IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2022, accepted 15 July 2022, date of publication 25 July 2022, date of current version 3 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193506

RESEARCH ARTICLE

A Data-Driven Scheme Based on Sparse Projection Oblique Randomer Forests for Real-Time Dynamic Security Assessment

YANFENG LIN[1] **AND XINYAO WANG**[2,3]

[1]College of International Communications, China Three Gorges University, Yichang 443002, China
[2]College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China
[3]Hubei Provincial Collaborative Innovation Center for New Energy Microgrid, China Three Gorges University, Yichang 443002, China

Corresponding author: Yanfeng Lin (lscsxdx@163.com)

This work was supported in part by the Open Foundation of Yichang Key Laboratory of Intelligent Operation and Security Defense of Power System under Grant 2020DLXY06, in part by the Open Foundation of Hubei Provincial Key Laboratory for Operation and Control of Cascaded Hydropower Station under Grant 2019KJX11, and in part by the National Natural Science Foundation of China under Grant 51607104 and Grant 51807109.

ABSTRACT With the wide interconnection of power systems and extensive application of phasor measurement units (PMUs), the secure operation of power systems is facing considerable challenges. To satisfy the demand of online dynamic security assessment (DSA) for modern power systems, a data-driven scheme based on sparse projection oblique randomer forests (SPORF) is proposed, which includes offline training, periodic update and online assessment. In the first stage, an improved adaptive synthetic sampling (ADASYN) method is developed to mitigate the class imbalance problem for the data-driven DSA approach. Then, the SPORF-based DSA model is trained using crucial features with low redundancy selected by a feature selection procedure based on the minimal-redundancy-maximal-relevance (MRMR) criterion. In the second stage, the periodic update of the DSA model for unseen system topologies is executed to enhance the robustness of the model. In the third stage, the trained model can provide the DSA result immediately when the real-time operation information of a system is received. The satisfactory performance of the proposed scheme is demonstrated through a series of tests and the comparisons on a 23-bus system and a practical 1648-bus system.

INDEX TERMS Dynamic security assessment, data-driven, data oversampling, feature selection, sparse projection oblique randomer forests.

I. INTRODUCTION

Dynamic security assessment (DSA) is of great importance in the power system planning and operation [1]. With the construction and development of smart grids, traditional power systems are undergoing substantial changes. The gradual expansion of the system scale and the increasing insertion of new electrical equipment may bring unprecedented risks to power systems [2]–[4]. To grasp the system operation status in a timely manner and reduce the economic loss caused by unforeseen electric power accidents, the achievement of real-time and accurate DSA for power systems is essential [5].

Traditional DSA for power systems commonly uses analytical physics methods, such as time-domain simulation (TDS) and direct method [6], [7]. The TDS method determines whether the system is stable by solving a set of high-dimensional nonlinear differential difference algebraic equations (DDAEs) that reflect the electromechanical transient process [8]. The direct method judges the transient stability of a system by constructing the transient energy function based on the Lyapunov stability principle [9]. However, due to the high computational complexity and time consumption, these methods may not satisfy the requirements of real-time DSA for modern large-scale power systems [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Akshay Kumar Saha.

In recent years, phasor measurement units (PMUs) have been developed as advanced measurement tools in power systems to collect system operating information simultaneously from different locations [11]. With the widespread adoption of PMUs, data-driven tools have been employed to assess system stability online. Some DSA methods have been proposed using data-driven tools based on the following idea. First, data-driven tools are trained to construct the mapping relationships between the operating variables and the corresponding security index. Second, online DSA results can be provided by the trained DSA model when the PMU measurements of pre-fault attributes are received [1]. Existing classic data-driven methods, such as decision tree (DT) [1], [12], artificial neural network (ANN) [11], [13] and support vector machine (SVM) [14], [15], have been proven effective for DSA. Additionally, an ensemble learning method that constructs a series of extreme learning machines (ELMs) is proposed in [16] to generalize the randomness of ELMs and provide more credible DSA results. In [17], a novel fast transient stability batch assessment framework using cascaded convolutional neural networks (CNNs) is constructed to terminate the TDS early to shorten the simulation time under the premise of accurate assessment. In [18], an online DSA approach based on random forest (RF) is proposed to accurately assess the unseen system operating conditions and indicate the confidence level of the security status one minute ahead of real time by means of an outlier identification method.

Although relevant studies have demonstrated the efficiency of these data-driven methods in the DSA of power systems, some issues remain in the online application to large-scale power systems for these methods. First, the captured security rules may sometimes lack credibility based on massive system operating variables that appear to be largely irrelevant to fault analysis [19]. Second, when applied to large-scale power systems, some methods may be limited by difficult and complicated parameter tuning, which will consume considerable computing time [20]. Third, with the increasing stability of modern power systems, more secure samples than insecure samples are obtained for DSA research, and the resulting class imbalance problem may bring unpredictable impacts to data-driven DSA methods, which exposes the secure operation of the system to great risks [21].

To address the above-mentioned issues, this paper proposes a novel data-driven scheme to achieve real-time and effective DSA for practical large power systems. Notably, stability control is beyond the scope of this paper. In this scheme, an improved adaptive synthetic sampling (ADASYN) method is employed to alleviate the class imbalance problem, a feature selection procedure based on the minimal-redundancy-maximal-relevance (MRMR) criterion [22] is introduced to select crucial features with low redundancy, and a DSA model based on sparse projection oblique randomer forests (SPORF) [23] is trained to extract the mapping relationships between the power system operating variables and the corresponding class label. The performance and effectiveness of the proposed scheme are validated on a 23-bus system and a practical 1648-bus system.

The main contributions in this paper can be summarized as follows:

1) A powerful data-driven tool called SPORF is utilized to construct an efficient DSA scheme that can provide accurate assessment results with a short computing time and low computational burden in online applications.
2) An improved ADASYN method is developed to effectively relieve the class imbalance problem, and a feature selection procedure based on the MRMR criterion is adopted to select the crucial features with low redundancy and improve the computational efficiency.
3) In view of the generalization ability and reliability of the DSA model in practical applications, a model update mechanism that responds to system network topology changes is proposed to ensure the effectiveness of the DSA model in online application.
4) The high computing speed of the proposed scheme is verified based on tests with a 23-bus system and a practical 1648-bus system. Moreover, the impact of the training dataset size is analyzed, and the good robustness to missing data is demonstrated.

The remainder of this paper is organized as follows. The problem statement and supporting methods are introduced in Section II. The technical details of the proposed online DSA scheme are described in Section III. An illustrative example using the 23-bus system is presented in Section IV. The application of the proposed scheme to a practical 1648-bus system is discussed in Section V. Finally, the conclusion is provided in Section VI.

II. PROBLEM STATEMENT AND SUPPORTING METHODS

A. RULE FOR CLASSIFICATION

The research on pre-fault DSA of power systems aims to assess the operation status of a power system with expected but not yet occurring faults based on the pre-fault steady-state operating variables, such as bus voltage amplitude, bus voltage phase angle, power generation and load [7], [24]. In general, pre-fault DSA can be defined as a two-class classification problem: secure and insecure [17].

The transient stability margin (TSM) is typically employed to describe the operation status of the system because it can represent the transient stability level of a power system [9]. Meanwhile, the critical clearing time (CCT), which can reflect the anti-interference ability of the system [24], [25], is the basis for TSM. The CCTs in different fault locations of a power system can be acquired through a series of dynamic simulations with different clearing times. In these simulations, the stability status of the power system can be judged by observing the rotor angle difference of any two generators. If the difference exceeds 360 degrees, the system is regarded as unstable. Subsequently, the TSM, which can be constructed by employing the CCT and the actual clearing time (ACT), is defined in formula (1):

$$\mathrm{TSM} = \frac{\mathrm{CCT} - \mathrm{ACT}}{\mathrm{CCT} + \mathrm{ACT}} \tag{1}$$

Theoretically, TSM is a continuous index that ranges from -1 to 1. Use a binary pair 1 and 0 to represent the secure and insecure status of a system operating point under a fault. Then a typical rule can be adopted to classify an unseen operating point, which is given by formula (2):

$$\begin{cases} \mathrm{TSM} \geq u \rightarrow \text{label} = 1 \text{ (secure)} \\ \mathrm{TSM} < u \rightarrow \text{label} = 0 \text{ (insecure)} \end{cases} \tag{2}$$

where u is a user-defined threshold. The value of u can be adjusted according to different requirements for the degree of security. In particular, when the system operators desire more conservative assessment results, the value of u can be increased.

B. IMPROVED ADASYN

ADASYN is an adaptive synthetic oversampling method used to address the class imbalance problem. The number of required synthetic samples generated by ADASYN for each minority class sample can be adaptively determined using the density distribution of minority class samples as a criterion [26]. However, due to the existence of noise, the new synthetic samples may not provide useful information, which will increase the possibility of overfitting for data-driven tools. Therefore, this paper develops an improved ADASYN method that can alleviate the negative impact of noise while synthesizing minority class samples. The main steps of the improved ADASYN method are summarized below.

1) Let p and q denote the number of minority class samples and the number of majority class samples in a dataset, respectively. Then, the total number of synthetic minority class samples M can be calculated by formula (3).

$$M = (q - p)\,\beta \tag{3}$$

where β is a user-defined parameter ranging from 0 to 1 for adjusting the number of synthetic minority class samples. $\beta = 1$ means that the training dataset achieves absolute class balance after data oversampling. Previous studies recommend a value of β between 0.15 and 0.30 [2].

2) Find the k nearest neighbors based on Euclidean distance for each minority class sample. The number of majority class samples among the k nearest neighbors is denoted as k_i'. If $k_i' = k$, all the k nearest neighbors of the minority class sample are majority class samples; therefore, the minority class sample is considered to be noise and is not used for oversampling. If $k_i' \neq k$, the minority class sample is considered as a useful minority class sample, and the following oversampling steps are executed.

3) Let $\boldsymbol{t}_i\,(i = 1, 2, \cdots h)$ represent each useful minority class sample. Then, the density distribution $\bar{r}_i$ of each useful minority class sample $\boldsymbol{t}_i$ can be calculated according to formula (4).

$$\begin{cases} r_i = \dfrac{k_i'}{k} \\ \bar{r}_i = \dfrac{r_i}{\sum\limits_{i=1}^{h} r_i} \end{cases} \tag{4}$$

4) Calculate the number of required synthetic samples d_i for each useful minority class sample $\boldsymbol{t}_i$ according to formula (5).

$$d_i = \bar{r}_i M \tag{5}$$

5) Randomly select a minority class sample t_k from the k nearest neighbors of each useful minority class sample t_i; then, the new synthetic sample $\boldsymbol{t}_{new}$ is constructed according to formula (6).

$$\boldsymbol{t}_{new} = \boldsymbol{t}_i + \lambda\,(\boldsymbol{t}_k - \boldsymbol{t}_i) \tag{6}$$

where λ represents a random number ranging from 0 to 1, $(\boldsymbol{t}_k - \boldsymbol{t}_i)$ represents the difference vector in n dimensional spaces, and n is the number of features in each sample.

6) Repeat step 5 until d_i samples are generated for each useful minority class sample $\boldsymbol{t}_i$.

C. MRMR CRITERION

In this paper, the MRMR criterion, which is based on mutual information (MI), is used to measure the quality of features [22]. The goal is to select the crucial features that have the minimum redundancy among features and the maximum relevance with the class. The MRMR criterion is described in detail below.

1) MAX-RELEVANCE

Max-relevance, which is used to select the features with the highest correlation with the class, can be calculated by formula (7).

$$\max D\,(\boldsymbol{S}, c)\,, D = \frac{1}{|\boldsymbol{S}|} \sum_{f_i \in \boldsymbol{S}} I\,(f_i, c) \tag{7}$$

where $\boldsymbol{S}$ is a subset consisting of the selected features, $|\boldsymbol{S}|$ is the size of the feature set, and $I\,(f_i, c)$ is the MI between feature f_i and class c. The stronger the relevance between the features and the class are, the higher the value of D will be.

2) MIN-REDUNDANCY

Min-redundancy, which is used to reduce the redundancy among the selected features, can be computed by formula (8).

$$\min R\,(\boldsymbol{S})\,, R = \frac{1}{|\boldsymbol{S}|^2} \sum_{f_i, f_j \in \boldsymbol{S}} I\left(f_i, f_j\right) \tag{8}$$

where $I\left(f_i, f_j\right)$ is the MI between the features f_i and f_j. The weaker the redundancy among the features are, the smaller the value of R will be.

3) MIN-REDUNDANCY AND MAX-RELEVANCE CRITERION

In practice, the max-relevance and the min-redundancy cannot always be achieved simultaneously. Therefore, the formulas (7) and (8) can be combined into a single MRMR criterion as formula (9).

$$\max \Phi\left(D, R\right), \Phi = D - \frac{R}{\min(H(f_i), H(f_j))} \tag{9}$$

where $H(f_i)$ and $H(f_j)$ are the entropy of the ith and jth features, respectively. The feature selection algorithm based on the MRMR criterion is implemented as an incremental search procedure [22], and the detail steps of the feature selection are summarized in part A of Section III.

D. SPORF

SPORF is a supervised learning algorithm that combines sparse random projections with the RF algorithm [23]. In contrast to RF constructed by ensembles of axis-aligned decision trees, SPORF is constructed based on oblique decision trees. As shown in Figure 1 (a), the axis-aligned split separates data by creating an orthogonal decision boundary [18]; however, the data are not linearly separable in most cases, which may lead to inaccurate classification results. As shown in Figure 1 (b), the oblique split divides the data by means of oriented hyperplanes, which can provide more accurate results and shallower trees [27]. Compared to RF, SPORF significantly improves the classification accuracy and reduces the consumption of computing resources.

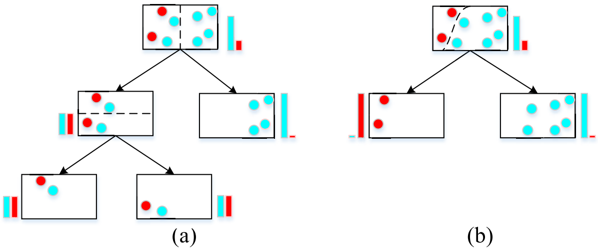

FIGURE 1. Different split types of decision trees (the bars represent the information gain). (a) Axis-aligned split, tree with depth 2. (b) Oblique split, tree with depth 1.

The main process of the SPORF algorithm consists of three steps: 1) Randomly project the input feature matrix into a new feature space; 2) Train the oblique decision trees by searching for the best split in the high-dimensional projected subspace. The best node split is usually determined by minimizing the Gini impurity or entropy. Moreover, an increment function is employed to update the counts of the left and right partitions when assigning the child nodes of the tree. 3) Adopt a majority voting mechanism to obtain the final classification result.

In this paper, SPORF is introduced into DSA to rapidly and accurately assess the security status of a power system. By constructing mapping relationships between the crucial features and the corresponding class label in the trained SPORF classifiers, the current security status of the power system can be immediately assessed when the real-time PMU data are obtained.

III. PROBLEM STATEMENT AND SUPPORTING METHODS

The proposed data-driven scheme based on SPORF is shown in Figure 2. The power system operating variables are acquired via the collection of PMUs and simulation data, and an initial knowledge base containing a large amount of system operation information and the corresponding class label is generated. Data processing, including data oversampling and feature selection, is designed to address the class imbalance problem and select the crucial features. Then, the SPORF-based DSA model is trained by employing the crucial features and the corresponding class label. Considering that the system topology often varies due to possible operational requirements, a model update mechanism is devised to accommodate to unseen system topologies and to improve the generalization of the DSA model. In practical applications, online DSA is executed based on the trained DSA model when the real-time operation information of the power system is collected by PMUs. Overall, the proposed scheme consists of three stages, namely, offline training, periodic update and online assessment, which are described in detail below.

A. OFFLINE TRAINING STAGE

1) INITIAL KNOWLEDGE BASE PREPARATION

The efficiency of the proposed scheme relies strongly on whether the mapping relationships between the operating variables and the corresponding class label are accurate. To construct accurate mapping relationships, an abundant, convincing and reasonable knowledge base must be established. The historical operation data of power systems collected by PMUs contain limited information since some potential system operation behaviors may not be recorded. Therefore, to record a greater variety of system operation behavior and establish an abundant initial knowledge base, several approaches are adopted to generate more operating points. One is linear interpolation based on the two close historical operating points [1]. Another is to add reasonable fluctuations to practical operating points [28]. Subsequently, more operating points that are consistent with the actual operation of the power system can be obtained. Then, to calculate the CCTs related to the obtained system operating points, a series of dynamics simulations by PSS/E are executed in consideration of multiple expected faults. Finally, for a fault location, according to the classification rule mentioned in Section II, the class label for each operating point is obtained, and an initial knowledge base containing massive pre-fault system operating variables and the corresponding class label is established.

Specifically, the initial knowledge base can be represented in matrix form as $T : [X_{N\times M}, Y]$, where $X_{N\times M}$ is the input feature matrix containing the system operating variables (e.g., bus voltage, branch power flows, loads and generations), N is the number of samples, M is the number of

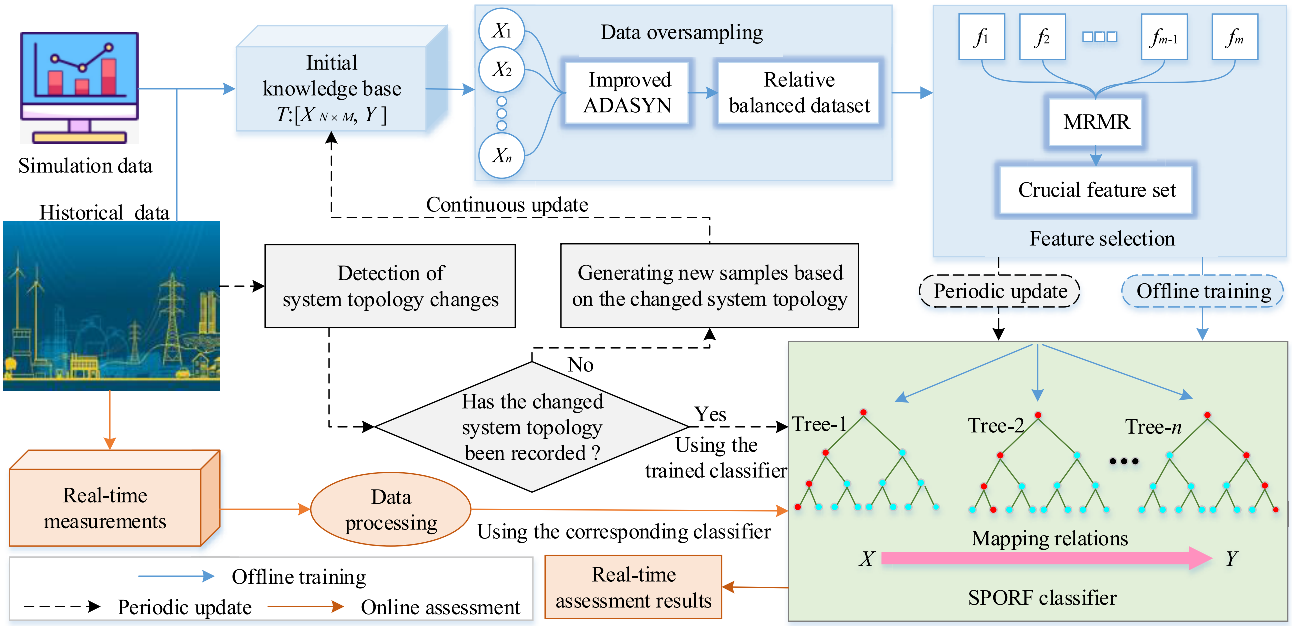

FIGURE 2. Proposed scheme for DSA.

features for each sample, and Y is the corresponding class label vector. The number of samples required for satisfactory performance on a specific power system can be determined experimentally.

2) DATA OVERSAMPLING

In practice, modern power system operation remains secure after most contingencies and becomes insecure in only a few cases [29]. Therefore, secure samples obtained for DSA studies usually constitute a much larger portion than insecure samples, which will lead to a serious class imbalance problem. If not addressed in an appropriate manner, the class imbalance problem may decrease the classification accuracy and generalization ability of the DSA model. Typically, misclassification in DSA is often related to the class imbalance problem [2].

In this paper, an improved ADASYN method is designed to process the initial knowledge base, and the class imbalance problem is effectively mitigated by adaptively synthesizing insecure samples. Consequently, a more abundant and balanced dataset is obtained. A schematic diagram of data oversampling is presented in Figure 3. Variables A and B are two pre-fault steady-state operating variables. After data oversampling, a DSA model trained on a relatively balanced dataset can more accurately distinguish insecure samples.

3) FEATURE SELECTION

The structure and operation mode of modern power systems tends to be complicated, which has resulted in the explosive growth of system operating features. For data-driven DSA of power systems, excessive input features not only lead to meaningless expansion of the input space and a waste of computing resources but also influence the classification accuracy of DSA model [10]. Feature selection is essential to reduce the dimensionality of the input features and improve the classification performance of the DSA model. In this paper, a feature selection procedure based on the MRMR criterion is adopted to select the crucial features with the minimum redundancy among features and the maximum relevance with the class.

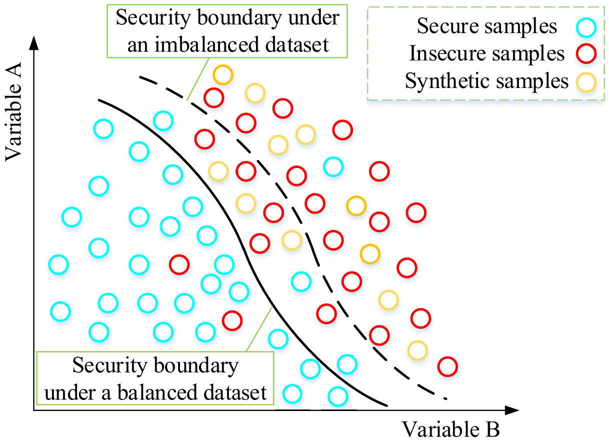
Security boundary under an imbalanced dataset
Secure samples ○
Insecure samples ○
Synthetic samples ○
Variable A
Variable B
Security boundary under a balanced dataset

FIGURE 3. Schematic diagram of data oversampling.

As shown in Figure 4, the crucial features are successively added into a null set S. After feature selection, a crucial feature subset with $m(m > 2)$ features is acquired. According to the demand in practical applications, the value of m can be tuned for different systems. The feature selection procedure is summarized in three steps.

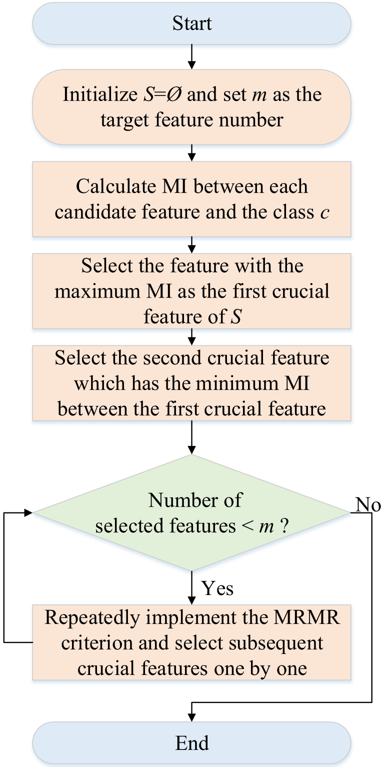
Start
Initialize S=Ø and set m as the target feature number
Calculate MI between each candidate feature and the class c
Select the feature with the maximum MI as the first crucial feature of S
Select the second crucial feature which has the minimum MI between the first crucial feature
Number of selected features < m ?
No
Yes
Repeatedly implement the MRMR criterion and select subsequent crucial features one by one
End

FIGURE 4. **Flow chart of the proposed feature selection procedure.**

1) The feature with the maximum MI is selected as the first crucial feature of S after calculating the MI between each candidate feature and the class in the knowledge base.
2) Then, the MI between the first crucial feature and other candidate features are calculated, and the feature with the minimum MI is chosen as the second crucial feature of S.
3) Subsequently, the max-relevance and the min-redundancy in S can be calculated by formulas (7) and (8). And the MRMR criterion, which is expressed as formula (9), can be implemented. The subsequent features are selected one by one by utilizing the MRMR criterion repeatedly till m crucial features are selected.

4) TRAINING OF THE SPORF CLASSIFIERS

The mapping relationships are constructed using the crucial feature subset and the corresponding class label as the input and output of SPORF. In general, the credible system topology list can be acquired from utility companies; thus, multiple training sample sets corresponding to such topologies can be generated to train a series of candidate SPORF classifiers offline to address various system topologies [28]. Finally, the various system topologies and the corresponding trained SPORF classifiers are formed into multiple matching pairs for invocation.

B. PERIODIC UPDATE STAGE

In practice, the system topology often changes due to possible operational requirements (e.g., scheduled maintenance, economic dispatch and optimal power flow) [30]. A knowledge base generated offline hardly captures all potential operation behaviors of a time-varying power system, and the offline trained classifiers may not provide accurate and reliable assessment results for unseen system topologies. Therefore, a model update mechanism is designed to improve the generalization ability of the proposed scheme. Generally, the update stage runs continuously in parallel with the online assessment. The model update mechanism shown in Figure 2 is described in detail below.

When the system topology changes in the application of the proposed scheme, if the changed topology has been recorded in the offline stage, the corresponding trained candidate SPORF classifier will immediately replace the currently used SPORF classifier to achieve DSA. If the changed topology has not previously been recorded, new samples based on the new system topology will be generated to train a new SPORF classifier, and the new matching pair will be acquired. By periodically executing the model update mechanism, the probability of encountering unseen system topologies will decrease, and seamless online assessment can be achieved in the future.

C. ONLINE ASSESSMENT STAGE

With the rapid deployment and development of PMUs in power systems, data acquisition of power systems has become much faster. When the real-time PMU measurements of the selected features are sent to the corresponding previously trained SPORF classifier, the online DSA result can be provided immediately. If the current status is determined as secure, the system will continue to be monitored; otherwise, risk warning signals will be sent to the system operators.

IV. ILLUSTRATIVE EXAMPLE

The 23-bus system provided by PSS/E is used as an illustrative system to evaluate the performance of the proposed scheme. A diagram of the 23-bus system is shown in Figure 5. The test system consists of 23 buses, 6 generators and 10 transformers [31], and the performance tests are conducted on an Intel Core i7 3.40-GHz CPU with 8 GB of RAM.

A. KNOWLEDGE BASE GENERATION

To capture more potential system operation behavior and establish an abundant and reasonable knowledge base, the generators/loads are initialized by randomly varying their original distributions within a range of 80-120%, and the load level varies from 70% to 130% of the initial value. Then, the most serious three-phase faults are considered in this paper, and the fault locations include buses and the middle points of the transmission lines [32], [33]. In this work, PSS/E and Python are used to automatically conduct power flow

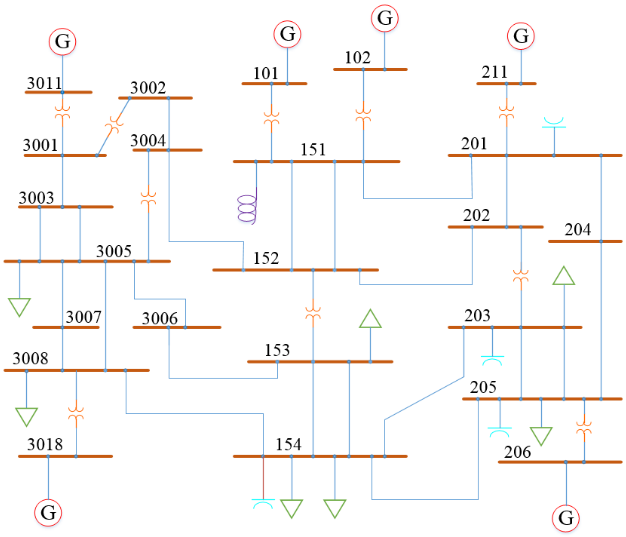

FIGURE 5. Diagram of the 23-bus system provided by PSS/E.

analysis and dynamic simulations. Notably, the power flows of all the generated operating points are solved with limit checking, and the conditions with any overloading or voltage limit violation are not included in the knowledge base [30]. Thus, 403 initial variables with the corresponding class label are obtained for each operating point. The variables include bus voltage amplitude, bus voltage phase angle, active power and reactive power of loads, active power and reactive power of generators, reactive power of shunts, active power and reactive power from bus i to j, and active power loss and reactive power loss from bus i to j. For the data oversampling, β is set to the recommended value of 0.30, and k is set to 5. After data oversampling, a total of 6925 samples are obtained. Then, the value of m is set to 30, and 30 variables are selected as the crucial features using the feature selection procedure. Eventually, for each fault location, a knowledge base consisting of the crucial features and the corresponding class label is established.

B. DSA TEST

The 5-fold cross-validation is employed to comprehensively test the performance of the DSA model. The samples after feature selection are randomly divided into 5 mutually exclusive subsets with the same size. Cross-validation is repeated 5 times, with each subset alternately acting as the testing set and the remaining as the training set [16].

1) EVALUATION METRICS

In this paper, the performance of the DSA model is evaluated using statistical metrics from the perspective of system status classification. As shown in Table 1, according to the predicted and actual classes, the classification results can be classified into four categories. Then, the classification accuracy and $F_{1-score}$ can be defined as follows.

TABLE 1. Confusion matrix.

Predicted class Actual class	Secure	Insecure
Secure	TP	FN
Insecure	FP	TN

(1) The classification accuracy, which denotes the ratio of correct classifications, is given by formula (10). A higher accuracy indicates that the DSA model has better classification performance.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{10}$$

(2) The $F_{1-score}$, given by formula (11), is defined as the harmonic mean of precision and recall. Precision is the ratio of correctly predicted secure samples to all the predicted secure samples. Recall is the ratio of correctly predicted secure samples to all the actual secure samples. The value of $F_{1-score}$ falls in (0, 1), and the classification performance of the DSA model is better when $F_{1-score}$ is closer to 1 [19].

$$\begin{cases} \text{precision} = \dfrac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \\ F_{1-score} = \dfrac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \end{cases} \tag{11}$$

2) TEST RESULTS

The proposed DSA model has been tested on the 23-bus system, and the test results of the classification accuracy and $F_{1-score}$ are shown in Table 2, which indicate that the DSA model has satisfactory classification ability on the 23-bus system.

TABLE 2. Test results of the DSA model for the two test systems.

Test System	Accuracy	$F_{1-score}$
23-bus	98.78%	98.67%
1648-bus	97.95%	96.51%

V. APPLICATION TO A LARGER SYSTEM

To further validate the performance of the proposed scheme, the DSA model is tested on a practical 1648-bus system consisting of 313 generators, 182 shunts and 2294 transmission lines [31]. The same method for generating the knowledge base as used in the 23-bus system is adopted. Tests and analyses are performed in the same simulation environment as that used in Section IV. For the data oversampling, the values of β and k are set to 0.30 and 5, respectively. A total of 13497 samples with 37439 variables are generated for the 1648-bus system. Then, the value of m is set to 800, and 800 crucial features are selected. As illustrated in Table 2, the DSA model exhibits excellent classification performance for the 1648-bus system.

A. SIGNIFICANCE OF DATA OVERSAMPLING

Table 3 shows the test results of classification accuracy and $F_{1-score}$ for the 23-bus system and the 1648-bus system without data oversampling when using the same testing set. According to the comparison results of the Table 2 and 3, the DSA model trained with the dataset after data oversampling shows better classification performance than the model trained with the original data. In fact, since the improved ADASYN method handles the class imbalance problem by interpolating data points in the feature space in a more general and adaptive manner, a satisfactory generalization performance can be achieved. Hence, it is significant to alleviate the class imbalance problem for improving the performance of the DSA model.

TABLE 3. Test results without data oversampling.

Test System	Accuracy	$F_{1-score}$
23-bus	94.01%	96.15%
1648-bus	91.26%	92.89%

In recent years, several attempts have been made to address the class imbalance problem in data-driven DSA, such as random oversampling (ROS) [34], the synthetic minority oversampling technique (SMOTE) [35], Borderline-SMOTE (B-SMOTE) [36] and ADASYN [26]. ROS simply copies insecure samples in the initial knowledge base, which easily results in overfitting of DSA model. SMOTE blindly synthesizes new samples for each insecure sample via linear interpolation without further information about neighboring insecure samples, which results in class overlapping. B-SMOTE only pays attention to borderline insecure samples and ignores the information of the other insecure samples, which deteriorates the generalization capacity of the DSA model. To avoid these issues, ADASYN adaptively determines the synthetic proportion of each minority class sample according to the importance of each minority class sample. However, the quality of synthetic samples is easily affected by noise. Hence, this paper proposes an improved ADASYN, which not only effectively alleviates the class imbalance problem but also significantly reduces the impact of noise on synthetic samples. Furthermore, the improved ADASYN improves the classification performance of the DSA model; therefore, the advantage of the improved ADASYN method is more obvious in dealing with the class imbalance problem.

B. PERFORMANCE COMPARISON WITH OTHER CLASSIFIERS

To compare the proposed SPORF-based DSA model against conventional classifiers, five other methods, namely, logistic regression (LR), ANN, SVM, DT and RF, are tested using the same input samples. The comparison results in terms of the classification accuracy and $F_{1-score}$ for the two systems are shown in Table 4. It can be observed that the SPORF-based DSA model achieves better classification performance for DSA. In addition, the SPORF-based classification model has special advantages over conventional classifiers, which are summarized as follows.

TABLE 4. Performance comparison of different classifiers.

Classifiers	23-bus System		1648-bus System	
	Accuracy	$F_{1-score}$	Accuracy	$F_{1-score}$
LR	96.35%	96.01%	95.49%	93.97%
ANN	97.42%	97.13%	96.74%	94.61%
SVM	97.64%	97.40%	96.92%	94.85%
DT	95.17%	94.96%	94.25%	92.38%
RF	97.81%	97.69%	97.16%	95.03%
SPORF	98.78%	98.67%	97.95%	96.51%

1) For LR, the dimension of the input space greatly affects the classification performance. In particular, the accuracy of LR may be unacceptable when a large number of input variables are considered [37]. Compared to LR, SPORF can train vast amounts of variables efficiently since it adopts multiple parallel trees to accommodate high-dimensional datasets [23].

2) For ANN and SVM, the high computational cost is an obvious problem when these classifiers are applied to actual large power systems. These two classifiers consume massive quantities of machine memories and computing time due to the complex calculation process and slow learning speed [38]. Compared to ANN and SVM, SPORF has good robustness derived from the default parameters instituted by experimental rules, and it adopts the sparse projections method to reduce the memory consumption, which can significantly improve the computing speed [23].

3) For ANN and DT, overfitting is a pervasive problem in model learning [12], [13]. DT easily fits noise data and unrepresentative data, and the growth of trees is not restricted reasonably in the building process. ANN is prone to overfitting due to complex model and superabundant parameters. Compared to ANN and DT, SPORF effectively reduces the possibility of overfitting via two mechanisms: 1) constructing each tree on random bootstrap samples of the original data; 2) making each split of each node with a random subset of the features.

4) For RF, the process of tree splitting is performed along the coordinate axes of the feature space, so the strength and diversity of trees are limited. In addition, axis-aligned splits may result in suboptimal trees [39]. Compared to RF, SPORF combines the expressive capacity of oblique trees with the benefits of axis-aligned trees and builds an ensemble of oblique, interpretable, and scalable trees [23].

C. COMPUTING TIME

The computing time is often as important as the classification accuracy in data-driven DSA, especially for large datasets. In practice, to assess the system security status at

TABLE 5. Computing time of the DSA model for the two test systems.

Test System	Training Time	Testing Time
23-bus	36.84 s (5540 samples)	1.47 s (1385 samples)
1648-bus	127.28 s (10798 samples)	3.26 s (2699 samples)

each snapshot, the processing time of PMU data should be less than 0.033 s [30]. Therefore, how to make full use of rapidly updated PMU data is critical to achieve real-time DSA. Table 5 summarizes the training and testing times for the 23-bus system and the 1648-bus system. As shown in Table 5, a new operating point can be assessed in less than 0.002 s for both test systems.

D. IMPACT OF TRAINING SET SIZE

The ever-increasing size and complexity of power systems make DSA extremely challenging [12]. In this section, the effects of different training set sizes on the classification accuracy are explored. Different proportions (30%, 50%, 70% and 90%) of the original training set are used to train the DSA model in each test, and the test results are shown in Figure 6. It can be observed that 50% of the original training set is sufficient for training a model with a classification accuracy greater than 95%. Moreover, the larger the training set is, the higher the classification accuracy will be. The system operator can choose an appropriate training set size according to the actual demand.

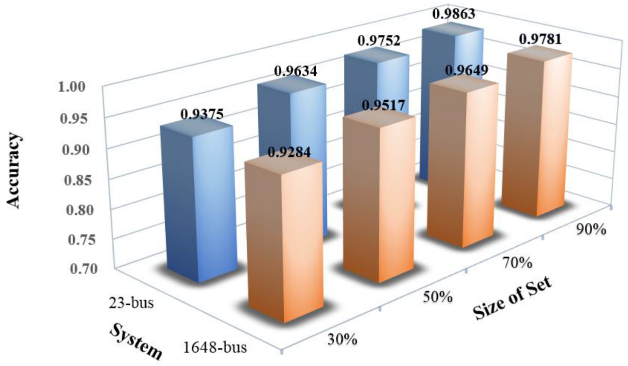

FIGURE 6. Classification accuracy for different training set sizes.

E. TESTS FOR UNSEEN NETWORK TOPOLOGIES

In practical applications, the accuracy of a DSA model may be affected by the change of network topology [30]. Therefore, a reliable online DSA model should have the capacity to adapt to previously unseen network topologies. To demonstrate the wide adaptability of the DSA model to the change of network topology, different topologies of the 23-bus system and 1648-bus system are considered. All the tests are performed to assess the unseen network topologies using the model trained on the system with the original topology. The corresponding test results are shown in Table 6.

TABLE 6. Classification accuracy for different unseen network topologies.

Test Type	Out of Service	Accuracy
23-bus N-1	G 102	96.37%
23-bus N-1	Line 203-205	96.56%
23-bus N-1	Line 4-16	96.44%
23-bus N-2	G1 and Line 5-18	95.83%
23-bus N-2	Line 203-205 and Line 3003-3005	95.89%
23-bus N-2	Line 151-152 and G 3011	95.86%
1648-bus N-1	G 107	97.03%
1648-bus N-1	Line 124-377	96.74%
1648-bus N-1	Line 6-10	96.88%
1648-bus N-2	Line 55-76 and 89-92	96.45%
1648-bus N-2	Line 134-138 and G181	96.68%
1648-bus N-2	Line 293-387 and Line 877-1045	96.31%

The classification accuracy of the DSA model decreases slightly when unseen network topologies are encountered, but it can still maintain a desirable classification accuracy in the tests of Table 6. Thus, the DSA model has good robustness to unseen network topologies.

F. TEST FOR MISSING DATA

Generally, due to various practical problems, including PMU malfunctions, communication delays or cyber-attacks, the input variables of the DSA model may be incomplete [7], [40]. To investigate the impact of missing data on the classification accuracy, five missing rates of input variables (10%, 20%, 30%, 40%, and 50%) are considered, and the missing variables are randomly selected.

The test results shown in Figure 7 indicate that the higher the missing data rate is, the lower the classification accuracy will be. However, the accuracy of the DSA model still maintains greater than 90% even the missing data rate is 50%. Therefore, the DSA model has good robustness to missing data.

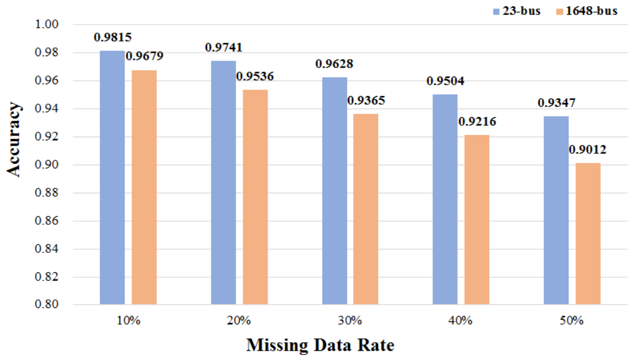

FIGURE 7. Classification accuracy for different data missing rates.

VI. CONCLUSION

Based on a powerful data-driven tool called SPORF, this paper proposes a data-driven scheme for real-time and reliable pre-fault DSA that includes three stages: offline training, periodic update and online assessment. In the offline stage, to address the class imbalance problem then reduce the impact of irrelevant and redundant features on the DSA model performance, an improved ADASYN method and a feature selection procedure based on the MRMR criterion are designed to process the initial knowledge base for efficient training of the SPORF-based DSA model. To address previously unseen system topologies and promote the generalization and adaptation of the proposed scheme, a model update mechanism is designed in the periodic update stage. In the online assessment stage, the DSA model is applied to rapidly provide reliable assessment results based on real-time PMU measurements.

The test results on two typical power systems (a 23-bus system and a 1648-bus system) demonstrate that the proposed scheme has superior DSA performance. In particular, the adaptability of the DSA model to topology changes is verified. Furthermore, the impact of missing data is studied, and the robustness to missing data is illustrated for the DSA model. Therefore, the proposed scheme is of great significance to the practical operation of power systems.

REFERENCES

[1] M. He, J. Zhang, and V. Vittal, “Robust online dynamic security assessment using adaptive ensemble decision-tree learning,” *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4089–4098, Nov. 2013.

[2] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, “Load forecasting, dynamic pricing and DSM in smart grid: A review,” *Renew. Sustain. Energy Rev.*, vol. 54, pp. 1311–1322, Feb. 2016.

[3] G. Wijeweera, U. D. Annakkage, W. Zhang, A. D. Rajapakse, and M. Rheault, “Development of an equivalent circuit of a large power system for real-time security assessment,” *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 3490–3499, Jul. 2018.

[4] M. de Jong, G. Papaefthymiou, and P. Palensky, “A framework for incorporation of infeed uncertainty in power system risk-based security assessment,” *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 613–621, Jan. 2018.

[5] Y. Zhang and L. Xie, “Online dynamic security assessment of microgrid interconnections in smart distribution systems,” *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3246–3254, Nov. 2015.

[6] F. J. Luo, Z. Y. Dong, G. Chen, Y. Xu, K. Meng, Y. Y. Chen, and K. Wong, “Advanced pattern discovery-based fuzzy classification method for power system dynamic security assessment,” *IEEE Trans. Ind. Informat.*, vol. 11, no. 2, pp. 416–426, Apr. 2015.

[7] C. Ren and Y. Xu, “A fully data-driven method based on generative adversarial networks for power system dynamic security assessment with missing data,” *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 5044–5052, Nov. 2019.

[8] F. Su, S. Yang, H. Wang, and B. Zhang, “Study on fast termination algorithm of time-domain simulation for power system transient stability,” *Proc. CSEE*, vol. 15, no. 15, pp. 4372–4378, Aug. 2017.

[9] T. L. Vu and K. Turitsyn, “Lyapunov functions family approach to transient stability assessment,” *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1269–1277, Mar. 2016.

[10] J. J. Q. Yu, D. J. Hill, A. Y. S. Lam, J. Gu, and V. O. K. Li, “Intelligent time-adaptive transient stability assessment system,” *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1049–1058, Jan. 2018.

[11] F. Hashiesh, H. E. Mostafa, A.-R. Khatib, I. Helal, and M. M. Mansour, “An intelligent wide area synchrophasor based system for predicting and mitigating transient instabilities,” *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 645–652, Jun. 2012.

[12] C. Liu, K. Sun, Z. H. Rather, Z. Chen, C. L. Bak, P. Thøgersen, and P. Lund, “A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees,” *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 717–730, Mar. 2014.

[13] A. N. Al-Masri, M. Z. A. A. Kadir, H. Hizam, and N. Mariun, “A novel implementation for generator rotor angle stability prediction using an adaptive artificial neural network application for dynamic security assessment,” *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2516–2525, Aug. 2013.

[14] F. R. Gomez, A. D. Rajapakse, U. D. Annakkage, and I. T. Fernando, “Support vector machine-based algorithm for post-fault transient stability status prediction using synchronized measurements,” *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1474–1483, Aug. 2011.

[15] Y. Zhou, J. Wu, Z. Yu, L. Ji, and L. Hao, “A hierarchical method for transient stability prediction of power systems using the confidence of a SVM-based ensemble classifier,” *Energies*, vol. 9, no. 10, p. 778, Sep. 2016.

[16] Y. Zhang, Y. Xu, Z. Y. Dong, Z. Xu, and K. P. Wong, “Intelligent early warning of power system dynamic insecurity risk: Toward optimal accuracy-earliness tradeoff,” *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2544–2554, Oct. 2017.

[17] R. Yan, G. Geng, Q. Jiang, and Y. Li, “Fast transient stability batch assessment using cascaded convolutional neural networks,” *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2802–2813, Jul. 2019.

[18] C. Liu, F. Tang, and C. L. Bak, “An accurate online dynamic security assessment scheme based on random forest,” *Energies*, vol. 11, no. 7, pp. 1914–1930, 2018.

[19] M. Sun, I. Konstantelos, and G. Strbac, “A deep learning-based feature extraction framework for system security assessment,” *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5007–5020, Sep. 2019.

[20] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, “A review of machine learning approaches to power system security and stability,” *IEEE Access*, vol. 8, pp. 113512–113531, 2020.

[21] B. Tan, J. Yang, Y. Tang, S. Jiang, P. Xie, and W. Yuan, “A deep imbalanced learning framework for transient stability assessment of power system,” *IEEE Access*, vol. 7, pp. 81759–81769, 2019.

[22] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[23] T. M. Tomita, J. Browne, C. Shen, J. Chung, J. L. Patsolic, B. Falk, C. E. Priebe, J. Yim, R. Burns, M. Maggioni, and J. T. Vogelstein, “Sparse projection oblique randomer forests,” *J. Mach. Learn. Res.*, vol. 21, no. 104, pp. 1–39, 2020.

[24] Y. Xu, Z. Y. Dong, J. H. Zhao, P. Zhang, and K. P. Wong, “A reliable intelligent system for real-time dynamic security assessment of power systems,” *IEEE Trans. Power Syst.*, vol. 27, no. 3, pp. 1253–1263, Aug. 2012.

[25] T. S. Aghdam, H. K. Karegar, and H. H. Zeineldin, “Transient stability constrained protection coordination for distribution systems with DG,” *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5733–5741, Nov. 2018.

[26] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1322–1328.

[27] B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht, “On oblique random forests,” in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Athens, Greece, 2011, pp. 453–469.

[28] S. Liu, L. Liu, Y. Fan, L. Zhang, Y. Huang, T. Zhang, J. Cheng, L. Wang, M. Zhang, R. Shi, and D. Mao, “An integrated scheme for online dynamic security assessment based on partial mutual information and iterated random forest,” *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3606–3619, Jul. 2020.

[29] L. Zhu, C. Lu, Z. Y. Dong, and C. Hong, “Imbalance learning machine-based power system short-term voltage stability assessment,” *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2533–2543, Oct. 2017.

[30] C. Zheng, V. Malbasa, and M. Kezunovic, “Regression tree for stability margin prediction using synchrophasor measurements,” *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1978–1987, May 2013.

[31] *PSS/E Users Manual*, Shaw Power Technol., Chennai, India, Aug. 2004.

[32] L. Zhu, C. Lu, and Y. Sun, “Time series shapelet classification based online short-term voltage stability assessment,” *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1430–1439, Mar. 2016.

[33] P. N. Papadopoulos and J. V. Milanović, “Probabilistic framework for transient stability assessment of power systems with high penetration of renewable generation,” *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3078–3088, Jul. 2017.

[34] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, p. 20–29, Jun. 2004.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[36] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new oversampling method in imbalanced data sets learning," in *Proc. Int. Conf. Adv. Intell. Comput.*, in Lecture Notes in Computer Science, Berlin, Germany, 2005, pp. 878–887.

[37] B. Zhao, J. Cao, Z. Zhu, and H. Zhang, "A new transient voltage stability prediction model using big data analysis," in *Proc. IEEE ISGT-Asia*, Nov. 2016, pp. 1065–1069.

[38] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[39] Y. Wang, Y. Li, W. Pu, K. Wen, Y. Y. Shugart, M. Xiong, and L. Jin, "Random bits forest: A strong classifier/regressor for big data," *Sci. Rep.*, vol. 6, no. 1, pp. 1–8, Sep. 2016.

[40] R. Deng, G. Xiao, R. Lu, H. Liang, and A. V. Vasilakos, "False data injection on state estimation in power systems—Attacks, impacts, and defense: A survey," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 411–423, Apr. 2017.

YANFENG LIN is currently a Researcher with the College of International Communications, China Three Gorges University, Yichang, China. His current research interests include artificial intelligence and knowledge engineering, security analysis, and risk control technology of power systems.

XINYAO WANG is currently a Researcher with the College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, China. She is also a member of the Hubei Provincial Collaborative Innovation Center for New Energy Microgrid, China Three Gorges University. Her research interests include knowledge engineering and security analysis of power systems.