Chen and Guan
Protection and Control of Modern Power Systems (2023) 8:5
https://doi.org/10.1186/s41601-023-00278-x

Protection and Control of Modern Power Systems

ORIGINAL RESEARCH

Open Access

Static information, K-neighbor, and self-attention aggregated scheme: a transient stability prediction model with enhanced interpretability

Liukai Chen and Lin Guan*

Abstract

Data-driven preventive scanning for transient stability assessment (DTSA) is a faster and more efficient solution than time-domain simulation (TDS). However, most current methods cannot balance generalization to different topologies and interpretability, with simple output. A model that conforms to the physical mechanism and richer label for transient stability can increase confidence in DTSA. Thus a static-information, k-neighbor, and self-attention aggregated schema (SKETCH) is proposed in this paper. Taking only static measurements as input, SKETCH gives several explanations that are consistent with the physical mechanisms of TSA and provides results for all generator stability while predicting system stability. A module based on the self-attention mechanism is designed to solve the locality problem of a graph neural network (GNN), achieving subgraph equivalence outside the k-order neighborhood. Test results on the IEEE 39-bus system and IEEE 300-bus system indicate the superiority of SKETCH and also demonstrate the rich sample interpretation results.

Highlights

- A fast TSA scheme for pre-failure scanning.
- A physical mechanism-based attention structure for dynamic graph pooling.
- A node regression model that responds to key physical mechanisms.
- Generator label for richer output information.
- Top performance and post-hoc interpretation.

Keywords Transient stability assessment (TSA), Data-driven, Explainable, Graph neural network (GNN), Self-attention

1 Introduction

1.1 Background

In recent years, transient stability assessment (TSA) models that rely on time-domain simulation (TDS) and expert experience have been challenged by the increased penetration of renewable energy sources and the increased flexibility required for power system operation. Although TDS can provide the most detailed dynamic

*Correspondence:
Lin Guan
lguan@scut.edu.cn
School of Electric Power, South China University of Technology, Guangzhou, China

profile of a given transient, it is time-consuming and cannot provide an analytic mapping from the input to the stability results. In other words, the mechanism of transient stability remains a black-box for engineers.

In this context, data-driven TSA (DTSA) has received a lot of attention because of its fast-scanning speed and ability to learn generic knowledge from TDS results. With both scientific and engineering requirements, it is a new proposition how to obtain more human-understandable supporting information from the black-box and add interpretability to the models.

1.2 Literature review and motivation

DTSA methods have made rapid progress in recent years, and the main target is to improve the interpretability and generalization performance of the methods.

Fully explainable models are the first to receive attention because they can provide a complete analysis process. Methods based on the assumption of feature independence, such as the general linear model (GLM) [1], provide explanations of the coefficients of feature effects on the stability results, but their assumptions do not explain power system transients with strongly coupled features. Rule-based methods, such as decision trees (DT) [2–4], and extreme gradient boosting (XGBoost) [5], model the stability margin as a tree-like structure with segmented discriminations of thresholds, which can clearly show the model decision basis and boundary. A causal theory-based feature selection approach [6] combined with DT can give more robust results. However, it is difficult for these models to improve performance because of strong model assumptions and failure to generalize to topologies. This means that different models are required for different topologies. On the other hand, in the process of real system operation, the dimensionality caused by the combination of multiple maintenance cases will lead to large numbers of different system models.

Models with strong generalization capabilities, represented by deep learning (DL), have subsequently attracted a lot of attention. Such models are generalized well enough that a single model can be trained and used for tasks with different modes of operation, different fault locations, and different topologies. Models with system dynamic information as input are developed first, represented by approaches such as Gated Recurrent Unit (GRU) [7], Long-Short-Term Memory (LSTM) [8], or reordering trajectories to form pictures [9], and temporal feature transformation using a Convolution Neural Network (CNN) [10], and have achieved success. The dynamic information can express the real-time results of the system components after they interact with the network, thus improving the generalizability of the model to the topology. However, the possibility of obtaining the mapping between the static operation mode with the steady-state results is hindered by the fact that the dynamic features come from the black-box that is the time-domain simulation. At the same time, the interpretability of the process of the model is greatly weakened because of the excessive number of parameters and overly complex mapping, thus limiting the trust of the operators.

With this in mind, researchers have developed two types of schemes to improve the interpretability of DTSA. The first is the post hoc interpretation methods that can provide the sensitivity performance of the model in the vicinity of the sample, such as differencing [11] or Local Interpretable Model-agnostic Explanations (LIME) [12]. However, they cannot provide the interpretation within the model. The second is to design models that better reflect the physical mechanisms of transient stability. Models designed in this way provide a portion of the internal structure with interpretability and often with enhanced performance. The Graph Neural Network (GNN) is by far the most promising solution. Considering the non-Euclidean properties of the power system, where buses refer to nodes and lines refer to edges, the power system can be represented by a time-varying graph. A GNN can take into account the topology of the power system and thus differentiate the node information extraction in different local networks [13, 14]. However, it is essentially a local aggregation model with the problem of the k-neighbor locality. The information extraction of a GNN relies on the stacking of layers, but the number of layers k cannot grow infinitely to ensure that it can encompass the entire network. Moreover, when the number of layers k is too large, it faces the problem of feature over-smoothing. To solve this problem, a method to dynamically extract information from outside the node's neighborhood is needed. So far, there is no method to pool the remaining graph signals based on the neighborhood characteristics of the nodes.

In this paper, a self-attention mechanism is used to address the challenges of interpretability and locality. The self-attention mechanism [15], as a new interpretable model, has the potential to solve both the locality and over smoothing problems. It captures global information based on the magnitude of each object's similarity to other objects as weights, thus giving the model the ability to distinguish essential information. In a fault diagnosis method [16], self-attentiveness is used to identify and select the most important features in all nodes. For the image description task, self-attention is applied to explain the high dependence of a word in the output description on a region in the image [17]. On the task of speech recognition as text, the work in [18] is well suited

to explain the correspondence between sound segments at the input and phonemes in the output sequence. The model proposed in [19] represents the degree of association of each word in the output sequence with a particular word in the input sequence, which explains the correspondence between French and English words.

1.3 Contribution of this paper

To combine generalization performance and interpretability, this paper proposes a static-information, k-neighbor, and self-attention aggregated schema, namely, SKETCH. This is a model with the interpretability of key information aggregation, and can provide more specific transient stability information.

The main contributions of this paper are:

- Proposing SKETCH, which can provide stability predictions for each generator and the system. The structural design of the proposed model reflects the physical mechanism that the transient stability results are determined by a combination of the generators' information, the local information of the network, and global information.
- Design of a unique self-attention mechanism for solving the local information extraction problem of GNN such that the interpretability of the model is enhanced. The internal weights of this mechanism represent the strength of information interactions between nodes and thus provide richer information about the model internals.
- Development of a node regression model to implement the above structure, responding to two physical mechanisms at the internal structure level. Benefiting from the design of the nodal regression model, it is possible to explain the reasons for the information interaction between nodes and provide a post-hoc explanation for the model.

The rest of the paper is organized as follows. Section 2 formally defines the problem and introduces the structure of the SKETCH model. Section 3 presents the feature extraction model, while Sect. 4 presents the downstream model. Details of model training and decision-making are described in Sect. 5, whereas Sect. 6 develops numerical experiments on the IEEE test systems. Finally, Sect. 7 concludes the paper.

2 The data-driven transient stability assessment sketch scheme

To achieve a higher level of interpretability, the concept of sketching from the field of painting is adopted. The sketch has both parts of the detail information and the frame information in the finished draft, and thus can reflect the new requirements for DTSA: more output information and more information about the model structure reflecting the physical mechanisms. This paper refers to this new problem as DTSA sketch (DTSAS).

2.1 Task definition

More output information is beneficial to improving the trustworthiness of the model. The rotor angle information of the generators in TSA can play this key role. Therefore, the generator label is introduced as the prediction target.

In this paper, two learning tasks of DTSAS and the corresponding labels are designed to improve the robustness of the model using multi-task learning.

- Task 1 (main task): to learn the system stability label y_s of the system after fault clearing, which is the primary target of TSA.
- Task 2 (auxiliary task): to learn the stability label y_g of the generators. When the system is found to be at risk of destabilization, the next step is to specify emergency control measures for the dominant destabilized generators.

The system and generator labels are calculated based on system rotor angle difference $\Delta\delta_s$ and generator rotor angle difference $\Delta\delta_g$, given as:

$$\begin{aligned} \Delta\delta_s &= \max_{i,j} \left(\delta_i - \delta_j\right) \\ \Delta\delta_{g_i} &= \max_{j} \left(\delta_i - \delta_j\right) \end{aligned} \tag{1}$$

where the rotor angle δ_i refers to the cumulative change from t_{0-} moment to the end of the simulation, which might exceed 360°. Note that no absolute value calculation is used in this definition, and this allows the index to distinguish between the steady states of different generators. The introduction of generator labels can distinguish richer information, such as the leading generator in the dynamic process. This is helpful in subsequent control. In transient stability, different generator rotor angle states indicate different system stability mechanisms, which cannot at present be reflected by the most commonly used system stability 0/1 labels or TSI labels [10] alone.

Figure 1 shows the results of labeling the same dataset with different labels. For real power systems, the instability case is very rare. If a 0/1 index is used to indicate whether the system is unstable or not, then serious category bias problems occur. With this in mind, continuous metrics are used to represent stability. The continuous value can reflect the slight stability differences among transient stable cases. This avoids the problem of uneven distribution of sample categories. Comparing Fig. 1a, b, it

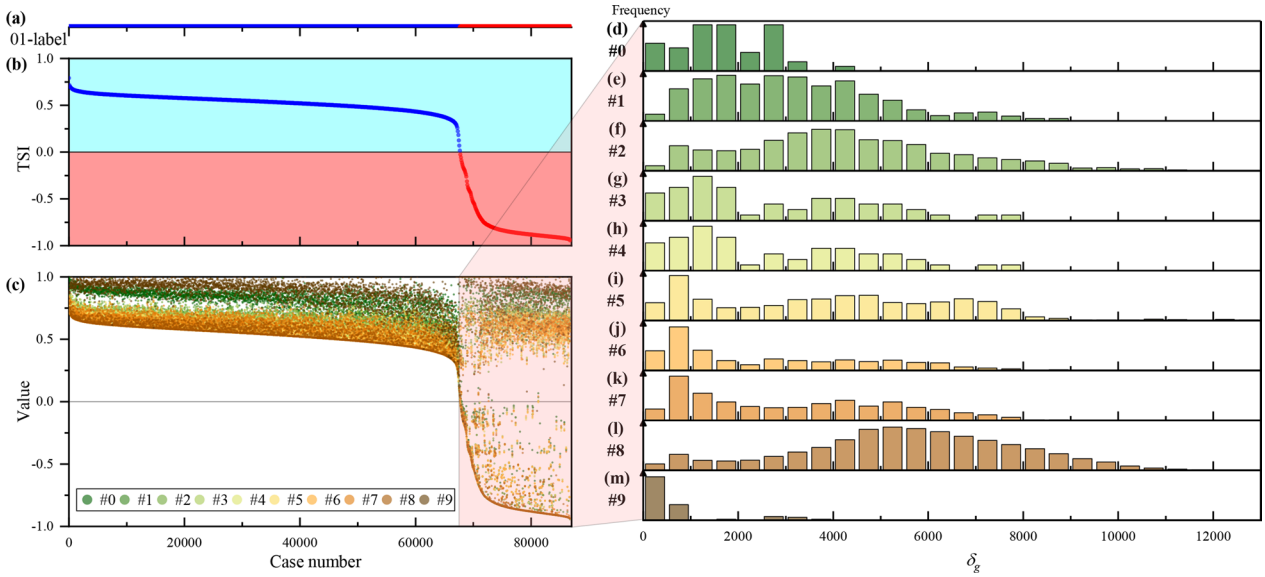

Fig. 1 Comparison of the amount of transient stability information delivered by different types of labels in the same data set. **a** The 01 label; **b** The TSI label of the system using $\Delta\delta_s$; **c** Replacing $\Delta\delta_s$ with the cumulate angle difference of generators $\Delta\delta_g$; **d**–**m** The distributions of the respective cumulative rotor angle differences for generators numbered 0–9 for all instability cases in the dataset

can be seen that the continuous value labels contain significantly more information than the 0/1 labels.

To further discriminate the differences in stability between samples, $\Delta\delta_g$ is introduced. The $\Delta\delta_g$ distribution of the unstable cases in the dataset (red boxed part in Fig. 1b) is shown in more detail in Fig. 1c. It can be seen that $\Delta\delta_g$ of the individual generators show very clear differences, which cannot be obtained by only the system label. Further details on how to obtain labels from a relative perspective are given in Sect. 5.

2.2 The structure of the SKETCH model

The proposed SKETCH model design is based on two key physical mechanisms.

Mechanism I. The stability of a node (generator) is determined by a combination of its own characteristics, local network, and external network characteristics.

Mechanism II. The system rotor angle stability is the result of the interaction of all nodes, but it is the generator node and not the other nodes that ultimately determine the stability.

The scheme is described formally using the language of graph deep learning as follows. The power system transient process can be described by a time-varying graph $\mathcal{G}_m = \mathcal{G}|_{t=t_m} = (\mathcal{A}_m, \mathcal{X}_m)$, where $\mathcal{A}$ denotes the topology of the graph and $\mathcal{X}$ denotes the parameterized information of the graph. The system in steady state $\mathcal{G}_{0-} = (\mathcal{A}_{0-}, \mathcal{X}_{0-})$ is perturbed by a perturbation $\mathcal{D}\ominus = \{\mathcal{A}_{0+}, \mathcal{A}_{c+}\}$ resulting in an actual stable outcome $y=\{y_s, y_{gi}\}$. The model f is designed to predict this result, given as $\hat{y}=\{\hat{y}_s, \hat{y}_{gi}\}$. In this paper, the graph information is denoted by a matrix $\mathcal{X}$, whose row index is the node number and column index is the feature of the node.

In this paper, the symbols are specified uniformly. For physical quantities, the subscript is the moment. For variables in the machine learning process, the subscript is the name of the associated module. All the variables are summarized in the NOMENCLATURE.

The proposed scheme is divided into three parts, i.e., the feature extraction model, the downstream model, and the decision-making process, as shown in Fig. 2.

The feature extraction part is designed with Mechanism I. First, the static feature $(\mathcal{G}_{0-}, \mathcal{D})$ is passed through the feature enhancement module f_D to obtain the distributed features $\mathcal{X}_D$. Second, the local network extraction module f_L computes the local network features $\mathcal{X}_L$ for each node. Then, the global attention module f_G computes the features $\mathcal{X}_D$ outside the local network. Finally, the three features are combined together as the transient stable features of the nodes.

The node-level downstream model is designed according to Mechanism II. First, the mask module eliminates the information of non-generator nodes

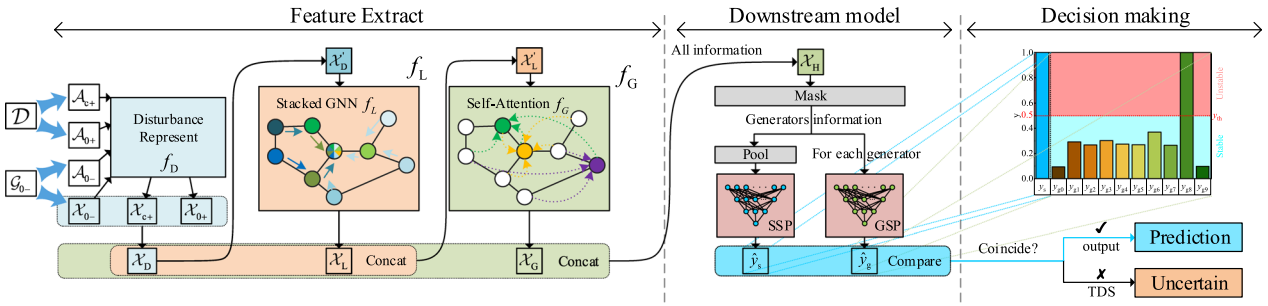

Fig. 2 The information flow of the proposed method. $\mathcal{A}$ denotes the topology of the graph and $\mathcal{X}$ denotes the parameterized information of the graph. In the feature extract session, the feature $\mathcal{X}_{0-}$ and the topology $\mathcal{A}_{0-}, \mathcal{A}_{0+}, \mathcal{A}_{c+}$ are used as input and processed by three modules to finally obtain $\mathcal{X}_H$. The downstream model takes $\mathcal{X}_H$ as input and obtains the system label y_s and generator label y_g by two predictors, respectively

and ensures that the information for the subsequent process comes only from the generator nodes. Then, on the one hand, these features are passed through the pooling step and the system stability predictor (SSP) to obtain the system stability prediction, while on the other hand, the features of each generator node are fed independently into the same generator stability predictor (GSP) to get the corresponding generator labels. Note that there is only one GSP, i.e., different generator features use a consistent identification logic to get their respective results.

In the decision-making process, the output of the model can be evaluated more precisely based on the correspondence that exists between the generator label and the system label. When the generator label agrees with the stability indicated by the system label, the predicted value is output, otherwise, this sample is uncertain and requires TDS to determine its stability.

SKETCH is distinguished from other models by the following two features.

- The model establishes an analytic mapping from static information to stable results and responds to physical Mechanisms I and II. The multi-layer perceptron (MLP) is not interpretable, because its parsed form has no physical meaning.
- The node-level feature extraction structure makes the information interactions between nodes transparently visible. This prevents the model from achieving its effect by over-fitting the features of non-generator nodes, while we do not consider that the model correctly learns the mechanism of transient stabilization in this case. The existing GNN-based work provides interpretability of local feature aggregation [13, 14, 20]. However, since the full graph features are used to discriminate the system stability, it is not possible to determine whether the key features come from the generator nodes.

2.3 Model application scheme

The application process of the model is divided into three stages: offline training → online prediction → decision-making.

In the offline training stage, a large amount of TDS data are pre-processed to form the dataset on which the overall fine-tuning phases of training are performed. The fully trained model in the offline phase will be used for online stabilization scanning tasks executed periodically (e.g., every 15 min). In the online prediction stage, a stability evaluation is performed based on static and disturbance data, and the evaluation results are validated by the online decision system to provide higher accuracy results, and decide which samples need to be further simulated accordingly.

Finally, a well-trained model will be used for online applications that can provide fast and high-resolution evaluation results for the current power system operating mode and specified disturbances.

3 Node-level global feature extraction model

3.1 Using static information as input

By solving the DAEs, TDS takes $(\mathcal{G}_{0-}, \mathcal{D})$ as input and obtains $\mathcal{G}_t, t > 0$ which indicates the stability. However, it cannot give any analytic mapping from $(\mathcal{G}_{0-}, \mathcal{D})$ to y, so it is regarded as a black-box. To avoid the hindrance of interpretability by using dynamic information generated by TDS, the steady-state information is used as input.

Using static information including fault occurrence and clearance information, TDS can represent transient processes triggered by a single fault of fixed duration, but for DTSAS there are two challenges, i.e., the sparsity of the disturbance location information and parameter sensitivity. The existing scheme performs well in modeling status at pre-fault (0−) and after-fault (0+), but is too simplified for the representation of fault clearance (c+) results. The perturbation at the t_{c+} moment is simply a change in two elements of the admittance matrix.

Here a power flow calculation is used for a distributed representation of the perturbation at the t_{c+} moment. This allows a distributed representation of the impact of disturbances on electrical quantities while requiring no TDS, making it a better choice. The online measured power flow $\mathcal{X}_{0-}$ is used to represent the state at t_{0-}, which contains the load active power $P^{(\mathrm{L})}$, load reactive power $Q^{(\mathrm{L})}$, generator active power $P^{(\mathrm{G})}$, generator reactive power $Q^{(\mathrm{G})}$, bus voltage magnitude V, bus voltage angle θ at each node, as:

$$\mathcal{X}_{0-}=\left\{P_{0-,i}^{(\mathrm{L})}||Q_{0-,i}^{(\mathrm{L})}||P_{0-,i}^{(\mathrm{G})}||Q_{0-,i}^{(\mathrm{G})}||V_{0-,i}||\theta_{0-,i}\right\}^{N} \tag{2}$$

where || denotes vector dimension splicing.

Without TDS, the deterministic conditions at t_{0+} and t_{c+} cannot be obtained, and therefore, the power flow state cannot be determined. Thus, a linear estimation of the short-circuit currents is used to represent the state $\mathcal{X}_{0+}$ at t_{0+}, and the detailed procedure of the method can be found in [20]. However, this method relies on the assumption of constant rotor angle at t_{0+} and cannot be used to estimate t_{c+}.

With the admittance matrix Y_{0-} at t_{0-}, the power flow equation is given as:

$$P_{0+}+jQ_{0+}=\dot{V}_{0-}Y_{0+}^{*}\dot{V}_{0-}^{*} \tag{3}$$

where * denotes conjugate and $\dot{V}$ denotes $V\angle\theta$.

For the t_{c+} moment, the only disturbance is a fault clearing operation, which is essentially a change in the Y matrix. From this perspective, a distributed representation of the variation of Y_{c+} is obtained with the help of (4) and $\dot{V}_{0-}$.

First, a group of powers are fictitiously represented as:

$$P^{'}+jQ^{'}=\dot{V}_{0-}Y_{c+}^{*}\dot{V}_{0-}^{*} \tag{4}$$

The difference between these powers and the t_{0-} state arises entirely from the change in the derivative matrix, so $P^{(\mathrm{imp})}+jQ^{(\mathrm{imp})}$ is computed to serve as an enhanced feature of t_{c+}, as:

$$P^{(\mathrm{imp})}+jQ^{(\mathrm{imp})}=P^{'}+jQ^{'}-\left(P_{0-}^{(\mathrm{L})}+jQ_{0-}^{(\mathrm{L})}\right) \tag{5}$$

The representation of disturbance at t_{c+} is obtained by:

$$\mathcal{X}_{c+}=\left\{P_{i}^{(\mathrm{imp})}\middle\|Q_{i}^{(\mathrm{imp})}\right\}^{N} \tag{6}$$

Finally, the three parts of information are combined together to obtain the pre-fault information and the results after encoding the disturbance information $\mathcal{X}_{\mathrm{D}}$, which is given as:

$$\mathcal{X}_{\mathrm{D}}=\mathcal{X}_{0-}\|\mathcal{X}_{0+}\|\mathcal{X}_{c+} \tag{7}$$

The disturbance information is computed by power flow and is decentralized to be encoded into the electrical characteristic quantities of the nodes. This coding method requires a small amount of storage, and the size is only related to the number of system buses rather than the number of lines. Another advantage is that this coding is dense and can provide more effective information than one-hot encoding.

3.2 The local information extraction module

The stability problem of generators after large disturbances is closely related to the local topology of the network. A GNN is a proper method for processing graph information [21].

A Graph Convolution Network (GCN) [22] is capable of aggregating node neighborhood information with a fixed weight. Let the input $\mathcal{X}_{\mathrm{D}}=\left\{x_{\mathrm{D},i}\right\}$. The GCN is represented as:

$$f_{\mathrm{GCN}}(\mathcal{X}_{\mathrm{D}},\mathcal{A})=\sigma\left(\sum\nolimits_{j\in\mathcal{N}_i\bigcup\{i\}}\frac{\mathcal{A}_{ij}}{\sqrt{\hat{d}_i\hat{d}_j}}W_{\mathrm{C}}x_{\mathrm{D},j}\right) \tag{8}$$

where $\sigma(\cdot)$ is the ReLU function, $\mathcal{N}_i$ and $\hat{d}_i$ represent the neighbors and degree of the ith node, respectively. W_{C} is a trainable weight matrix, and the edge weight $\mathcal{A}_{ij}$ is the element of the adjacency matrix $\mathcal{A}$.

The original GCN is essentially an average aggregation of neighbors because $\mathcal{A}$ is an 0/1 matrix. It considers that the admittance of the line plays a decisive and differentiated role in the state propagation of the buses, so a physically meaningful one is used instead of this average aggregation, whose element is:

$$\mathcal{A}_{ij}=\left|y_{ij}\right|/\left|y_{ii}\right| \tag{9}$$

Besides the information propagation path with fixed weights, A Graph Attention Network (GAT) [23] is introduced to allow the model to autonomously construct other information propagation strengths:

$$\begin{aligned}f_{\mathrm{GAT}}(\mathcal{X}_{\mathrm{D}},\mathcal{A})&=\sigma\left(\sum\nolimits_{j\in\mathcal{N}_i}\alpha_{ij}^{(\mathrm{A})}W_{\mathrm{A}}x_{\mathrm{D},j}\right)\\ \alpha_{ij}^{(\mathrm{A})}&=\frac{\exp\left(\sigma_{\mathrm{Leaky}}\left(a_{\mathrm{A}}^{\mathrm{T}}\left[W_{\mathrm{A}}x_i\|W_{\mathrm{A}}x_{\mathrm{D},j}\right]\right)\right)}{\sum_{j\in\mathcal{N}_i}\exp\left(\sigma_{\mathrm{Leaky}}\left(a_{\mathrm{A}}^{\mathrm{T}}\left[W_{\mathrm{A}}x_i\|W_{\mathrm{A}}x_{\mathrm{D},j}\right]\right)\right)}\end{aligned} \tag{10}$$

where a_{A} and W_{A} are learnable weights, and $\sigma_{\mathrm{Leaky}}(\cdot)$ refers to the LeakyReLU function. To combine a channel for information aggregation consistent with physical mechanisms and a channel for autonomous learning of

models, the outputs of GAT and GCN are added together as a layer of GNN transformations f_{GNN}, as:

$$f_{\mathrm{GNN},k}(\mathcal{X},\mathcal{A}) = f_{\mathrm{GCN},k}(\mathcal{X},\mathcal{A}) + f_{\mathrm{GAT},k}(\mathcal{X},\mathcal{A}) \quad (11)$$

The cascaded residual GNN structure [14] is used to extract information about the k-order neighbors, expressed by:

$$f_{\mathrm{GNN}}^{(k)}(\mathcal{X},\mathcal{A}) = f_{\mathrm{GNN},k}\left(f_{\mathrm{GNN}}^{(k-1)}(\mathcal{X},\mathcal{A}),\mathcal{A}\right) + f_{\mathrm{GNN}}^{(k-1)}(\mathcal{X},\mathcal{A}) \quad (12)$$

In this paper $k = 3$. The information flow of the stacked structure is shown in Fig. 3.

The above GNN structure for the t_{0-} topology $\mathcal{A}_{0-}$ and the t_{c+} topology $\mathcal{A}_{c+}$ are developed, respectively. Taking $\mathcal{X}_{\mathrm{D}}$ and $(\mathcal{A}_{0-},\mathcal{A}_{c+})$ as inputs, the neighborhood information extraction module is expressed by:

$$\begin{aligned}\mathcal{X}_{\mathrm{L}} &= f_{\mathrm{L}}(\mathcal{X}_{\mathrm{D}},\mathcal{A}_{0-},\mathcal{A}_{c+}) \\ &= \frac{1}{2}\cdot\left(f_{\mathrm{GNN}}^{(k)}(\mathcal{X}_{\mathrm{D}},\mathcal{A}_{0-}) + f_{\mathrm{GNN}}^{(k)}(\mathcal{X}_{\mathrm{D}},\mathcal{A}_{c+})\right)\end{aligned} \quad (13)$$

Note that this is information smoothing on the k-neighborhood subgraph, where the differences between nodes fade away as the depth deepens, and the depth is not deep enough for the nodes to get information about the full graph.

3.3 Subgraph equivalence by self-attention

Generally, TSA is a global problem. What affects the stability of the generator is not only its k-neighbor nodes, but also other nodes in the distant area.

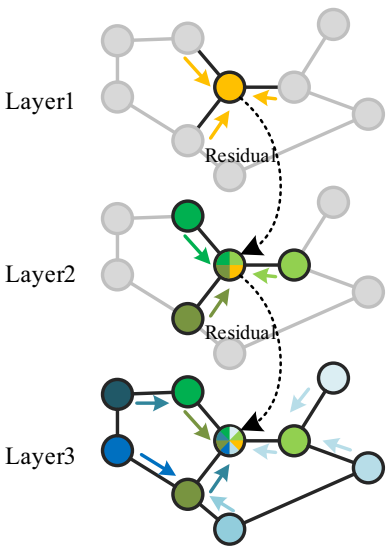

Fig. 3 The information flow of stacked residual GNN layers. As the number of GNN layers increases, the information of the local network is aggregated at the center of the node

Therefore, although the GNN-based feature extraction layer can capture the key information among several orders of the node's neighbors that affect its stability, a means of aggregating the information of distant nodes is still needed.

A self-attention-based method is thus proposed to solve this problem. The core of the self-attention mechanism is to reorganize the global information based on the similarity of information among the different nodes. In this paper, the similarity measure is replaced with a dissimilarity measure, so that the self-attention mechanism captures not similar information between nodes, but dissimilar information between nodes.

Let the input $\mathcal{X}_{\mathrm{L}}^{'} = \mathcal{X}_{\mathrm{D}} \| \mathcal{X}_{\mathrm{L}}$, the self-attention mechanism first calculates $\mathcal{X}_{\mathrm{Q}}$, $\mathcal{X}_{\mathrm{K}}$ and $\mathcal{X}_{\mathrm{V}}$ by three different linear transformations, as:

$$\begin{aligned}\mathcal{X}_{\mathrm{Q}} &= W_{\mathrm{Q}}\mathcal{X}_{\mathrm{L}}^{'} \\ \mathcal{X}_{\mathrm{K}} &= W_{\mathrm{K}}\mathcal{X}_{\mathrm{L}}^{'} \\ \mathcal{X}_{\mathrm{V}} &= W_{\mathrm{V}}\mathcal{X}_{\mathrm{L}}^{'}\end{aligned} \quad (14)$$

Then the attention matrix $\mathbf{A}$ is calculated by a normal function f_{Norm} based on $\mathcal{X}_{\mathrm{Q}}$ and $\mathcal{X}_{\mathrm{K}}$, as:

$$\mathbf{A} = f_{\mathrm{Norm}}\left(\mathcal{X}_{\mathrm{Q}}\mathcal{X}_{\mathrm{K}}^{\mathrm{T}}\right) \quad (15)$$

Finally, the output is obtained by weighting $\mathcal{X}_{\mathrm{V}}$ according to $\mathbf{A}$, as:

$$\mathcal{X}_{\mathrm{G}} = \sigma(\mathbf{A}\mathcal{X}_{\mathrm{V}}) \quad (16)$$

The normal function f_{Norm} is a scaling normalization function to keep the intensity of the graph signal constant before and after the transformation. The SoftMax function is the most commonly used f_{Norm}. Let the intermediate results $\mathcal{X}_{\mathrm{Q}}\mathcal{X}_{\mathrm{K}}^{\mathrm{T}} = \mathbf{A}^{'} = a_{\mathrm{A}}^{'}$ and $\mathbf{A} = f_{\mathrm{Norm}}\left(\mathbf{A}^{'}\right) = \{a_{\mathbf{A}}\}$. Then the SoftMax function can be expressed by:

$$a_{\mathbf{A},i} = \exp\left(a_{\mathbf{A},i}^{'}\right) \Big/ \sum\nolimits_{j} \exp\left(a_{\mathbf{A},j}^{'}\right) \quad (17)$$

Note that the scaling normalization of the existing work is performed through the SoftMax function, which essentially encourages nodes to aggregate information that is similar to a high degree.

From the point of view of capturing the impact of other nodes of the power system on the generator node, the dissimilar rather than similar information should be captured, so the use of the SoftMin function is proposed:

$$a_{\mathbf{A},i} = \exp\left(-a_{\mathbf{A},i}^{'}\right) \Big/ \sum\nolimits_{j} \exp\left(-a_{\mathbf{A},j}^{'}\right) \quad (18)$$

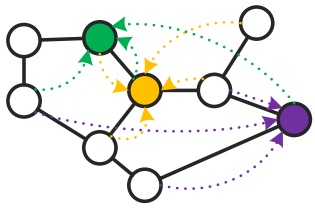

Fig. 4 The information flow of attention mechanism in GNN perspective. Each node extracts differentiated information from other nodes in the full graph without scope restrictions, and the strength of the extraction depends on the values of the elements in the attention matrix

The role of node-level attention can be explained from the perspective of GNN, as:

$$f_{\mathrm{G}}\left(\mathcal{X}_{\mathrm{L}}^{'}, \mathbf{A}\right)=\sigma\left(\sum\nolimits_{1 \leq j \leq N} \mathbf{A}_{ij} W_{\mathrm{V}} x_{\mathrm{L},j}^{'}\right) \tag{19}$$

It can be seen as performing GNN on a weighted and directed strongly connected graph, and thus it enables the extraction of information outside the k-neighborhood of a node. Therefore, the attention matrix $\mathbf{A}$ is equivalent to $\mathcal{A}$ in (8) and $\alpha^{(\mathrm{A})}$ in (10), and thus serves as a global information extraction for other nodes, as shown in Fig. 4.

Taking the local network information $\mathcal{X}_{\mathrm{L}}^{'}=\mathcal{X}_{\mathrm{D}}^{'} \| \mathcal{X}_{\mathrm{L}}$ as input, the global graph equivalence progress can be expressed as:

$$\mathcal{X}_{\mathrm{G}}=f_{\mathrm{G}}\left(\mathcal{X}_{\mathrm{L}}^{'}\right)=\sigma\left(f_{\mathrm{Norm}}\left(W_{\mathrm{Q}} \mathcal{X}_{\mathrm{L}}^{'} \mathcal{X}_{\mathrm{L}}^{'\mathrm{T}} W_{\mathrm{K}}\right) \cdot W_{\mathrm{V}} \mathcal{X}_{\mathrm{L}}^{'}\right) \tag{20}$$

So far, the input information is sequentially encoded by the disturbance representation, the local and global encoding of the GNN, and finally becomes a hidden feature $\mathcal{X}_{\mathrm{H}}$, as:

$$\mathcal{X}_{\mathrm{H}}=\mathcal{X}_{\mathrm{D}} \| \mathcal{X}_{\mathrm{L}} \| \mathcal{X}_{\mathrm{G}} \tag{21}$$

4 Downstream model

The downstream model implements node-level regression of stability labels to predict the stability of each generator based on its hidden features. This design distinguishes SKETCH from other DTSA schemes. Based on the downstream model, strong interpretability and decoupling of the number of model parameters from the system size are achieved. The detailed model is shown in Fig. 5.

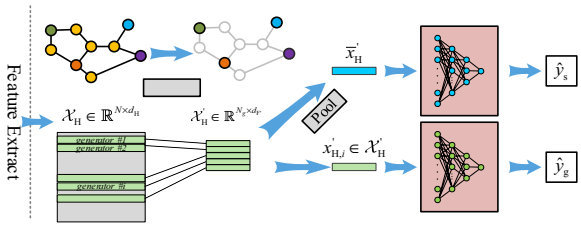

Fig. 5 The structure and feature shape of the downstream model

4.1 Node-level mask

Physically, the stability results of the system are influenced by the characteristics of all nodes, but the final judgment of whether the system is stable or not depends only on the state quantities of the generator rather than the other nodes. In other words, after an effective representation learning process, the information of other nodes is noise for the stability judgment process, and cannot be used as input, and thus needs to be explicitly excluded.

However, all the existing DTSA models based on the stability information make use of the hidden features of other nodes to some extent. For example, they spread the features of all nodes into MLP [14] or downscale the features of all nodes before feeding them into MLP [20]. These structures make the interpretability of the model limited because it is not possible to determine whether the stability-determining information comes from the generator or non-generator nodes. The former approach also makes the number of parameters of the downstream model grow with the size of the system.

A masking mechanism is proposed to implement this physical mechanism as follows:

$$\mathcal{X}_{\mathrm{H}}^{'}=\left\{W_{\mathrm{M}} x_{\mathrm{H},g_i}^{\mathrm{T}} \middle| x_{\mathrm{H},g_i} \in \mathcal{X}_{\mathrm{H}}, i=0,1,\ldots,N_g\right\} \tag{22}$$

where N_g is the number of generators, g_i is the index of the ith generator, $W_{\mathrm{M}} \in \mathbb{R}^{d_{\mathrm{H}} \times d_{\mathrm{H}}^{\prime}}$ is a learnable matrix to reduce dimension, and $d_{\mathrm{H}}^{\prime}=d_{\mathrm{H}}/4$.

After masking, the latent feature $\mathcal{X}_{\mathrm{H}} \in \mathbb{R}^{N \times d_{\mathrm{H}}^{\prime}}$ is reduced to $\mathcal{X}_{\mathrm{H}}^{'} \in \mathbb{R}^{N_g \times d_{\mathrm{H}}^{'}}$, which only contains the features of the generator nodes.

4.2 Graph pooling

To further make the scale of $\mathcal{X}_{\mathrm{H}}^{'}$ independent of the system scale, $\mathcal{X}_{\mathrm{H}}^{'}$ is reduced into a vector $\overline{x}_{\mathrm{H}}^{'}$ using a global mean pooling operation, as:

$$\overline{x}_{\mathrm{H}}^{'}=\frac{1}{N_{\mathrm{g}}} \sum\nolimits_i x_{\mathrm{H},i}^{'} \tag{23}$$

4.3 Stability predictor based on MLP

Finally, two predictors based on full-connected (FC) layers, system stability predictor (SSP) and generator stability predictor (GSP), are proposed to carry out the system stability results and generator stability results, respectively. The parameters of the two predictors are 21-21-16-1.

The input of SSP is $\mathcal{X}_{\mathrm{S}}$ and output is y_s, whereas the input of GSP is $x'_{\mathrm{H},i} \in \mathcal{X}'_{\mathrm{H}}$ and output is $y_{\mathrm{g},i}$. Note that the same GSP is used to predict the stability of different generators.

5 Training and decision making

5.1 Enhance separability of labels

To make full use of the detailed information of the stable labels, this paper does not directly classify the input samples as 0 or 1, but returns the predicted values to the vicinity of the labels and then makes a classification judgment on stability by a threshold value. For this purpose, the labels proposed in Sect. 2.1 are processed.

First, the original value $\Delta\delta_{\mathrm{X}}, \mathrm{X} \in \{\mathrm{s}, \mathrm{g}_i\}$ is normalized to [0, 1] interval by the tanh function, as:

$$y_{\mathrm{X}} = \tanh\left(\tanh^{-1}\left(y_{\mathrm{th}}\right) \cdot \frac{\Delta\delta_{\mathrm{X}}}{\delta_{\mathrm{th}}}\right) \tag{24}$$

where $\delta_{\mathrm{th}} = 180^{\circ}$ is the stable threshold of angle, and $y_{\mathrm{th}} = 0.5$ is the stable threshold of the labels. Under these parameters, the label $y_{\mathrm{X}} \le y_{\mathrm{th}}$ indicates that the system is stable.

The advantage of this processing method is that the parameters of the labels are independent of the data set, and their numerical significances before and after processing are order-preserving for different data sets.

5.2 Loss function

In this paper, the smooth L1 loss function J_{SL} is used to fit the labels, and a directional loss function J_{D} is proposed to improve the classification performance near the threshold. For each label y_{X}, the loss function J can be expressed as:

$$J\left(\hat{y}_{\mathrm{X}}, y_{\mathrm{X}}\right) = J_{\mathrm{SL}}\left(\hat{y}_{\mathrm{X}}, y_{\mathrm{X}}\right) + J_{\mathrm{D}}\left(\hat{y}_{\mathrm{X}}, y_{\mathrm{X}}\right) \tag{25}$$

The SmoothL1 loss function is:

$$J_{\mathrm{SL}}\left(\hat{y}_{\mathrm{X}}, y_{\mathrm{X}}\right) = \begin{cases} \left(\hat{y}_{\mathrm{X}} - y_{\mathrm{X}}\right)^2/(2\beta) & \left|\hat{y}_{\mathrm{X}} - y_{\mathrm{X}}\right| < \beta \\ \left|\hat{y}_{\mathrm{X}} - y_{\mathrm{X}}\right| - \beta/2 & \text{otherwise} \end{cases} \tag{26}$$

It considers the advantages of robust regression of L1 loss for outliers and convergence of MSE loss through a segmented loss for the β-error interval, taking $\beta = 0.1$ in the training phase. Considering the L2 loss of the model parameters, the total loss function of the model is:

$$J_{\mathrm{Total}} = J_{\mathrm{SL}}\left(\hat{y}_{\mathrm{g}}, y_{\mathrm{g}}\right) + N_{\mathrm{g}} J_{\mathrm{SL}}\left(\hat{y}_{\mathrm{s}}, y_{\mathrm{s}}\right) + \eta \sum\nolimits_f \left\| W_f \right\|_2^2 \tag{27}$$

where $\left\| W_f \right\|_2$ denotes the 2-Norm of the parameters of the module f and η=0.0005 is its coefficient.

5.3 Decision enhancement module

In the decision phase, the stability classification labels are calculated based on the continuous values of the model output $\hat{y}_{\mathrm{X}}$:

$$\mathrm{U}_{\mathrm{X},m} \Leftrightarrow y_{\mathrm{X},m} \ge y_{\mathrm{th}} \tag{28}$$

A decision enhancement module (DEM) is proposed to enhance the performance based on model output. This strategy improves the accuracy of the final decision by identifying samples that are difficult to discriminate by the model and selecting them to further determine the stability using TDS.

The generator stability label is causally related to the system stability label, and the cases are labeled as uncertain when the predictions conflict with each other. These uncertain cases are then sent to TDS to determine stability. Overall, this strategy trades a small additional cost of TDS for an increase in decision accuracy. For case m, the judgment of DEM η_{m} can be expressed as:

$$\eta_m = \left(\mathrm{US}_m \oplus \mathrm{UG}_m\right) \vee \mathrm{CO}_m \tag{29}$$

where $\oplus$ refers to logical operator XOR, $\vee$ is logical, and US_m, UG_m and CO_m are logical values to be used in classification:

$$\begin{aligned} \mathrm{US}_m &= \mathrm{U}_{\mathrm{s},m} \\ \mathrm{UG}_m &= \bigcup\nolimits_i \mathrm{U}_{\mathrm{g}_i,m} \\ \mathrm{CO}_m &\Leftrightarrow \left|y_{\mathrm{X}} - y_{\mathrm{th}}\right| < \beta \end{aligned} \tag{30}$$

5.4 Performance metrics

To test the performance of the regression, the mean squared error (MSE) is introduced:

$$\mathrm{MSE}_{\mathrm{X}} = \frac{1}{N_{\mathrm{m}}} \sum\nolimits_i \left\| \hat{y}_{\mathrm{X}} - y_{\mathrm{X}} \right\|_2^2 \tag{31}$$

where N_{m} is the case number of the dataset.

Here, accuracy (ACC), recall (REC), and precision (PRE) are introduced to evaluate the performance of the classification. To measure the combined effect of

Table 1 Confusion matrix

	Actually unstable	Actually stable
Predicted unstable	TP	FN
Predicted stable	FP	TN

the model for the imbalance categories, the kappa index (KAP) is also used in the evaluation.

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{32}$$

$$\mathrm{REC} = 1 - \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \tag{33}$$

$$\mathrm{PRE} = 1 - \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} \tag{34}$$

$$\mathrm{KAP} = \frac{\mathrm{ACC}-p_e}{1-p_e}$$
$$p_e = \frac{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})-(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2} \tag{35}$$

These metrics are calculated based on the confusion matrix in Table 1.

6 Results and discussion

In this section, SKETCH is demonstrated and compared to other node-level methods in the IEEE 39-bus system and IEEE 300-bus system.

6.1 Dataset generation and parameter setting

The datasets are generated by TDS on PSD-BPA, where all generators use the 6^{th}-order model with the excitation system of IEEE model type I. Following the principle of not creating islands, different topologies are obtained by cutting 0/1/2 lines respectively. The global load and generator levels start at 75% and grow to 120% in 5% steps, and these increments are randomly assigned to all generators and loads. The disturbance is a three-phase short-circuit fault, occurring at the first section or end of the line, with a duration of 0.1 s before the protection operates and isolates the faulty line. The dataset contains 86,756 cases, of which 19,040 are unstable and 67,595 are stable, and are randomly divided into training set, validation set, and test set according to 6:2:2 ratios. Note that the developed model does not preset the knowledge of the dynamic properties of the system components, while this knowledge is extracted from the data and solidified in the structure and parameters of the model through a training process. If the dataset reflects a system containing power electronic components, then a stability discriminative model adapted to these components can be trained with these data.

Between the different schemes compared below, a linear layer without nonlinear transformation capability will be added for dimensional transformation when there is a difference between the input and model dimensions, if not specifically stated. This approach minimizes the variation in the number of model parameters caused by differences in the dimensions of a particular layer, thus enhancing the comparability of the results.

All models are optimized by an Adam optimizer with batch size 256. The learning rate starts from 0.002 and decays by 10% every 10 batches, for a total of 100 epochs of training. If not differently specified, the proposed strategy is leveraged to train all the models and the DEM module is disabled in the proposed scheme in the remaining subsections.

The program is implemented using the PyTorch [24] and PyTorch-Geometric frameworks [25], and the computing platform is Intel i7-9700 CPU and Nvidia GTX 1660Ti.

6.2 The hyper-parameter setting of the GNN structure

After disturbance representation, the transient features of the system are fed to the GNN module for feature extraction from the local network. The values of the relevant hyper-parameters are compared in this section, including the number of layers of GNNs and the connection patterns of different GNNs at each layer.

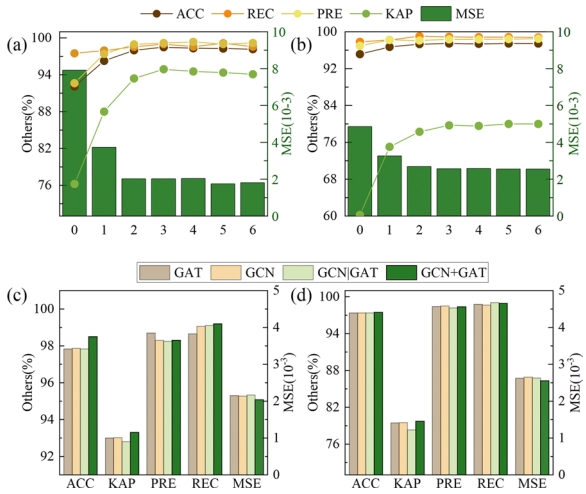

Fig. 6 Comparison on different GNN structures. **a**, **b** Show the variation of model performance with the number of GNN layers, while **c**, **d** show the effect of different structure within each GNN layer

The results of the comparison of different layers according to the connection of (12) are presented in Fig. 6a, b. As k grows from 0 to 3, the metrics obtain a substantial improvement, and the optimal value is achieved at $k = 3$, indicating that a certain amount of local information perception helps to improve the model performance. However, when k continues to grow, the metrics no longer show significant growth. Therefore, $k = 3$ is used.

Then, several other ways of connecting GNNs within layers are compared around (11), using only GAT, GCN, and splicing using both, respectively. As shown in Fig. 6c, d, the connection proposed in (11) is the best form.

6.3 Verification of disturbance representation

To demonstrate the efficiency of the proposed disturbance representation method, the real and estimated values are mixed to form the following sets of input value schemes. Each input scheme consists of three letters representing the source of information at t_{0-}, t_{0+}, t_{c+}. R denotes the actual value, E denotes the estimated value, and a short bar − denotes that this is not used. Detailed scenario information is shown in Table 2.

The results are shown in Fig. 7. As seen, overall, the performance metrics are almost always better for the solutions with more temporal information entered. Compared with R - -, the kappa of RR - and RRR increase by 7.12% and 9.61%, respectively. When the inputs are estimates (R - -, RE -, REE), kappa improves by 7.32% and 8.71% in that order. A similar conclusion also holds for other performance metrics.

It is noted that when the real value is not used (RE - and REE), the results close to those obtained using TDS (RR - and RRR) for the features are also obtained using the estimates proposed in this paper. This reflects the potential of this estimation method for the application of fast pre-fault scanning.

6.4 Overall performance of SKETCH on the test set

This section shows the statistics of the output on the IEEE 39-bus test set. Figure 8a shows the prediction of the system label y_s. The orange bars represent the distribution of the test set, while the box plots and blue data points at the corresponding locations represent the performance of the model for this subset. In the subset with more data (0.2–0.4), the regression error is relatively small, while in the subsets near y_{th} or with fewer cases, the error is relatively large, indicating the lack of data.

Table 2 Scenario details and KAP on task 1

Scenario	Input information $\mathcal{X}_D$	KAP (%)
R–	$\mathcal{X}_{0-}$	86.22
RE-	$\mathcal{X}_{0-}, \mathcal{X}_{0+}$	93.54
RR-	$\mathcal{X}_{0-}$, actual value at t_{0+}	93.34
REE	$\mathcal{X}_{0-}, \mathcal{X}_{0+}, \mathcal{X}_{c-}$	94.93
RRE	$\mathcal{X}_{0-}$, actual value at $t_{0+}, \mathcal{X}_{c-}$	94.50
RRR	$\mathcal{X}_{0-}$, actual value at t_{0+}, t_{c+}	95.83

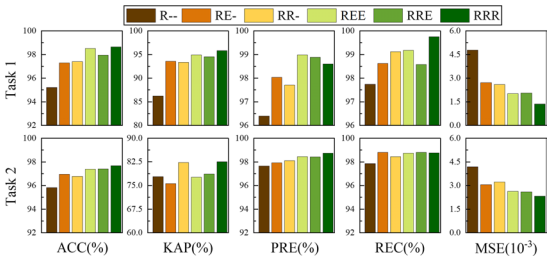

Fig. 7 Model performance with different input information

To demonstrate how well the model learns the overall distribution, Fig. 8b shows the means and quartiles of the sketch on the test set, and Fig. 8c represents the error distribution for each label. Overall, the model achieves the regression of y_s and y_{gi} with a relatively low and consistent error level.

6.5 Explanation of the internal mechanism

From the perspective of information flow, the mechanism of the model can be further analyzed. In the above experiments, the SKETCH model exhibits

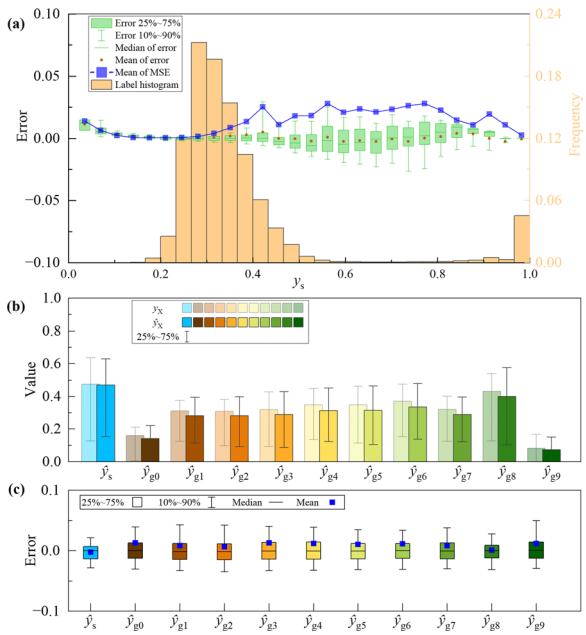

Fig. 8 Statistic results of SKETCH output on the IEEE 39-bus test set. Here are the statistics of the system labels (**a**), the sketch (**b**) and the error of labels (**c**), respectively

superior performance in task 2. This implies that the model extracts the key information that determines transient stability.

Since each generator shares the same evaluator GSP, it means that the nodes' differentiated information can only come from the feature extraction model: self-features, neighborhood information, and global attention information.

The transient stability problem is a global problem, and the perceptual field of the three GNN layers in the model cannot cover the whole network, implying that the information outside the k-order neighborhood of the generator can only be obtained by the attention module.

The weights of the attention matrix represent the intensity of information importance. The i^{th} row and j^{th} column of the matrix represent the importance of node j to node i.

The information capture of the model is explored for different cases by visualizing the attention matrix.

Both a stable and an unstable case are shown in Fig. 9. In case-1, line #17–#27 is disconnected for maintenance and the system is operating at 75% load level with bus #16 being the central node of the system. At $t_0 = 0$ s, a short circuit occurs on line #23–#24 near bus #24, lasting for 0.1 s before the protection operates and the line is disconnected. During 4 s, the system remains stable with a maximum angle difference of 99.3 degrees, which occurs between generator #0 (bus #30) and generator #6 (bus #36), as shown in Fig. 9b. The actual output sketch of the model is shown in Fig. 9c, which indicates that the model predicts the actual sketch of the system with a very small error. The attention matrix of the case is shown in Fig. 9d, buses #16, #24, #38, and #39 are given prominent weights.

In Fig. 9a, e–g, case-2 shows an unstable case. In this case, lines #5–#6 and #8–#9 are disconnected for maintenance and the system is operating at 105% load level with bus #8 being a new central node of the system. At $t_0 = 0$ s, a short circuit occurs on line #26–#29 near bus #29. The system is then destabilized with a maximum angle difference of 8033.7 degrees, which occurs between generator #0 (bus #30) and generator #8 (bus #38), as shown in Fig. 9e. Figure 9f demonstrates that the predicted results are very close to the actual values. In the attention matrix of this example, buses #8, #29, and #38 receive larger weights.

Physically, the fault bus, the generators with the largest angle difference, and the topologically significant bus are three key factors in system stability. In case-1 and case-2, the attention module shows greater attention weights to the above three factors, suggesting that SKETCH may have learned this mechanism. Note that the first two are different for each sample and the model can distinguish them effectively, indicating that this part of the model is knowledge that can be generalized.

The attention between nodes is not the same in each sample. However, for similar topologies, effective models always give consistently high attention to topologically important nodes. The attention entropy and the mean value of attention are used on the test set to measure the consistency of this conclusion and examine whether the model can identify important nodes.

The entropy of the node i is calculated by:

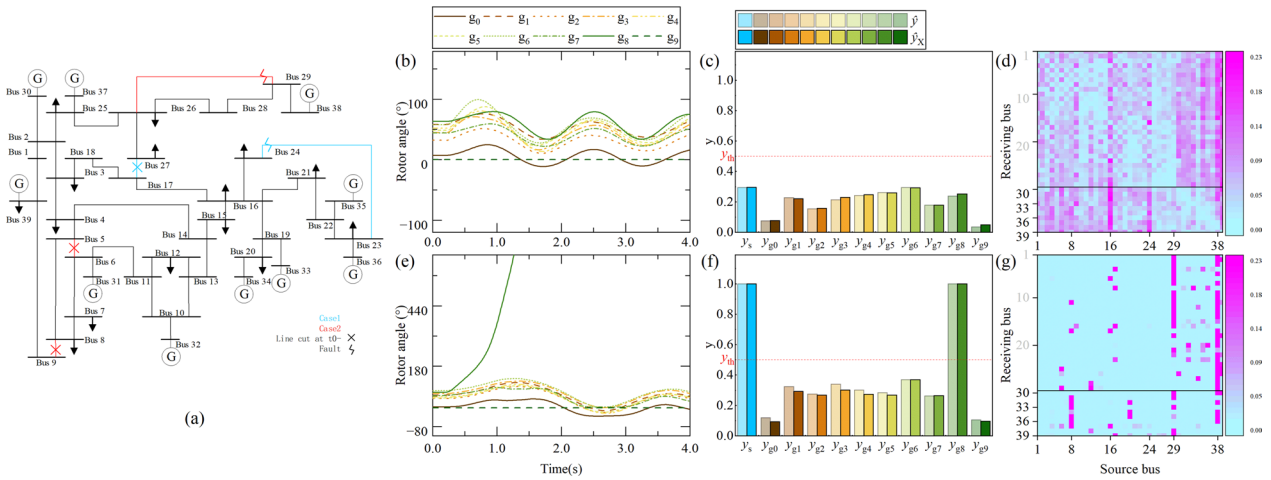

Fig. 9 Explanations of two cases. **a** Is the topology of the system, where the lines and markings with specific colors (case-1 in blue and case-2 in red) indicate the difference on the original system. **b**–**d** are the rotor angle curve, output labels and attention weights of case-1, respectively, while **e**–**g** are those of case-2

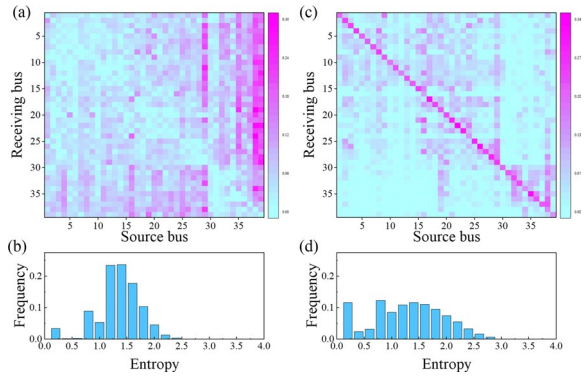

Fig. 10 Consistency of case-level explanations. **a**, **b** are results of the proposed SoftMin-based attention module, while **c**, **d** are results of the SoftMax-based attention module

Table 3 Performance on task 1 in ablation experiments

Part	ACC	PRE	REC	KAP
Original	98.50	98.98	99.18	94.93
Input[1]	92.84	84.67	82.53	80.71
GNN[2]	88.59	90.43	70.07	71.80
Attention[3]	95.79	93.15	89.95	85.63
SoftMin[4]	95.14	94.75	95.79	86.61

[1] Replace (7) with $\mathcal{X}_D = W\mathcal{X}_0 \_ \mathcal{X}_D = W\mathcal{X}_0 \_$

[2] Replace (13) with $\mathcal{X}_L = W\mathcal{X}_D \mathcal{X}_L = W\mathcal{X}_D$

[3] Replace (16) with $\mathcal{X}_G = W\mathcal{X}_L' \mathcal{X}_G = W\mathcal{X}_L'$

[4] Replace (18) with (17)

$$e_i = -\sum\nolimits_j \alpha_{ij} \log\left(\alpha_{ij}\right) \tag{36}$$

It can be seen that uniform attention has the highest entropy $e_{\max} = \log(N) \approx 3.66$. Low entropy indicates high attentional focus. Ideally, the attention matrix of the model should be a distribution with low entropy, i.e., a few nodes are much more important than others.

The results are shown in Fig. 10a, b. The average attention score of buses #4, #16 and #29 are high across samples and the attention entropy of all nodes is well below the maximum value.

To verify the effectiveness of the proposed SoftMin-based attention module for global information extraction, the results of using SoftMax are analyzed, as shown in Fig. 10c, d. Although the SoftMax results have similar low entropy in Fig. 10d, it can be seen from the visualized Fig. 10c that the reason for low entropy is that each node unnecessarily pays great attention to its own features, which hinders the extraction of information from other nodes. The impact of the SoftMax function on the performance is analyzed in Table 3.

Table 4 Model performance on the IEEE 39-bus system

Model	ACC	PRE	REC	KAP
SKETCH	98.50	98.98	99.18	94.93
DNN[1]	96.63	97.81	97.18	88.89
ResGAT	97.96	99.00	98.52	93.04
PE-MAGCN	99.28	99.37	98.86	96.23
RGCN	96.99	97.15	96.23	90.11

[1] The scale of the DNN is {546-273-117-39-10-2}

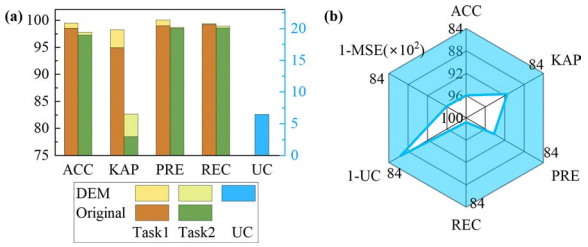

Fig. 11 Further enhancement by DEM on the IEEE 39-bus system

6.6 Ablation experiments

The technical details mentioned above are all ablated and tested separately to verify their effectiveness. Specifically, when ablating a module, a linear alignment of the features is performed to align with the dimensionality of the other parts. The results are also shown in Table 3.

The performance of the model after the ablation of all three modules is degraded to different degrees. This reflects the fact that these modules are essential to the overall performance. Among them, removing disturbance encode gives the largest decrease in precision (about 15%), indicating that the node instability information mainly comes from the disturbed situation. In contrast, removing GNN gives the largest boost to the other metrics, reflecting the importance of considering topological information in the TSA problem. Removing the attention module reveals that performance degradation still occurs and it verifies that information outside the *k*th-neighborhood is equally indispensable.

6.7 Comprehensive performance

Several baseline models are employed for comparison with the SKETCH model on task 1. A DNN is added to the comparison as representative of models that perform well but do not have topology adaptation capabilities. The existing best models, ResGAT [14] and PE-MAGCN [20], have also been added for comparison. Meanwhile, it focuses on solutions that use dynamic information as

input and therefore introduce the best performing model RGCN [13] for comparison. The results are shown in Table 4.

6.8 Decision enhancement strategy

In the decision-making phase, DEM is activated to improve the accuracy of the model by using the consistency of the model on the output of task 1 and task 2 to filter the uncertain samples.

As shown in Fig. 11a, after activating DEM, the performance of the model on both tasks is improved to different degrees. Among them, the accuracy of task 1 is improved from 98.50% to 99.51%, and the effectiveness of the model is further improved. After activating DEM, 6.49% of the samples are subjected to additional time-domain simulations to determine their stability.

Model robustness under larger topological disturbance on account of extreme weather conditions is considered in Fig. 11b. In extreme weather, the system is more likely to suffer from larger topology changes and power distribution variations, and such extreme cases are used to further test the stability of the model. Overall, 7000 samples are generated from the N-3 topology with 20% random variation in power distribution and SKETCH is tested directly without retraining. As shown in Fig. 11, SKETCH illustrates strong adaptability to unknown topologies. In such extreme cases, SKETCH identifies confusing cases, which account for 13.59% of the total samples, and maintains 95.94% accuracy over the remaining cases.

6.9 Test results in a larger system

A larger and more complex IEEE 300-bus system is employed to validate the effectiveness and scalability of SKETCH. The system, with 300 buses, 69 generators, 203 loads, and 411 transmission lines, is comparable in size to the China Southern Grid (500 kV). Overall, 52,210 samples are generated, of which 46,516 are stable and 5,694 are unstable. A retrained model is required, and its optimal settings are listed in Table 5.

The statistical results of the SKETCH output are shown in Fig. 12. Consistent with the results on the IEEE 39-bus test set, the model achieves an overall prediction error of less than 5%, which is in good agreement with the real data distribution. The model performs very well in the data-rich subset, and the error increases in the data-sparse part. A sampling of important data and augmentation of the dataset are still issues that need further consideration.

Table 5 Model performance on the IEEE 300-bus system

Model	ACC	PRE	REC	KAP
SKETCH	98.48	98.73	99.43	94.65
DNN 1	92.18	92.29	71.69	75.37
ResGAT	95.94	97.40	97.81	86.66
PE-MAGCN	98.88	99.14	99.53	96.32
RGCN	96.63	97.81	98.20	88.89

[1] To match the input and output dimensions, the scale of the DNN is modified to {4200-2100-900-300-69-1}

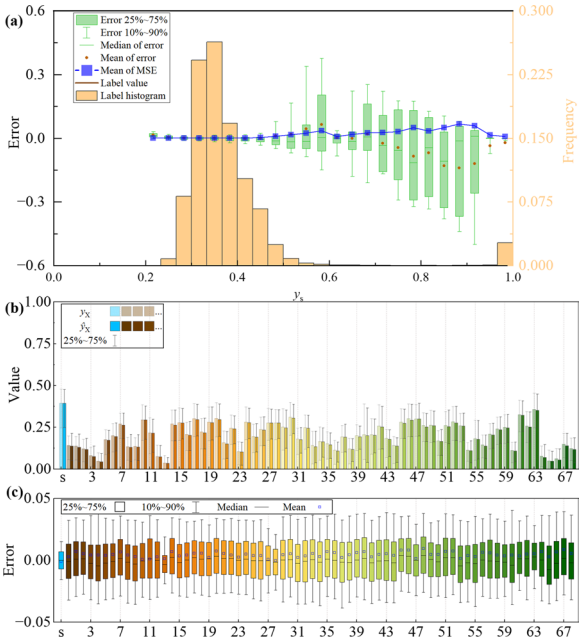

Fig. 12 Statistic results of SKETCH output on the IEEE 300-bus test set. Here are the statistics of the system labels (**a**), the sketch (**b**) and the error of labels (**c**), respectively

Note that without changing the number of model parameters, the model still shows high performance on a much larger system. Tests show that the model maintains its performance in small systems with little degradation, indicating the potential of the model to be applied to large systems.

7 Conclusion

This paper presents the DTSAS problem and a corresponding solution, namely, SKETCH, to confirm the physical mechanism for critical model interpretation and to enrich output information.

The developed scheme uses only static measurement as input and proposes a representation of features at the moment of fault clearance. This is proven to obtain effective enhancement. A module based on the self-attention mechanism is designed to solve the locality problem of the GNN, achieving subgraph equivalence outside the k-order neighborhood. At the same time, the interpretability of the model is enhanced because the model is

structurally designed to conform to the physical mechanism of transient stability.

Test results on the IEEE 39-bus system and IEEE 300-bus system show that SKETCH exhibits better performance than other models and that the performance improvement can be drilled down to black-box models for qualitative interpretation.

Future work will concentrate on exploring the potential of this node-number-independent mechanism to investigate models that can account for variations in the number of nodes in the system.

Acknowledgements
This work is supported by the National Natural Science Foundation of China (52077080).

Authors' information
LIUKAI CHEN received the B.S. degree and master's degree in electrical engineering from the South China University of Technology (SCUT), Guangzhou, China, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include explainable deep learning in power system security and transient stability assessment. LIN GUAN (Member, IEEE) received the B.S. and Ph.D. degrees in electric power engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1990 and 1995, respectively. Her research interests include application of artificial intelligence technology in electrical engineering, power system security and control, and power system planning and reliability. She is currently a Professor with the Electric Power College, South China University of Technology, Guangzhou, China. From 2014 to 2015, she is a Visiting Scholar with Stanford University. She is the author of more than 120 articles and a Principal Investigator of more than 50 projects.

Author contributions
All authors contributed to the study conception and commented on previous versions of the manuscript. LC proposed the methodology for the article, conducted the experiments and is responsible for the writing of the paper. LG guided the direction of the paper and put forward modification suggestions. All authors read and approved the final manuscript.

Funding
This work is supported by the National Natural Science Foundation of China (52077080).

Availability of data and materials
Please contact author for data and material request.

Declarations

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

Received: 6 September 2022 Accepted: 6 January 2023
Published online: 30 January 2023

References

1. Lv, J., Pawlak, M., & Annakkage, U. D. (2017). Prediction of the transient stability boundary based on nonparametric additive modeling. *IEEE Transactions on Power Systems, 32*(6), 4362–4369. https://doi.org/10.1109/TPWRS.2017.2669839
2. Mazhari, S. M., Khorramdel, B., Chung, C. Y., Kamwa, I., & Novosel, D. (2021). A Simulation-based classification approach for online prediction of generator dynamic behavior under multiple large disturbances. *IEEE Transactions on Power Systems, 36*(2), 1217–1228. https://doi.org/10.1109/TPWRS.2020.3021137
3. Zheng, C., Malbasa, V., & Kezunovic, M. (2013). Regression tree for stability margin prediction using synchrophasor measurements. *IEEE Transactions on Power Systems, 28*(2), 1978–1987. https://doi.org/10.1109/TPWRS.2012.2220988
4. Liu, C., Sun, K., Rather, Z. H., Chen, Z., Bak, C. L., Thogersen, P., & Lund, P. (2014). A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees. *IEEE Transactions on Power Systems, 29*(2), 717–730. https://doi.org/10.1109/TPWRS.2013.2283064
5. Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., & Liu, R. (2019). XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. *IEEE Access, 7*, 13149–13158. https://doi.org/10.1109/ACCESS.2019.2893448
6. Bellizio, F., Cremer, J. L., Sun, M., & Strbac, G. (2021). A causality based feature selection approach for data-driven dynamic security assessment. *Electric Power Systems Research, 201*, 107537. https://doi.org/10.1016/j.epsr.2021.107537
7. Chen, Q., & Wang, H. (2021). Time-adaptive transient stability assessment based on gated recurrent unit. *International Journal of Electrical Power and Energy Systems, 133*, 107156. https://doi.org/10.1016/j.ijepes.2021.107156
8. Azman, M. S., Isbeih, Y. J., El Moursi, M. S., Elbassioni, K., Azman, S. K., Isbeih, Y. J., El Moursi, M. S., & Elbassioni, K. (2020). A unified online deep learning prediction model for small signal and transient stability. *IEEE Transactions on Power Systems, 35*(6), 4585–4598. https://doi.org/10.1109/TPWRS.2020.2999102
9. Zhu, L., Hill, D. J., & Lu, C. (2020). Hierarchical deep learning machine for power system online transient stability prediction. *IEEE Transactions on Power Systems, 35*(3), 2399–2411. https://doi.org/10.1109/TPWRS.2019.2957377
10. Gupta, A., Gurrala, G., & Sastry, P. S. (2019). An online power system stability monitoring system using convolutional neural networks. *IEEE Transactions on Power Systems, 34*(2), 864–872. https://doi.org/10.1109/TPWRS.2018.2872505
11. Wang, Z., Zhou, Y., Guo, Q., & Sun, H. (2021). Interpretable neighborhood deep models for online total transfer capability evaluation of power systems. *IEEE Transactions on Power Systems*. https://doi.org/10.1109/TPWRS.2021.3091710
12. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, 13–17-August*, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778
13. Huang, J., Guan, L., Su, Y., Yao, H., Guo, M., & Zhong, Z. (2020). Recurrent graph convolutional network-based multi-task transient stability assessment framework in power system. *IEEE Access, 8*, 93283–93296. https://doi.org/10.1109/ACCESS.2020.2991263
14. Huang, J., Guan, L., Su, Y., Yao, H., Guo, M., & Zhong, Z. (2021). A topology adaptive high-speed transient stability assessment scheme based on multi-graph attention network with residual structure. *International Journal of Electrical Power and Energy System*. https://doi.org/10.1016/j.ijepes.2021.106948
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems, 2017-December*, pp. 5999–6009.
16. Fahim, S. R., Sarker, S. K., Muyeen, S. M., Sheikh, M. R. I., Das, S. K., & Simoes, M. (2021). A robust self-attentive capsule network for fault diagnosis of series-compensated transmission line. *IEEE Transactions on Power Delivery, 36*(6), 3846–3857. https://doi.org/10.1109/TPWRD.2021.3049861
17. Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning, ICML 2015*, Vol. 3, pp. 2048–2057.
18. Chorowski, J. (2015). *Attention-based models for speech recognition*, pp. 577–585.
19. Bahdanau, D. (2015). *Neural machine translation by jointly learning to align and translate.*
20. Huang, J., Guan, L., Su, Y., Yao, H., Guo, M., & Zhong, Z. (2021). System-scale-free transient contingency screening scheme based on steady-state information: A pooling-ensemble multi-graph learning approach. *IEEE Transactions on Power Systems*. https://doi.org/10.1109/TPWRS.2021.3097331

21. Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2019). Graph neural networks: A review of methods and applications. *AI Open*. https://doi.org/10.1016/j.aiopen.2021.01.001
22. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017—Conference Track Proceedings*, pp. 1–14.
23. Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., & Bengio, Y. (2018). Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018—Conference Track Proceedings*, *abs/1710.1*, pp. 1–12. https://doi.org/10.1007/978-3-031-01587-8_7
24. Torch Contributors. (2019). *PyTorch documentation—PyTorch 1.8.1 documentation*. Torch Contributors. https://pytorch.org/docs/stable/index.html, https://pytorch.org/docs/1.8.1/
25. PyG Team. (n.d.). *PyG Documentation—pytorch_geometric 2.0.1 documentation*. Retrieved October 20, 2021, from https://pytorch-geometric.readthedocs.io/en/latest/