Transient Voltage Stability Assessment Method based on gcForest

SUN Zhiyuan\textsuperscript{1,5}, LI Mingpo\textsuperscript{1,5}, ZHANG Jie\textsuperscript{2,3,4,5}, HU Binjiang\textsuperscript{2,3,4,5}, GUO Qi\textsuperscript{2,3,4,5}, ZHU Yihua\textsuperscript{2,3,4,5}

\textsuperscript{1} Guangxi Power Grid Electrical Power Research Institute
\textsuperscript{2} State Key Laboratory of HVDC, Electric Power Research Institute, CSG, China
\textsuperscript{3} Guangdong Provincial Key Laboratory of Intelligent Operation and Control for New Energy Power System
\textsuperscript{4} CSG Key Laboratory of Power System Simulation
\textsuperscript{5} Supported by the Science and Technology Project of Guangxi Power Grid (Contract No(0470002019030101DL00006)).

Email: zhangjie5@csg.cn

Abstract. Aiming at the fact that the traditional method cannot quickly and accurately judge the power system transient voltage stability, and the deep neural network method needs many sample data sets and has a large number of hyperparameters, a power system transient voltage stability assessment method based on deep forest (gcForest) is proposed. gcForest effectively extracts features through ensemble learning, multi-layer feature transformation, and dynamically adjusting the number of model layers. Finally, this method is applied to transient voltage stability assessment of an electrolytic aluminium load-intensive area in Guangxi Power Grid. The results show that the method has high accuracy, the feasibility and effectiveness of this method are verified. This method can assist dispatchers to judge the transient voltage stability of the power grid.

1. Introduction
With the rapid economic development, huge changes have taken place in the composition and characteristics of power grid loads. The heavy loads in industrially-intensive areas and transient voltage stability problems cannot be ignored. The time-domain simulation method has a slow calculation speed, while the accuracy of the direct method applied to complex models in large power grids remains to be studied. In recent years, artificial intelligence technology has developed rapidly, and the research and application of machine learning technology in the field of power grid operation has gradually deepened. Some scholars have introduced machine learning into transient voltage stability assessment \cite{1-7}, and \cite{4-5} uses deep neural networks for assessment. But the algorithm based on deep neural network is based on a large data set, and has a large number of hyper-parameters, and the model structure is very complex; \cite{4-5} proposed a method based on U/P/Q time series trajectory feature learning for transient voltage stability assessment, but the trajectory features are too complex, and the adaptability to the actual power grid is insufficient; \cite{6-7} uses a decision tree algorithm for transient voltage stability assessment, which is a shallow machine learning algorithm, so its ability to fit complex power grids is weak. gcForest (multi-Grained Cascade forest) is a tree learning algorithm based on ensemble learning proposed by Zhou Zhihua \cite{8}. The algorithm is a supervised ensemble machine learning algorithm based
on the tree algorithm inspired by the deep learning theory. It has the advantages of small model complexity and small sample size requirements, and is used in many fields [9,10,11] and achieved good results. To the best of our knowledge, no scholar has applied the gcForest method to the field of power grid operation. In view of the good design and effect of gcForest, this paper proposes a transient voltage stability assessment method based on gcForest, and verifies the effectiveness and feasibility of this method in a regional subsystem of Guangxi Power Grid.

2. gcForest

Given its smashing success, deep neural networks also have the following shortcomings:

(1) Due to the large number of parameters of the deep neural network, a large amount of data is required to obtain a good fitting result. When it comes to supervised learning problems, a lot of manpower is required to complete data labelling;

(2) To achieve good results, deep neural networks need appropriate hyperparameters, such as learning rate, optimizer type, which are usually large in number. Appropriate hyperparameters are very important for the final outcome. Therefore, training neural networks usually requires a lot of energy for researchers to fine-tune hyperparameters;

(3) Deep neural networks lack universal interpretability methods.

At the same time, the deep neural network has the following elements: 1) multi-layer feature calculation; 2) feature conversion within the model; 3) sufficient model complexity.

In view of the above problems of deep neural networks and the characteristics that are worthy of reference, gcForest is cascaded by multiple layers of tree learners, so it is called Cascade Forest, in order to obtain better results through layer-by-layer processing and transformation. And it also uses Multi-Grained Scanning to enrich input features.

![Figure 1. Cascade Forest](image)

2.1. Cascade Forest

The cascade forest structure is shown in Figure 1 [8]. Each level in the cascade structure takes the processing result of the previous level as input, and outputs the converted feature information of this level to the next level. Each level of the cascade can include multiple basic learners. The illustration here shows two different learners in blue and black colours. Here are two types of tree model, completely-random tree forest (CRTF) and random forest (RF). Each CRTF has 100 completely random trees, and the features are randomly selected from the feature set. The node splits until each leaf node contains the same type of samples or less than 10 samples. The same is true for RF, which contains 100 trees. First, the square root of the number of input features is randomly selected and put into the candidate set, and the features of the Gini coefficient are calculated from them to split. Now there are two classes in the problem we are studying, namely instability and stability. Correspondingly, the class vector length is two, and then the input feature and the result are connected to form a new feature that becomes the input of the next layer.

Among them, about the class probability vector generation: the forest will count the proportion of each class of results on its leaf nodes, and then average the statistical results of each category on all trees, so that the forest generates a category vector to reflect class distribution. In terms of transient stability problems, there are two classes, and each forest will output a two-dimensional class vector, so there are
8 = 2×4 augmented feature outputs, and the dimension of the vector passed down is (2×4 + input feature length in the previous layer).

In order to avoid over-fitting affecting the generalization performance, the class vector of each level will be determined by k-fold cross-validation. The data will be equally divided into k parts, each of which k-1 parts will be used for training, and one part will be used as the validation set for k times of training. For each data point, they will be treated as k-1 training data to generate k-1 vector results, and finally the k-1 are averaged and input to the next layer. After entering the new level, the model will be validated on the validation set. If the effect does not improve after the custom round, the training will end. Therefore, the number of cascaded layers is dynamic according to the data set definite. This measure is called early stopping. On the contrary, the complexity of most deep neural networks is already determined at the time of design. This paradigm makes it easy to overfit in small data scenarios. In contrast, gcForest dynamically determines the size of the model, which enables it to adapt to data sets of different sizes.

It can be seen that the design ideas of gcForest have the following highlights:
(1) Compared with the end-to-end training characteristics of deep neural networks, this makes it only receive the supervision signal in the last layer, while gcForest is trained layer by layer, which is equivalent to using the supervision signal in each layer. In this way, gcForest can dynamically control the size of the model according to the size and distribution of the data set.
(2) gcForest is robust to hyperparameters. Because the depth is dynamic, the parameters of the model are determined dynamically. Basically, as long as the gcForest layer and the number of trees within the layer are set, the final effect is almost the same as the optimal one. However, there are many hyperparameters that deep neural networks need to set and adjust. It is difficult to know in advance what structure should be used for a specific problem. Even using neural architecture search and other techniques is time-consuming. As the BP algorithm is adopted, the optimization algorithm also needs to carefully select and adjust its parameters.
(3) With the advantages of ensemble learning, the advantages of small sample learning are outstanding. This part can be seen from the results in Section 3.4.

Figure 2. Multi-Grained Scanning

2.2. Multi-Grained Scanning
Multi-Grained Scanning is similar to the convolution operation in the convolutional neural network. The difference is that its function is data augmentation with a sliding window, and then feed the data into the forest to obtain the classification result, and then concatenate them to obtain new features.

As shown in Figure 2 [8], a window slides a one-dimensional vector with a length of 400 to augment the data, and 301 one-dimensional vectors with a length of 100 are obtained.

And then the obtained features are fed into CRTF and RF respectively. Assuming there are three categories, the result of the sliding window is 301 3-dimensional vectors, which are 1806-dimensional feature vectors after being connected. More specifically, taking the transient voltage stability assessment problem as an example, 301 2-dimensional class vectors are concatenated to obtain a vector with a length of 1204. Multi-Grained Scanning, which uses sliding windows to extract features, is commonly used in fields such as image and voice, and the difference lies in whether it has parameter. This technique is based on the assumption of the local relevance of features, so as long as the feature has a certain local structure, this technique is useful, but when judging transient voltage stability, the features we use are
not local spatial structure, so Multi-Grained Scanning has little effect. This is also reflected in the experimental results in Section 3.4. How to adopt better sampling features needs to be left to future work.

3. Transient voltage stability assessment method based on gcForest

3.1. Input feature selection
Since the rotor kinetic energy, power angle and some other state variables are difficult to measure, and have little effect on transient voltage stability, this article will select the active power P, reactive power Q, and bus voltage amplitude U and phase angle $\theta$ as the sample feature [12]. These steady-state power flow information before the fault has high measurement accuracy and can be directly obtained online in the SCADA system.

3.2. Sample generation method
Since there are very few transient voltage instability samples in the actual power grid, it is difficult to meet the requirements of machine learning. In order to obtain sufficient balanced samples, this paper uses the PSD-BPA developed by the China Electric Power Research Institute to conduct transient time-domain simulations based on the anticipated accident set. First select a set of operating mode data, then randomly adjust the load, and finally perform a fault scan for each load level.

3.3. Transient voltage stability assessment process and model performance evaluation
The transient voltage stability assessment process based on gcForest include offline training and online application. Offline training includes sample generation, model training, and model evaluation. Online applications include input feature data acquisition and online transient voltage stability prediction. In order to evaluate the performance of the model, this paper uses two indicators of accuracy and recall to evaluate the quality of the model.

4. Platforms and experiments

4.1. Platform
In order to improve the speed of sample generation and artificial intelligence model training, this paper designs and builds a set of artificial intelligence software and hardware platforms. The platform consists of 5 Sugon high-performance computing workstations. Each workstation has two Intel E5 2643 v4 processors, 128G memory and 2T intel solid state hard drives. There are 5 NVIDIA Tesla P100 GPU cards in one of the workstations for model training.

4.2. Sample generation process
There are a large number of aluminium loads in the Guangxi power grid, and the transient voltage stability problem is prominent in areas where the power supply is insufficient. There are 149 busbars, 5 generator sets, 6 aluminium loads, 92 220kV lines, and 28 loads in this area. The ordinary load is composed of 50% constant impedance load + 50% induction motor load, and the electrolytic aluminium load is composed of 90% constant current load + 10% induction motor load. The anticipated accident set includes three-phase short-circuit faults at 0% of all 220kV lines, and the fault lasts for 0.12s. The criterion of transient voltage instability refers to the "Guide on Security and Stability Analysis for CSG", specifically, the busbar voltage is continuously lower than 0.75pu for more than 1s, and after the dynamic transient process, the bus voltage in 220kV and above substations shall not be lower than 0.9p.u. [13]. The specific steps of sample generation are as follows:

(1) Set load level and network topology of the power grid.

(2) Select 10 load scaling factors from 0.8 to 1.25 [0.80 0.85 0.90 0.95 1.00 1.05 1.10 1.15 1.20 1.25], and apply the load scaling factors on the load to generate 10 load levels.
(3) The 10 load levels generated in step (2) vary randomly, the vary coefficient range is [0.5, 1.5], and each load level generates 1000 different new load levels. A total of 10,000 load levels are generated.

(4) For each load level, call the PSD-BPA electromechanical transient simulation program to calculate whether the simulation result is transient voltage instability according to the guidelines.

A total of 10,000 samples were generated by the above method, of which 2,238 were transient voltage instability samples and 7,762 were stable samples.

4.3. Experiment and result analysis
In the model training process, after three rounds of training, if there is no further improvement in the result of the validation set, early stopping is adopted. The basic learner of each layer uses XGBoost and RF, each with 100, and the maximum number of layers is set to 100. Due to the cross-validation mechanism, the data are automatically divided into 5 parts within the model, so the validation set is no longer divided manually.

First, the experiment explored the role of Multi-Grained Scanning. The ratio of the test set in the data set was set to 0.7. The multi-granularity scanning used three window sizes of 7, 13, and 21. The results are shown in Table 1. It can be seen that when it comes to the data without local spatial structure, the Multi-Grained Scanning does not show the advantages that it should have. Therefore, the Multi-Grained Scanning is turned off by default later to save training time.

| Type   | Multi-grained scanning turned on | Multi-grained scanning turned off |
|--------|--------------------------------|---------------------------------|
| Accuracy | 0.9864 | 0.9855 |
| Recall  | 0.9742 | 0.9755 |

It can be seen from the ratio of positive and negative samples that the unstable samples are only one third of the stable samples, so we choose to increase the weight of the unstable samples so that the weight of the positive and negative samples is close to 1:1 in front of the model. The ratio of the test set in the data set is set to 0.7. Table 2 shows the results. It can be seen that after the weight of the unstable samples is increased, the recall increases slightly, and the accuracy increases more significantly, approaching 3%.

| Type   | Increased negative sample weight | Normal negative sample weight |
|--------|---------------------------------|------------------------------|
| Accuracy | 0.9855 | 0.9593 |
| Recall  | 0.9755 | 0.9693 |

In order to explore the effect of gcForest on different data ratios, the ratio of the test set in the data set is set to 0.5 to 0.9. The experimental results are shown in Figure 3. It can be seen that from the perspective of data utilization, 0.7 is a appropriate setting, but even when the test set proportion reaches 0.9, the effect is still not very significant. So, we further test the conditions for smaller training set, and set the ratio of the test set in the data set from 0.91 to 0.99. When it reaches 0.91, it means that there are only 900 training samples. By 0.99, there are only 100 training samples. Since cross-validation will be
done inside gcForest, the actual model only uses 80 samples. The result is shown in Figure 4. It can be seen that even with only 100 training samples, gcForest still shows excellent accuracy and recall.

![Figure 3. The effect of different test set ratios](image)

![Figure 4. gcForest accuracy in small dataset scenarios](image)

Furthermore, Table 3 lists the comparison of the effects of other common machine learning algorithms with limited samples. The ratio of the test set in the data set is set to 0.9. Except for gcForest, other models can completely fit the assigned training data. However, gcForest only used four-fifths of the samples due to cross-validation, but it still achieved better results than other models.

| Algorithm Type | Accuracy | Recall |
|----------------|----------|--------|
| gcForest       | 0.9643   | 0.9631 |
| RF             | 0.9649   | 0.9497 |
| SVM            | 0.9023   | 0.8866 |
| KNN            | 0.8785   | 0.8127 |
| XGBoost        | 0.9613   | 0.9443 |

5. Conclusion

This paper introduces gcForest into the transient voltage stability assessment of power system for the first time, and proposes a method of power system transient voltage stability assessment based on gcForest. Considering that traditional time-domain simulation method has long calculation time, the direct method is difficult to adapt to the complex power grid model, and the deep neural network method
requires a large data set, the method based on gcForest has the advantages of appropriate model complexity and less sample demand.

In this paper, based on a regional subsystem of Guangxi Power Grid with more electrolytic aluminium load, 166-dimensional voltage stability features are selected, and a power system transient voltage stability assessment model based on gcForest is constructed. It has achieved better accuracy and recall than other models.

6. References

[1] YANG Weiquan, ZHU Yuanzhen, LIU Yutian. Fast Assessment of Transient Voltage Stability Based on Convolutional Neural Network[J]. Automation of Electric Power Systems, 2019, 43(22):46-52+139.

[2] LAI Wenqing, GONG Qingwu, GAO Chunhui, LIU Huibin, LIU Weiming, WU Liuchuang, LIU Xu. Transient voltage stability evaluation based on feature and convolutional neural network[J]. Engineering Journal of Wuhan University, 2019, 52(09):815-823+835.

[3] Xu Hao, Wang Gang, Wang Longjun. Online Identification of Transient Voltage Stability Based on In-depth Learning[J]. Electrical Automation 2019, 41(01):41-43+67.

[4] ZHU Lipeng, LU Chao, HUANG He, LIU Yingshang. Transient Voltage Stability Assessment Based on Sequential Trajectory Feature Learning[J]. Power System Technology, 2019, 43(06):1922-1931.

[5] ZHU Lipeng, LU Chao, HUANG He, SU Yinseng, WANG Jifeng. Wide-Area Time Series Data Mining Based Transient Voltage Stability Assessment[J]. Power System Technology, 2016, 40(01):180-185.

[6] SHI Yang, LI Yong, CAO Yijia, WANG Weiyu, ZUO Jian. Transient Voltage Stability Assessment Method Based on Morphologically Similar Distance[J]. Proceedings of the CSU-EP-SA, 2020, 32(05):105-112.

[7] ZHU Lipeng, LU Chao, SUN Yuanzhang, HUANG He, SU Yinseng, LI Zhihuan. Data Mining Based Regional Transient Voltage Stability Assessment[J]. Power System Technology, 2015, 39(04):1026-1032.

[8] Zhou Zhi Hua, Feng J. Deep forest: Towards an alternative to deep neural networks[C]. IJCAI, 2017.

[9] LIN Yang, DONG Bao-liang, LIU Ze-ping. Method of military target recognition based on CGAN and GcForest[J]. Information Technology, 2020, 44(03):134-138.

[10] WANG Lun, LI Lu. Multi-factor Quantitative Stock Selection Strategy Based on gcForest[J]. Computer Engineering and Applications, 2020, 56(15):86-91.

[11] Zhao Pu. Research and Application of Imbalanced Classification Technology Based on gcForest[D]. Chongqing University of Posts and Telecommunications, 2019.

[12] Wang Tongwen, Guan Lin, Zhang Yao. A Survey on Application of Artificial Intelligence Technology in Power System Stability Assessment[J]. Power System Technology, 2009, 33(12):60-65+71.

[13] 2009 Q/CSG. Guide on Security and Stability Analysis for CSG[S].