT METHODS ON THE IEEE 145-BUS SYSTEM

Model	SVM	CNN	LSTM	TCN	GCN	Transformer-GCN	**Proposed**
ACC/%	93.16	93.32	94.33	95.86	97.08	97.45	**98.33**

The results indicate that the proposed Transformer-GAT model still achieves the highest accuracy and stability in the large-scale system, with an ACC of 98.33%, outperforming all other methods. This verifies the model's scalability and robustness.

IV. CONCLUSION

The main conclusions of this paper are as follows:

1) The spatiotemporal fusion framework based on Transformer-GAT is proposed: the framework leverages Transformer to capture global temporal dependencies in bus voltage and phase angle sequences, while GAT adaptively models the power network topology. This effectively represents the spatiotemporal coupling in system transients and significantly enhances feature representation capability.

2) The hybrid loss function is designed to improve learning under class imbalance: By combining BCE loss with similarity coefficient loss, the model's ability to distinguish boundary and minority class samples is enhanced.

3) Simulation results verify the superiority of the model: On the IEEE 39-bus and 145-bus systems, the proposed method outperforms existing approaches in terms of accuracy and other key metrics, demonstrating higher evaluation precision and robustness.

4) The model shows strong scalability and engineering applicability: under noise interference, topology changes, and large-scale systems, the proposed method consistently outperforms state-of-the-art approaches and maintains stable performance, confirming its generalization ability and practical feasibility in complex power system scenarios.

In future work, we will further consider N-1-1/N-k fault scenarios and focus on improving robustness under missing or corrupted input conditions.

ACKNOWLEDGMENT

This work is supported by Smart Grid-National Science and Technology Major Project (No. 2024ZD0800800) and science and technology project of SGCC (State Grid Corporation of China): Low-carbon and Reliable Urban Power Distribution System Demonstration Project (SGTJDK00DWJS2400298).

REFERENCES

[1]T. Zhao, J. Wang, X. Lu, *et al.*, "Neural Lyapunov Control for Power System Transient Stability: A Deep Learning-Based Approach," *IEEE Trans. Power Syst.*, vol. 37, no. 2, pp. 955–966, Mar. 2022.

[2]S. Li, C. Ye, Y. Ding, *et al.*, "Reliability Assessment of Renewable Power Systems Considering Thermally-Induced Incidents of Large-Scale Battery Energy Storage," *IEEE Trans. Power Syst.*, pp. 1–15, 2022.

[3]C. Ren, Y. Xu, and R. Zhang, "An Interpretable Deep Learning Method for Power System Transient Stability Assessment via Tree Regularization," *IEEE Trans. Power Syst.*, vol. 37, no. 5, pp. 3359–3369, Sept. 2022.

[4]H. Wang, L. Hu, and Y. Zhang, "SVM based imbalanced correction method for Power Systems Transient stability evaluation," *ISA Trans.*, vol. 136, pp. 245–253, May 2023.

[5]C. Zhang, Y. Li, Z. Yu, *et al.*, "Feature selection of power system transient stability assessment based on random forest and recursive feature elimination," in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Xi'an, China: IEEE, Oct. 2016, pp. 1264–1268.

[6]M. Massaoudi, T. Zamzam, M. E. Eddin, *et al.*, "Fast Transient Stability Assessment of Power Systems Using Optimized Temporal Convolutional Networks," *IEEE Open J. Ind. Appl.*, vol. 5, pp. 267–282, 2024.

[7]Z. Shao, Q. Wang, Y. Cao, *et al.*, "A Novel Data-Driven LSTM-SAF Model for Power Systems Transient Stability Assessment," *IEEE Trans. Ind. Inform.*, vol. 20, no. 7, pp. 9083–9097, July 2024.

[8]F. Luo, S. Wang, Y. Lv, *et al.*, "Domain knowledge-enhanced graph reinforcement learning method for Volt/Var control in distribution networks," *Appl. Energy*, vol. 398, p. 126409, Nov. 2025.

[9]S. Xia, C. Zhang, Y. Li, *et al.*, "GCN-LSTM Based Transient Angle Stability Assessment Method for Future Power Systems Considering Spatial-Temporal Disturbance Response Characteristics," *Prot. Control Mod. Power Syst.*, vol. 9, no. 6, pp. 108–121, Nov. 2024.

[10]X. Zhou, X. Guan, D. Sun, *et al.*, "Transient Stability Assessment Based on Gated Graph Neural Network With Imbalanced Data in Internet of Energy," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9320–9331, June 2022.

[11]X. Zhan, S. Han, N. Rong, *et al.*, "A hybrid transfer learning method for transient stability prediction considering sample imbalance," *Appl. Energy*, vol. 333, p. 120573, Mar. 2023.

[12]"High-Quality Sample Generation for Power System Transient Stability Assessment Based on Data-Driven Methods," *CSEE J. Power Energy Syst.*, 2025, Accessed: Nov. 10, 2025.

[13]J. Fang, L. Zheng, C. Liu, *et al.*, "A Data-Driven Case Generation Model for Transient Stability Assessment Using Generative Adversarial Networks," *IEEE Trans. Ind. Inform.*, vol. 20, no. 12, pp. 14391–14400, Dec. 2024.

[14]B. Li, P. Fan, Q. Chen, *et al.*, "High-Quality Sample Generation for Power System Transient Stability Assessment Based on Data-Driven Methods," *CSEE J. Power Energy Syst.*, 2025.

[15]H. Zhao and R. Ni, "Power System Transient Stability Assessment Based on Spatio-Temporal Broad Learning System," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 10343–10353, 2025.

[16]M. Massaoudi, H. Abu-Rub, and A. Ghrayeb, "FLACON: A Deep Federated Transfer Learning-Enabled Transient Stability Assessment During Symmetrical and Asymmetrical Grid Faults," *IEEE Open J. Ind. Appl.*, vol. 5, pp. 253–266, 2024

979-8-3315-8138-1