A Quantile Regression Based Approach for Online Probabilistic Prediction of Unstable Groups of Coherent Generators in Power Systems

Seyed Mahdi Mazhari, *Member, IEEE*, Nima Safari, *Student Member, IEEE*, C.Y. Chung, *Fellow, IEEE*, and Innocent Kamwa, *Fellow, IEEE*

***Abstract*—This paper addresses a novel framework for probabilistic data-driven based prediction of unstable groups of coherent generators in large interconnected power systems. In contrast with the existing techniques in which deterministic classification or forecasting approaches are applied to an offline database, the current study is relying on a prediction interval (PI)-based method to tackle with prediction uncertainties. First, similarity coefficients (SCs) are considered as internal outputs and calculated for all offline cases based on rotor angle difference between any pair of generators. Then, at some generator terminals, selected via a feature selection process, voltage values are measured and exerted as the input features of the prediction tool. Quantile regression forest is conducted to generate PIs, in which several intervals with certain probabilities, are predicted for SCs between any pair of generators. Afterwards, the obtained PIs are used to shape an empirical cumulative distribution function of SCs; a Monte Carlo simulation is conducted therewith to find a reliable estimate of possible grouping patterns. Finally, a decision making phase is devised to draw clear distinction among various parts of the most plausible grouping pattern with respect to a reliability index. It can offer power operators a wider flexibility to select the corrective control strategy. On several IEEE test systems, the effective performance of the developed approach is put on show, followed by a discussion on results.**

***Index Terms*—Coherent generators, feature selection, Monte Carlo (MC) simulation, phasor measurement unit (PMU), prediction interval, probabilistic prediction, quantile regression forest (QRF), transient stability.**

I. Introduction

MODERN power systems are operating much closer to stability boundaries to improve efficiency and decrease operation and planning costs. It compels operators to employ diverse preventive, corrective, and emergency control strategies to ensure reliable operation of the grid, as power networks are recurrently exposed to various contingencies [1]. Meanwhile, gigantic size of the system has imposed serious challenges to stability analysis and control. Thus, various approaches have been introduced to either reduce the computational burden or facilitate opting a proper control action [2]. Generator grouping, known as coherency analysis, is one of the most popular tools which provides priceless inputs to dynamic model reduction and system aggregation techniques [3], network partitioning [4], and wide-area control studies [5]–[6].

Once a disturbance occurs in an interconnected power system, some generators, known as coherent groups, have tendency to swing in unison, that is, the generators in each group maintain an almost constant angular difference from each other [7]. In model reduction literature, this data is widely used to develop lower-order, but approximate, equivalent models of the original system [3]. With respect to network partitioning and wide-area control, coherent groups are considered as the basis of islanding and emergency control strategies [5]. This paper contributes to the generator groupings literature with regards to the latter applications.

Considering the importance of this problem, there are several research conducted in the literature to either identify or predict stability status and coherent groups of generators [7]–[24]. In overall, these studies can be classified into two categories:

- Slow coherency identification [7]–[12]: The system remains stable following a disturbance, wherein some generators swing together, but the angle or speed difference do not exceed the stability limits. Successful applications of fast Fourier transform [8], empirical mode decomposition [9], principal component analysis [10], and clustering techniques [11]–[12] are reported in the specialized literature.
- Unstable generator grouping [19]–[24]: The system lose synchronism after fault clearance and the grouping pattern allows operators to trigger prompt emergency actions such as generator tripping, fast-valving, and islanding to avoid blackouts or cascading failures [2]. Since the system may be exposed to widespread supply interruption in this situation, fast prediction of unstable groups of generators is crucial. It has various advantages over grouping identification approaches which are not suitable for real-time applications [19]. This paper zeros in on developing an online prediction technique for coherent generators that are subjected to system instability.

Prediction of unstable groups of generators consists of two basic stages, i.e., instability prediction and grouping prognostication. There are several research conducted in the literature to address early prediction of rotor angle instability [13]–[18]. Among them, data-driven methods, in which a prediction model is developed using a large set of training data, have drummed up interest, owing to their blessings in real-time applications [14]–[18]. As

Manuscript sent May 17, 2018. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Saskatchewan Power Corporation (SaskPower).

S.M. Mazhari, N. Safari, and C.Y. Chung are with the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon S7N 5A9, Canada(e-mail: s.m.mazhari@usask.ca; n.safari@usask.ca; c.y.chung@usask.ca).

I. Kamwa is with the Hydro-Québec/IREQ, Power System and Mathematics Varennes QC J3X 1S1, Canada (e-mail: kamwa.innocent@ireq.ca)

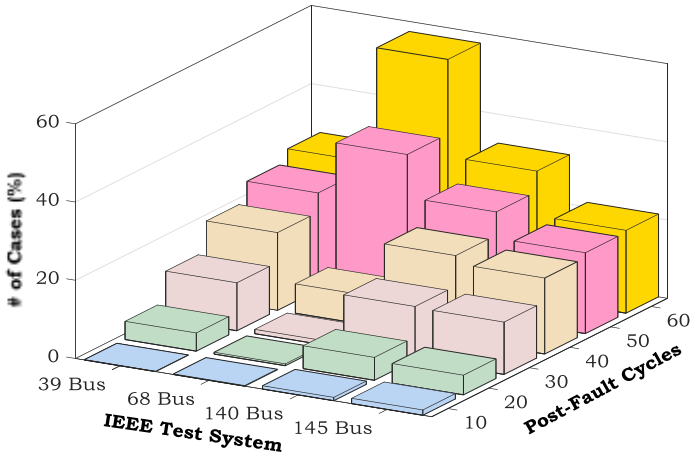

Fig. 1. Percentage of unstable cases that lose synchronism in different post-fault cycles for a database prepared in [9].

An instance, efficacy and robustness of extreme learning machine [16], core vector machine [17], and decision tree (DT) [17]–[18] have been reported. Although the aforesaid methods describe stability status of the network, they do not offer any information about the system dynamic behavior.

Much research has been carried out in the specialized literature to confront with the discussed matter [19]–[24]. Prediction of coherent groups of generators is considered as a multi-class classification problem in [19]–[23]. In [19], different prediction engines, including DT, random forest (RF), and support vector machine (SVM) are used with post-fault rotor angles as the input features. In [18], maximum Lyapunov exponents are calculated using online voltage magnitudes measured by phasor measurement units (PMUs) and used as inputs of an SVM based prediction model. Besides, efficacy of disparate input features are investigated in [21]–[22]. In [23], a decision forest method is exploited to solve the problem; furthermore, the sequence in which generator groups lose synchronism is also considered. Although studies [19]–[23] have attained commendable results using large window of input data, i.e. above 30 post-fault cycles, their accuracy hardly exceeds 90% for lower cycles. In order to resolve this issue, authors in [24] put into practice a prediction model for rotor angles of generators which is consequently served to ascertain generator groupings. Even though the developed method returned decent outcomes using 6 cycles of post-fault data, it is neither tested on large systems, nor considered topology changes in the process which can concretely affect the overall precision [23].

Considering the above mentioned studies, one can note that prediction accuracy of generators grouping is not perfect. Moreover, despite the fact that precision increases by extending length of predictors, there are several cases that may become unstable before the required amount of inputs can be gathered. Figure 1 depicts the percentage of unstable cases that lost synchronism at various post-fault time spans, based on the database prepared in [18]. As it can be seen in this figure, more than 20% of cases, in all test systems, would become unstable in less than 1 s (60 cycles); thus, from the practical point of view, the efficacy of the prediction models significantly degrades in larger post-fault spans. Functionality is an important factor that was rarely reflected in previous case studies, and may merit consideration.

Taking into account both accuracy and functionality, power system operators prefer to receive stability warning signs close to the fault clearing time [14], [18]. This requires inclusion of less post-fault information into the decision making process which subsequently may impose greater uncertainty in the prediction phase. To tackle with uncertainty related issues, probabilistic frameworks are employed in [17] and [23] to consider probability of various failures or uncertainty of system components. However, those methods are incapable of offering confidence level of a predicted state or pattern. Thus, in order to dispel such deficiency, it seems helpful to apply probabilistic estimation-based approaches to handle prediction errors.

Moreover, the structure of data-driven based techniques to solve the current problem has remained essentially unchanged in recent years. Majority of the proposed methods train a multi-class classifier for a network in which any unique grouping pattern is linked to a single cluster [19]–[23]. In online application, if the classifier fails to predict the exact class of a scenario, the outcome would be a wrong pattern; there is no idea which part of the proposed layout is erroneous, even though its distinction with the correct solution is minor. Since even the slightest error in generator groupings may lead to fallacious restorative control strategy, the system may be exposed to abrupt instability. Thus, a framework, which can separate different parts of a predicted grouping layout on the basis of a reliability index, would be of interest to power system operators. It allows them to utilize different control schemes for distinct parts of the network founded on the confidence level.

In addition, it is shown that a simple generator grouping may not be sufficient to preserve a network from instability [25]. Taking this into account, critical generators [19] and the order, in which critical generators lose synchronism [23], are also declared in past methods. However, the time span wherein each group of generators becomes unstable is not reported yet, though it can affect the adopted control action [25].

Aimed at addressing the aforementioned shortcomings of the coherency prediction methods proposed to date, a novel framework is put forward. A prediction interval (PI) based approach is developed for probabilistic estimation of similarity coefficients (SCs) defined for each pair of generators, as well as the generators' time span to instability (TSI). To the best of our knowledge, this is the first effort to adopt PIs into the transient stability domain. To such aim, quantile regression forest (QRF) is used to build the prediction models. The obtained quantiles are then used to from an empirical cumulative distribution function (CDF) for SCs. Thereupon, a Monte Carlo (MC) simulation is carried out via the procured CDFs, and a coherency identification technique is conducted to identify generator groupings for any set of SCs. The outcome of the MC process is used to evaluate SCs in reference to a reliability index and consequently to shape possible grouping patterns in conformity with the desired confidence levels. Finally, functionality of the proposed approach is appraised and compared with the existing methods via several IEEE test systems, including 10-, 16-, 48-, and 50-machine networks.

II. Identification of Generators Dynamic Behavior

As expressed in the past section, dynamic behavior of an interconnected power system in post-fault scenarios plays a significant role in opting a satisfactory remedial action. It can be fairly determined through two major modules, i.e. generators coherency identification and critical generators ordering [19]–[23]; the engaged methods therein are explained next.

A. Coherency Identification

There are several methods introduced in the literature for coherency identification of a multi-area interconnected power system [7]–[12], [19]–[24]. In overall, the general procedure triggers by calculating a similarity coefficient, SC, between any pair of generators:

$$SC_{ij} = \Upsilon(i,j), \qquad \forall\, i,j \in \Omega^G \tag{1}$$

where $\Upsilon(\cdot)$ and Ω^G represent a similarity function and the set of generators, respectively. With respect to low-frequency oscillations of a stable network, Υ is widely defined by the correlation between two dynamic signals [7], or using various transformations applied to generators speed or rotor angle values [9]. However, in an unstable situation, which is the primary focus of this study, the similarity function is mainly specified based on the rotor angle difference between generators, as it provides a consistent approach for coherency recognition [19].

The hierarchical clustering, which is performed well in previous studies, is employed in this paper [19], [23]. To do so, an agglomerative (bottom-up) strategy is conducted to form a hierarchical cluster tree. Each generator is initially considered as a single cluster, and then iteratively merged to the nearby clusters based on a linkage criterion. The final clusters are identified by cutting the tree using a predefined threshold for the linkage criterion, set to $360°$ in this work. Detailed process of the hierarchical clustering is explained in [19], [21], and [23].

B. Critical Generators Ordering

Once the generators coherency is known, critical groups of generators can be ordered based on the sequence in which they lose synchronism [23]. In the present work, generators' TSI is utilized to form the succession of critical generators. TSI of a generator indicates a time interval in the post-fault state in which the generator loses synchronism with the remaining part of the network.

In this respect, starting with the fault clearing moment as a reference time, $t = 0$, the maximum rotor angle difference between any pair of generators is calculated. If the angle difference violates the stability criteria in an instant of time, $t = \gamma$, the pair of generators associated with the largest angle difference is selected. Consider t^* as the final simulation time, angle difference between the current time (γ) and t^* is calculated for both generators of the selected pair. Among them, the generator with the higher value is chosen and its TSI is set to γ. Then, this generator is omitted and the same process is repeated till the stability criterion is met. Afterwards, a similar procedure is conducted for the next instances while either TSI of the remaining generators are obtained, or all time samples are observed. The TSI of a generator is assumed equal to the last time sample if the above process cannot assign a value to the generator.

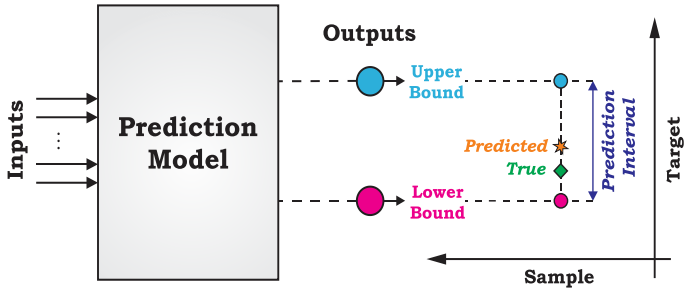

Fig. 2. A schematic predictor resulting PIs for a target value.

III. Prediction Interval and Quantile Regression Forest

To date, a vast majority of the stability prediction methods has restricted their attention solely to point forecasting based approaches. A point forecast represents a single number which is the most likely realization of the unknown true future value. Nevertheless, it does not deliver any information about variability around that predicted value, which is usually referred to as the degree of uncertainty involved in the estimation. PIs compensate for this deficiency by providing a range of values that one can be confident, the true value is encircled with a certain probability [26]. Such intervals can be remarkably informative as they represent a way of evaluating the extent to which a new observation may deviate from the deterministically predicted value. As an example, for a given 95% PI, one can be 95% confident that the new observation will fall within this range. Fig. 2 illustrates a symbolic predictor resulting PIs for a target value. It is shown that PIs are of greater importance to decision-makers compared to point estimates in practical applications, as they allow for a detailed assessment of the future uncertainty [26]–[28].

There are several methods introduced in the literature to construct PIs [28]; QRF is utilized in this study, as it has a common basis with RF, which has been successfully tested on stability prediction so far [12], [19], [23]. Moreover, it gives a nonparametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables, which is the aim of this study [29]. However, any other algorithm can be incorporated without loss of generality, as the employed PI method is not a part of the contribution of this paper. The basic methodology behind QRF and its main components, i.e., quantile regression (QR) and RF, are presented next.

A. Quantile Regression

In machine learning, regression is used as an approach to model the relationship between a response variable, $Y \in \mathbb{R}$, and a set of explanatory variables, $X \in \mathbb{R}^n$. For a given $X = x$, standard regression analysis attempts to exploit a least square technique to find an appropriate estimator, $z(x)$, in order to predict the conditional mean, $E(Y|X = x)$, via minimizing the expected square error loss as follows [28]:

$$E(Y|X = x) = \arg\min_{z} E\{(Y - z(x))^2 | X = x\} \tag{2}$$

Since the conditional mean may not convey sufficient information about power system dynamic behavior, other forms of regression should be considered to tackle with the uncertainties. QR is a specific type of regression analysis, which is originally used for construction of confidence intervals for a set of data; the

goal is to find arbitrary quantiles of the conditional distribution of y, as defined bellow [28]:

$$F(y|X=x) = P(Y \le y|X=x) \tag{3}$$

where $F(\cdot)$ and $P(\cdot)$ denote the CDF and probability distribution function (PDF), respectively. The τ quantile of a random variable x, $Q_\tau(x)$, can be described as (4), which means with a given probability of τ, observation x will be equal or less than $Q_\tau(x)$:

$$Q_\tau(x) = \inf\{y : F(y|X=x) \ge \tau\}, \quad 0 \le \tau \le 1 \tag{4}$$

B. Random Forests

RF is an ensemble learning method that performs by constructing a large number of decision trees, each capable of offering a response when fed by a set of predictor values [29]. In order to reduce the correlation between trees, RF employs a bootstrap aggregating technique in which a random subsample is extracted from the data with replacement for the process of tree growing. Furthermore, a random subset of predictors is selected as input features for each tree.

As a mean to elucidate the RF decision making process, assume a training data set as follow:

$$\mathcal{L} = \{(X_i, Y_i)_{i=1}^N |\, X_i \in \mathbb{R}^M, Y_i \in \mathbb{R}\} \tag{5}$$

where N and M are the number of training samples and number of features, respectively. Moreover, suppose ϑ_k as a random parameter vector which determines how the k^{th} tree, $T(\vartheta_k)$, is grown. For a typical sample from $\mathcal{L}$, like x, let $\mathcal{K}_\ell$ be a set of leaves, $\ell(x, \vartheta_k)$, that can be identified along the decision tree. The estimation of a $T(\vartheta_k)$ for a sample X_i, denoted by $\hat{Y}^k$, is obtained by calculating weighted average of the original observations [29]:

$$\hat{Y}^k = \sum_{i=1}^{N} \omega_i(x, \vartheta_k) \cdot Y_i \tag{6}$$

$$\omega_i(x, \vartheta_k) = \frac{1_{\{X_i \in \mathcal{K}_{\ell(x,\vartheta_k)}\}}}{\#\{j :\, X_j \in \mathcal{K}_{\ell(x,\vartheta_k)}\}} \tag{7}$$

where ω_i is a weight vector which gives a positive constant if observation X_i is part of $\ell(x, \vartheta_k)$, and 0 otherwise. The "#" symbol stands for numbers, and the ω_i weights sum to one. Finally, the conditional mean $E(Y|X=x)$ is obtained by average prediction of all trees (Ω^T), as show in (8):

$$\hat{Y} = E(Y|X=x) = \sum_{i=1}^{N} \omega_i(x) \cdot Y_i \tag{8}$$

$$\omega_i(x) = \frac{1}{|\Omega^T|} \cdot \sum_{k \in \Omega^T} \omega_i(x, \vartheta_k) \tag{9}$$

C. Quantile Regression Forest

QRF is an extension of RF and uses the same process to grow the trees. However, in comparison to RF which calculates and stores average observation for the entire leaves of the whole trees, it retains all relevant predictions. In other words, QRF preserves a raw distribution of all predictions at each leaf node. Such information can be used in a straightforward manner to assess the full conditional distribution of Y, for a given $X = x$, as bellow [29]:

$$F(y|X=x) = P(Y \le y|X=x) = E\left(1_{\{Y \le y\}} \middle| X = x\right) \tag{10}$$

Similar to RF, the conditional distribution can be approximated by the weighted mean over the observations of $1_{\{Y_i \le y\}}$ as represented in (11), in which $\omega_i(x)$ is calculated through (9).

$$\hat{F}(y|X=x) = \sum_{i=1}^{N} \omega_i(x) \cdot 1_{\{Y_i \le y\}} \tag{11}$$

Using (11), the conditional quantile $Q_\tau(x)$ can be estimated for a given τ, with $0 \le \tau \le 1$ [23]:

$$Q_\tau(x) = \hat{F}^{-1}(\tau) = \inf\{y : \hat{F}(y|X=x) \ge \tau\} \tag{12}$$

IV. The Proposed Solution Framework

Since the current study seeks to solve coherency prediction problem for unstable networks, stability status is an important factor which must be provided before triggering the process [19]. It is shown in the literature that the stability of a network can be reliably predicted with accuracies over 96% using pre-fault and during fault data [18], and over 99% using less than 10 cycles of post-disturbance data [14]–[15]. Thus, the amount of error which is imposed to the coherency prediction phase through the stability status identification is negligible in comparison to the overall prediction error [19]. In this paper, stability status is predicted based on [18] and fed into the proposed grouping framework, though any other method can be used in a straightforward manner.

The proposed coherency grouping framework consists of a training stage, MC simulation, and a decision making phase. Detailed description of each phase is delineated below.

A. Phase I: Training of Prediction Models

In order to train the prediction models, voltage magnitudes received from PMUs, fault type, and fault duration are considered as the predictors. The selected voltage values contain a cycle before fault occurrence, a cycle after fault time, a cycle before fault clearing time, and a few cycles of post-fault data which is set by the user preference. Absolute rotor angle difference is employed as the similarity function in (1), and SCs are calculated for any pair of generators. The total number of SCs is equal to all possible combinations of the two generators. Besides, TSI values are calculated for generators based on the procedure explained in Section II.B. The QRFs are utilized to predict both SCs and TSIs; in overall, for a network with n_g machines, the total number of prediction models, N_{PM}, which can be solved in parallel, is as follows:

$$N_{PM} = \frac{n_g \cdot (n_g - 1)}{2} + n_g \tag{13}$$

It might be helpful to mention that the QRF method selects various sets of features for different leaves; therefore, it has a suitable level of resistivity against inappropriate input features [29]. However, in this paper, a mutual information based feature selection algorithm is also applied to the input features in order to decrease both the number of trees as well as sensitivity to PMU delays [18].

B. Phase II: Monte Carlo Simulation

Since the number of quantiles which can be obtained from a PI model is limited, the PI models are not still able to fully represent

all the possible realization of SCs. To thoroughly capture the uncertainty in SCs, first, CDF of the entire SCs are estimated. There are several methods in the specialized literature which can efficiently generate random numbers with a desired precision, by relaying on the quantile values or a given empirical CDF [30]–[31].

Then, an MC simulation is performed to consider the effects of prediction uncertainty of SCs on the grouping pattern of coherent generators. In this study, a technique proposed by [31] is put into practice for CDF estimation; it requires a set of four quantiles which are obtained from the PI models. At any MC iteration, random values are generated based on the predicted quantiles and assigned to SCs. Once all SCs are generated, the hierarchical clustering technique, explained in Section II.A, is applied to find the grouping pattern. Based on an obtained layout at iteration k, a set of binary variables, ζ_{ij}^{k}, is initialized; ζ_{ij}^{k} is set equal to 1 if generators i and j are in a same group, and 0 otherwise. The stopping criterion of the MC simulation is calculated as (14):

$$\max_{i,j\in\Omega^G}\left(\frac{\sqrt{\mathcal{Z}_{ij}}}{\sqrt{NS\cdot\mathcal{M}_{ij}}}\right)\le\sigma \tag{14}$$

$$\mathcal{M}_{ij}=\frac{1}{NS}\cdot\sum_{k=1}^{NS}\zeta_{ij}^{k} \tag{15}$$

$$\mathcal{Z}_{ij}=\frac{1}{NS}\cdot\sum_{k=1}^{NS}\left(\zeta_{ij}^{k}-\mathcal{M}_{ij}\right)^2 \tag{16}$$

where NS is the number of MC iterations, and σ is set to 0.002 in this paper [32]. $\mathcal{M}_{ij}$ and $\mathcal{Z}_{ij}$ respectively calculate mean and variance of the binary variables. Eq. (14) certifies that sufficient number of simulations has been carried out for all SCs.

C. Phase III: Decision Making

The coherent groups of generators can be represented by a graph structure; the vertices indicate generators, and each edge denotes that the vertices at both ends belong to a same group. Thus, a connected graph, which has a path between every pair of vertices, portrays a stable network; and a disconnected graph stands for an unstable system. In this respect, the amount of $\mathcal{M}_{ij}$, calculated in (15), represents the probability of existing an edge between i and j vertices. Based on the outcomes of the MC simulation, a reliability index can be assigned to each edge, $\mathcal{R}_{ij}$, indicating the solution consistency in all MC iterations:

$$\mathcal{R}_{ij}=100\%\cdot\begin{cases}1-\mathcal{M}_{ij}, & \mathcal{M}_{ij}<0.5\\ \mathcal{M}_{ij}, & \mathcal{M}_{ij}\ge 0.5\end{cases} \tag{17}$$

If $\mathcal{M}_{ij}=1$, generators i and j are assigned to the similar coherent groups in the entire MC scenarios; similarly, $\mathcal{M}_{ij}=0$ implies that no edge should be drawn between $i-j$. In both cases, $\mathcal{R}_{ij}=100\%$. If the QRF fails to yield sharp PIs for a target SC, the prediction uncertainty will be high and the random generator, which uses the predicted quantiles, will be incapable of providing realizations close to the actual values. Thereby, there might be severe discrepancies in ζ_{ij}^{k} at distinct MC iterations which consequently lead to lower reliability index. In the worst case, $\mathcal{R}_{ij}=50\%$, that symbolizes the cut-off used by the point forecast methods.

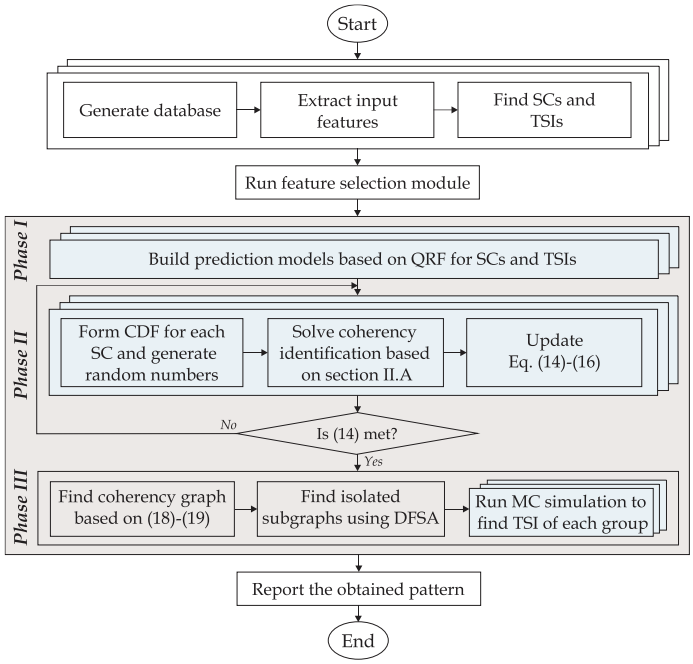

Fig. 3. Overall process of the proposed method applied to offline database.

By setting a reliability cut-off, $RL^*\in[50\ \ 100]$, a graph can be formed using those arcs which meet the confidence criterion:

$$\mathbb{G}_{RL^*}=\left(\Omega^G,[\mathcal{E}]_{|\Omega^G|\times|\Omega^G|}\right) \tag{18}$$

$$\mathcal{E}_{ij}=\begin{cases}0, & \mathcal{R}_{ij}<RL^*\\ \begin{cases}0, & \mathcal{M}_{ij}<0.5\\ 1, & \mathcal{M}_{ij}\ge 0.5\end{cases}, & \mathcal{R}_{ij}\ge RL^*\end{cases},\quad \forall\, i,j\in\Omega^G \tag{19}$$

where $\mathbb{G}_{RL^*}$ is a graph which represents coherent groups of generators with RL^* confidence interval. The number of groups and a list of generators belonging to each group can be found using a depth-first search algorithm (DFSA) [33].

Depending on the input features and the selected RL^*, some edges may not pass the reliability condition, which consequently may leave some generators as single groups. In this regard, (18)-(19) confirm a generator is not assigned to a multi-member coherent group unless the predefined reliability level is met. Such an attribute certifies a significant privilege of the developed framework in comparison to the past studies; it decreases the misclassification rate and restricts instability of the network caused by an erroneous grouping pattern. Based on (19), decreasing RL^* leads to more edges in $\mathbb{G}_{RL^*}$; $RL^*=50\%$ reports a deterministic solution, similar to previous methods, in which prediction uncertainty is ignored.

With the objective to find ordering of the critical generators for a selected layout, a simple MC simulation is carried out; similar to Phase II, different values are randomly generated for TSIs based on the predicted quantiles. Then, an average TSI is calculated for each group and the MC process is repeated till the stopping criterion is met [23]. Afterwards, the coherent groups are sorted in an ascending order based on the obtained average TSIs.

The overall process of the proposed method is shown in Fig. 3. As it can be seen in this flow diagram, major portions of the procedure, illustrated with collateral blocks, can be solved in parallel resulting in a substantial decrease of computational burden.

TABLE I
DATA FOR THE NETWORKS USED IN SIMULATIONS

Network	# of buses	# of cases in database (unstable %)	# of coherency patterns
10-machine	39	10000 (27.11%)	49
16-machine	68	10000 (12.72%)	20
48-machine	140	15000 (18.57%)	31
50-machine	145	15000 (15.45%)	63

D. Online Application Procedure

In case of online application, the proposed coherency prediction module triggers by a sign received from the central control indicating a fault occurred in the network. The algorithm starts collecting the required input data as mentioned in Section IV.A. Once the protecting switches disconnected the faulted line, a stability status prediction algorithm is conducted to check whether the system remains stable or not. This process takes a fraction of a cycle [18]. If the system is identified as unstable, the gathered predictors apply to the prediction models trained in Section IV.A. Then, MC simulations are carried out in parallel over the obtained PIs. Finally, the decision making process is executed as stated in Section IV.C. It is empirically seen in simulations that the entire process can be accomplished in a couple of cycles for different test systems, if PMU delays are ignored.

V. TESTS AND RESULTS

A. Description of Test Systems and Simulation Tools

Aimed at evaluating performance of the proposed approach, computer routines are performed in a MATLAB environment. The developed method is applied to several test networks, including IEEE 36-, 68-, 140-, and 145-bus systems. The data required for offline analysis are generated via the power system toolbox (PST) package [34]. In the database generation phase, different fault types, fault locations, fault resistances, and fault durations (2–15 cycles) are considered. Besides, the load demand at each bus is randomly changed between 0.65–1.25 of the base value. Furthermore, offline analysis is conducted so that the nominal power network topology as well as $N-1$ and $N-2$ contingencies, respectively cover 85, 14, and 1% of the whole database [18]. All simulations are carried out for 20 s after the fault clearance time and the PMUs are assumed to measure two samples per cycle [19]. Coherency identification algorithm, introduced in Section II.A, is applied to the whole database, and distinct unstable coherency patterns are identified. A summary of the generated database is reported in Table I. The computer used in simulations featured an Intel 3.4–GHz CPU with 16 GB of RAM.

B. Prediction of Transient Stability Status

As the first step, a DT based solution framework proposed in [18] is conducted to predict unstable cases. A stratified 5-fold technique is employed to divide the whole database; the evaluation process is repeated five times using different training sets, and 20% of the database is accounted as test samples in each iteration. Different post-fault cycles (PFCs) of voltage samples are considered as inputs of the classifiers. The obtained results are reported in Table II; as it can be seen in this table, the DT classifier is capable of predicting over 93% of unstable cases without any PFC data. The mean accuracy exceeds 97% using 60 cycles

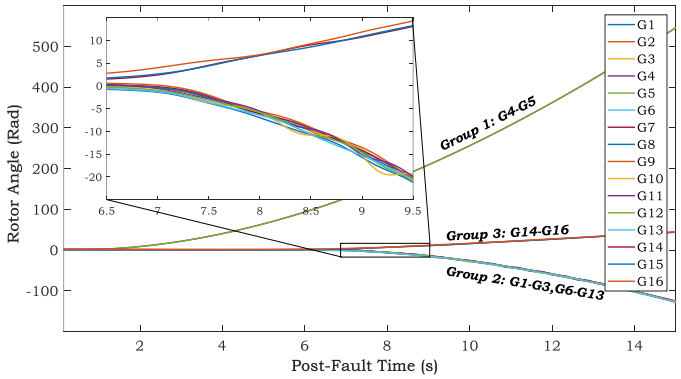

Fig. 4. Rotor angle variation of generators for a line-to-ground fault occurred near bus 16 of the IEEE 68-bus test system.

TABLE II
MEAN ACCURACY OF THE DT BASED METHOD ([18]) TO PREDICT UNSTABLE CASES USING DIFFERENT POST-FAULT CYCLES

Length of predictors	39-bus	68-bus	140-bus	145-bus
No PFC	95.33%	95.01%	99.63%	93.38%
20 PFC	95.91%	96.19%	99.66%	94.80%
40 PFC	96.70%	97.52%	99.75%	96.14%
60 PFC	98.27%	98.75%	99.84%	97.66%

of post-fault voltage samples. Thus, one can note that unstable cases can be predicted with suitable accuracy, which makes it reasonable to solve the coherency prediction in two stages [19]. Therefore, only unstable scenarios are considered to evaluate the proposed grouping framework in the subsequent simulations [19]–[24].

C. Coherency Prediction for a Test Scenario

In order to elaborate the procedure of the proposed method, obtained results of the developed algorithm is reported in detail for a sample test scenario. A line-to-ground fault is applied close to bus 16 of the IEEE 68-bus system, on a line connecting buses 16 and 19. The transmission line between buses 46 and 49 is considered to be out of service before the contingency taken place; load demands at network buses are randomly changed based on Section V.A. The fault is cleared after 5 cycles and the obtained rotor angles of generators are illustrated in Fig. 4. As it can be seen in this figure, the system is unstable and the generators are clearly separated into three coherent groups.

The input features of this test case are formed based on Section IV.A and fed into the trained QRF models. There are 16 generators in this study; hence, based on (13), 136 prediction models are trained in parallel, consisting of 120 models for SCs and 16 models related to TSIs. Both SCs and TSIs are normalized before running the training phase. The developed QRF models predict different quantiles for a target value, which consequently is used to estimate a probability distribution for the expected output. Fig. 5 represents the predicted 90% PI bounds for SC_{15} as well as PDF of several samples which are randomly generated based on the predicted quantiles. As it can be deduced from this figure, the target value of SC_{15} is about 0.04 and the predicted 90% PI is [0.0091,0.0667]; the user can be 90% confident that the value of SC_{15} would fall within this boundary. PDF of the samples produced based on the predicted quantiles shows that the similarity coefficient between generators 1 and 5 may hardly surpasses 0.08; this subsequently unveils that these generators belong to

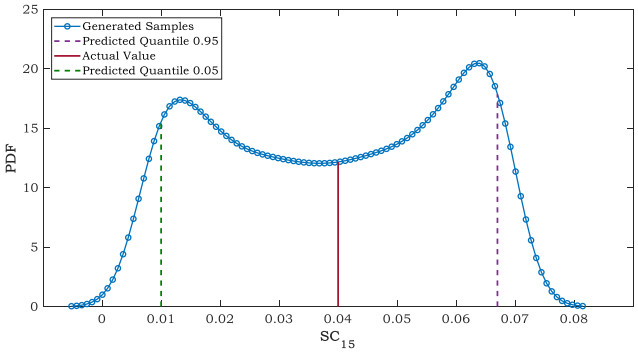

Fig. 5. PDF of samples generated for SC_{15} based on the predicted quantiles.

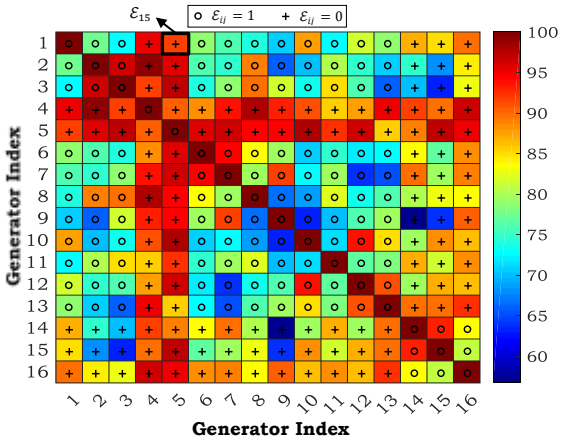

Fig. 6. Reliability index for all possible edges of the coherency graph of Fig. 4.

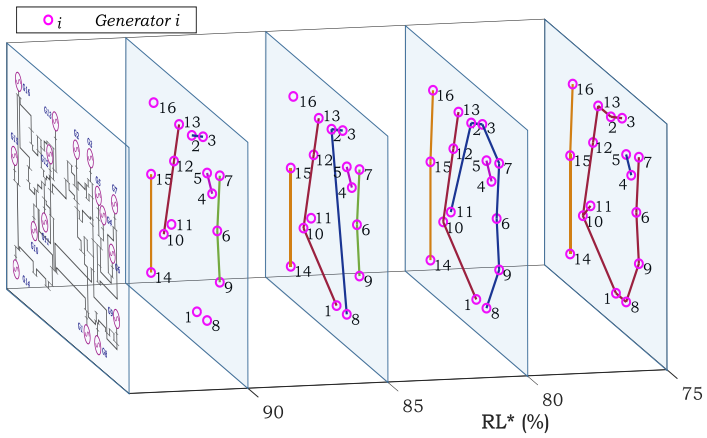

Fig. 7. Grouping patterns of generators for different RL^*.

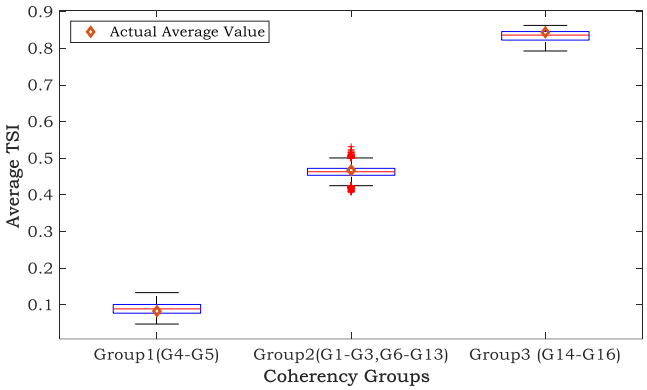

Fig. 8. Average TSI of coherency groups for $RL^* = 75\%$.

different coherency groups, as shown in Fig. 4.

Similarly, quantile values of the entire SCs are extracted from the prediction models and the MC simulations are carried out. Based on the MC outputs, the reliability index of (17) is calculated for the test scenario, and depicted in Fig. 6. This figure indicates the status of an edge between any pairs of generators $\mathcal{E}_{ij}$, and its reliability in accordance with the MC simulations. As an instance, it declares that generators 1 and 5 are classified in separate coherency groups ($\mathcal{E}_{15} = 0$) in around 94% of all MC cases.

Having the edge status and its reliability for all possible combinations of two generators, final grouping patterns can be obtained with respect to the user confidence priorities. To such aim, Fig. 6 is injected into the decision making phase and the obtained coherency layouts are shown in Fig. 7 for different reliability cut-offs. As it can be seen in this figure, for $RL^* = 90\%$, there are five groups with more than one member ({G2,G3}, {G4,G5}, {G5,G7,G9}, {G10,G12,G13,}, {G14,G15}), and four groups with a single generator, i.e. G1, G8, G11, G16. By checking the rows of Fig. 6 which are associated with G1, G8, G11, and G16, it would be clear that there is no edge with a reliability over 90% to connect these generators to others; so, each formed a single group. By decreasing RL^*, there are more edges that meet the confidence requirement; it is shown in Fig. 7 that $RL^* = 75\%$ can lead to the correct grouping pattern, same as Fig. 4. It should be mentioned that the edges depicted in Fig. 7 are solely used to represent coherent groups. Thus, there might be several other edges which meet the reliability criterion, though they are omitted to avoid confusion. However, the neglected edges do not change the coherency patterns. The groupings depicted in Fig. 7 offer valuable information to the operators. They draw clear distinction between layouts with different reliability levels. Moreover, they report confidence level for different parts of a selected pattern. As an example, while reliability of finding the correct pattern is around 75%, the algorithm detects G4 and G5, which is a critical coherency group, with over 93% confidence. Furthermore, by reporting conservative solutions for $RL^* > 50\%$, the proposed method does not assign a generator to other groups, unless it fulfills the reliability requirements. This helps the operator to avoid adopting an improper control strategy that may push the system towards instability.

The obtained patterns of each reliability cut-off can be sorted based on the trained TSI models. The coherency groups are fed into a simple MC simulation and the average TSI of each group is reported in Fig. 8 via a boxplot representation. It can be seen that the obtained results could sharply bound the actual values in all groups with high reliability. Therefore, the predicted average TSIs clearly distinguish the consequence of the critical groups.

The entire simulations, required to predict the coherency groupings and TSIs of this test case, i.e. phases II and III of Section IV, are accomplished in 0.0071 s. It is empirically seen in all simulations that the longest calculation time for these two phases, which occurred in the IEEE 140-bus system, did not exceed 0.0184 s which is around 1 cycle; this processing time decreased to 0.0109 s while running on a 64 processor Intel E5-2660 2.0-GHz CPU with 64 GB of RAM. It might be helpful to mention that this test system contains 48 machines and represents the electrical system of the northeast power coordinating council,

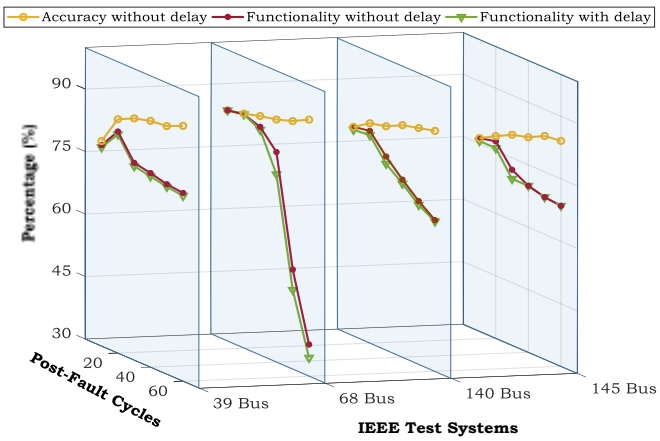

Fig. 9. Performance of the proposed method using different post-fault cycles for $RL^* = 50\%$.

TABLE III
COMPARISON OF ACCURACY OF DIFFERENT METHODS FOR COHERENCY PREDICTION USING $RL^* = 50\%$

Prediction model	PFC	39-bus	68-bus	140-bus	145-bus
DT based multi-class classification [19]	20	86.10%	82.23%	81.67%	75.72%
	60	90.44%	91.26%	86.12%	83.09%
ANN based rotor angle forecasting [24]	20	77.63%	74.65%	73.52%	69.93%
	60	83.77%	84.18%	77.49%	73.95%
Proposed method without PMU noise	20	**86.31%**	**82.78%**	**82.80%**	**78.51%**
	60	**91.32%**	**91.65%**	**87.71%**	**84.13%**
Proposed method with PMU noise	20	84.39%	80.42%	80.59%	76.47%
	60	90.25%	90.07%	85.33%	83.86%

which supplies a relatively large-area of North America; thus, the obtained simulations illustrate that the developed method is suitable for real-time applications of practical systems.

D. Performance of the Proposed Framework in Various Conditions

In order to assess efficacy of the proposed framework, the developed method is conducted on different IEEE test systems and tested for several PFCs. The reliability cut-off is set to 50% to enables meaningful comparison with the traditional multi-class classification techniques [19]–[24]. The obtained results are depicted in Fig. 9. As it can be seen in this figure, the accuracy of prediction increases by expanding PFCs. Functionality is also shown in this figure as an evaluation metric. It refers to the percentage of cases which are correctly predicted before losing synchronism. The figure presents that increasing PFCs over 20 cycles substantially decrease functionality in all networks. As an instance, functionality decreases from 86.1% to 59.32% by increasing PFCs from 20 to 40 in IEEE 68-bus test system. Since functionality is a dominant factor to evaluate a module from the operators' perspective, coherency prediction should be made with few PFCs in real-life applications. However, few PFCs lead to more prediction uncertainty and consequently lesser accuracy. This proves the necessity of employing probabilistic based approaches for coherency prediction, which is the main contribution of this paper.

Moreover, the effects of PMU delays are also included in the simulation results. To this end, different delays up to 200 ms are randomly added to the input signals. 2000 cases are generated for each scenario in the database and the average functionality is

TABLE IV
DETAILED COMPUTATIONAL TIME OF THE PROPOSED METHOD WITH RESPECT TO THE EXISTING TECHNIQUES FOR IEEE 140-BUS TEST SYSTEM ($PFC = 60$)

Application	Function		DT [19]	ANN [24]	Proposed Method
Offline (Training Data)	Generate database (h)		16.1458		
	Model development (min.)		3.7528	4.6001	17.8150
Online (Test Data)	Average CPU Time (s)	Without delay	0.0008	0.0013	0.0159
		With 200 ms delay	0.1604	0.2013	0.0526
	Prediction Accuracy	Without noise	86.12%	77.49%	87.71%
		With noise	82.89%	73.06%	85.33%
	Outputs	Generator grouping	✓	✓	✓
		Critical generators	-	-	✓
		Time to instability	-	-	✓
		Prediction uncertainty	-	-	✓

shown in Fig. 9. Compared to the situation in which PMU delays are ignored, the average prediction functionality of all networks decreases by 2.1%. It might be helpful to mention that QRF develops several trees which operate by different input features. Thus, it is able to provide suitable responses even if a couple of features are missed.

For the sake of comparison with the existing methods, a DT classifier proposed in [19] and the ANN algorithm reported in [24] are employed. The developed algorithms utilize different sizes of post-fault voltage samples as inputs and triggered by the same stability status prediction module as that of the proposed method, for better consistency. The obtained results are reported in Table III. It can be observed that the proposed method with $RL^* = 50\%$, could result comparable solutions in contrast to both methods, in all cases; this is while improving the accuracy of deterministic coherency prediction is not the main focus of this paper. Besides, the obtained results confirm that indirect identification of coherent groups via the predicted SCs is an effective approach. Comparing the outcomes of [19] and [24], it can be deduced that the classification based technique could bring about better prediction accuracies than the regression based framework used in [24]. The accuracy of the ANN solutions are less than what reported in [24] as the topology changes, ignored in that study, are also considered while creating database in this paper

The performance of the developed method is also investigated in the presence of PMU noise. White Gaussian noise is added to whole database with a signal-to-noise ratio equal to 34 dB [18], and the training process is repeated. The obtained results are shown in Table III; on average, classification accuracy is decreased by 1.73% in all cases in presence of PMU noise. Since the amount of noise considered in this study is almost twice of the requirement of IEEE C37.118.1-2011 standard, the developed method can realize superior accuracies in real-life situations compared to this severe test.

VI. DISCUSSION

The current study has proposed the inclusion of PI-based approach for coherency prediction of power systems. It helps the operators to come to grips with prediction uncertainties by providing extra information regarding confidence levels of dif-

ferent grouping patterns. Since such framework may give rise to more computational complexity, it seems helpful to discuss the course of dimensionality with more details and compare it with the exiting techniques.

Detailed computational burden associated with the IEEE 140-bus tests system is reported in Table IV. As it can be seen in this table, the time required to generate the database is similar for all methods and is around 16 h. Transient stability program is conducted in parallel for 15,000 separate scenarios using MATLAB parallel processing toolbox. Inasmuch as an actual power system may have hundreds of generation units, database generation may take several hours and it can be considered as a practical drawback of the entire machine learning based algorithms. To address this concern, dynamic equivalencing techniques might be used to reduce the original network and consequently shorten the simulation times [35]; however, this is beyond the scope of current study.

Computational burden of the model development phase is also illustrated in Table IV. It takes around 17 min for the proposed method to train PI models, in parallel, for all SCs and TSIs. Although this time is above 4 times greater than that of [19] and [24], it is practically reasonable as the prediction models are constructed once in the process and can be in function for several weeks.

Based on Table IV, average CPU time of the proposed method is around 0.0159 s in online applications. This time is significantly larger than the existing techniques, i.e. around 20 times more than [19], but clearly meets the realistic expectations, as it is still less than a cycle. The MC simulations are the main culprit for this computational burden, though they are conducted in parallel. Besides, with the inclusion of 200 ms delay for PMUs, CPU time of the proposed method is increased to 0.0526. However, since the developed framework is less sensitive to PMU delays, due to the QRF nature and involvement of feature selection, it is more than three time faster than [19], [24]. It might be helpful to mention that a 3% tolerance is considered in this study; i.e. the accuracy of the delayed scenario should not be less than 97% of the solutions without delay.

Considering the above discussions and with respect to Table IV, one can note that the proposed method meets the practical computational requirements in online applications. In addition, it is more resistive to PMU noises and delays with respect to the existing techniques. Besides, the proposed algorithm provides more information to operators such as time to instability and prediction uncertainties which can be included into the emergency control approaches. Thus, since the developed framework can result in better prediction accuracies than the existing methods, and taking into account the mentioned side benefits, it might be a helpful tool for real-life systems.

VII. Conclusion

This paper introduced a novel framework for probabilistic coherency prediction of unstable generators in power systems. QRF, as a PI based approach, was developed to estimate similarity coefficients for all possible combinations of two generators. The quantiles obtained from the trained predictors were incorporated to form probable distribution of the target response. The coherency groups were molded into a graph theory based representation and MC simulations were carried out to find reliability of each edge. In addition, a decision making phase was performed to find the grouping patterns based on the predefined confidence levels. The method developed was successfully tested on various IEEE test systems; the results and comparisons revealed that the proposed method can be used as an effective tool for generators grouping prediction of power systems.

Further research may be conducted to increase the prediction accuracy of the proposed framework through applying different machine learning techniques. Moreover, the application of the proposed probabilistic approach in finding optimal corrective control strategy can be investigated.

References

[1] E. Vaahedi, *Practical Power System Operation,* Piscataway, NJ: Wiley, 2014.

[2] Z. Li, G. Yao, G. Geng, Q. Jiang, "An efficient optimal control method for open-loop transient stability emergency control," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2704–2713, 2017.

[3] Y. Liang, X. Lin, A. M. Gole, and M. Yu, "Improved coherency-based wide-band equivalents for real-time digital simulators," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1410–1417, 2011.

[4] F. Raak, Y. Susuki, and T. Hikihara, "Data-driven partitioning of power networks via Koopman mode analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2799–2807, 2016.

[5] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparsity-promoting optimal wide-area control of power networks," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2281–2291, 2014.

[6] A. Thakallapelli, S. J. Hossain, and S. Kamalasadan, "Coherency and online signal selection based wide area control of wind integrated power grid," *IEEE Trans. IEEE Trans. Indus. App.*, vol. 54, no. 4, pp. 3712–3722, 2018.

[7] A.M. Khalil, R. Iravani, "A dynamic coherency identification method based on frequency deviation signals," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1779–1787, 2016.

[8] A. Vahidnia, G. Ledwich, E. Palmer, and A. Ghosh, "Generator coherency and area detection in large power systems," *IET Gener., Transm., Distrib.*, vol. 6, no. 9, p. 874, 883, 2012.

[9] S. Avdakovic, E. Becirovic, A. Nuhanovic, and M. Kusljugic, "Generator coherency using the wavelet phase difference approach," *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 271–278, 2014.

[10] K. K. Anaparthi, B. Chaudhuri, N. F. Thornhill, and B. C. Pal, "Coherency identification in power systems through principal component analysis," *IEEE Trans. Power Syst.*, vol. 20, no. 3, pp. 1658–1660, 2005.

[11] R. Agrawal and D. Thukaram, "Support vector clustering-based direct coherency identification of generators in a multi-machine power system," *IET Gener., Transm., Distrib.*, vol. 7, no. 12, p. 1357, 1366, 2013.

[12] M.H.R. Koochi1, S. Esmaeili, and R. Fadaeinedjad, "New phasor-based approach for online and fast prediction of generators grouping using decision tree," *IET Gener. Transm. Dis.*, vol. 11, no. 6, pp. 1566–1574, 2017.

[13] P. Bhui and N. Senroy, "Real-time prediction and control of transient stability using transient energy function," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 923–934, 2017.

[14] 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. pp, no. 99, pp. 1–10, 2017.

[15] R. Zhang, Y. Xu, Z.Y. Dong, and K.P. Wong, "Post-disturbance transient stability assessment of power systems by a self-adaptive intelligent system*," IET Gener. Transm. Dis.*, vol. 19, no. 5, pp. 296–305, 2015.

[16] B. Wang, B. Fang, Y. Wang, H. Liu, and Y. Liu, "Power system transient stability assessment based on big data and the core vector machine," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2561–2570, 2016.

[17] T. Guo and J.V. Milanovic, "Probabilistic framework for assessing the accuracy of data mining tool for online prediction of transient stability," *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 377–385, 2014.

[18] S.M. Mazhari, N. Safari, C.Y. Chung, and I. Kamwa, "A hybrid fault cluster and Thévenin equivalent based framework for rotor angle stability prediction," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5594–9603, 2018.

[19] T. Guo and J.V. Milanovic, "Online identification of power system dynamic signature using PMU measurements and data mining," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1760–1768, 2016.
[20] M. Rashidi and E. Farjah, "LEs based framework for transient instability prediction and mitigation using PMU data," *IET Gener. Transm. Dis.*, vol. 10, no. 14, pp. 3431–3440, 2016.
[21] P.N. Papadopoulos, J.V. Milanovic, P. Bhui, and N. Senroy, "Fast online identification of power system dynamic behavior," in *Proc. IEEE Power & Energy Society General Meeting*, Chicago, IL, US, pp. 1-5, 2017.
[22] E.A. Frimpong, J. Asumadu, and P.Y. Okyere, "Real time prediction of coherent generator groups," *Journal of Electrical Engineering*, vol. 16, no. 3, pp. 47-56, 2016.
[23] P.N. Papadopoulos, Tingyan Guo, and J.V. Milanovic, "Probabilistic framework for online identification of dynamic behavior of power systems with renewable generation," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 45–53, 2018.
[24] S.A. Siddiqui, K. Verma, K.R. Niazi , and M. Fozdar, "Real-time monitoring of post-fault scenario for determining generator coherency and transient stability through ANN," *IEEE Trans. Indus. App.*, vol. 54, no. 1, pp. 685–692, 2018.
[25] J.D. Morales, P.N. Papadopoulos, and J.V. Milanović, "Feasibility of different corrective control options for the improvement of transient stability," *IEEE PowerTech Conference*, pp. 1-6, 2017.
[26] A. Khosravi, and S. Nahavandi, "Combined nonparametric prediction intervals for wind power generation," *IEEE Trans. Sustain. Energy*, vol. 4, no. 4, pp. 849–856, 2013.
[27] N. Safari, S.M. Mazhari, C.Y. Chung, "very short-term wind power prediction interval framework via bi-level optimization and convex cost function," *IEEE Trans. Power Syst.*, vol. pp, no. 1, pp. 1–11, 2018.
[28] R. Koenker, and K.F. Hallock, "Quantile regression," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 143–156, 2001.
[29] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Res.*, vol. 7, no. 1, pp. 983–999, 2006.
[30] J.F. Lawless, *Statistical Models and Methods for Lifetime Data*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 2003.
[31] N.L. Johnson, "Systems of frequency curves generated by methods of translation," *Biometrika*, vol. 36, no. 1, pp. 149–176, 1949.
[32] S.M. Mazhari, H. Monsef, R. Romero, "A multi-objective distribution system expansion planning incorporating customer choices on reliability," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1330–1340, 2016.
[33] R. Tarjan, "Depth-first search and linear graph algorithms," in *Proc. 12th Annual Symposium on Switching and Automata Theory*, vol. 1, pp. 114–121, 1971.
[34] J.H. Chow and K.W. Cheung, "A toolbox for power system dynamics and control engineering education and research," *IEEE Trans. on Power Syst.*, vol. 7, no. 4, pp. 1559-1564, 1992.
[35] G. Chavan, M. Weiss, A. Chakrabortty, S. Bhattacharya, A. Salazar, and F. Habibi-Ashrafi, "Identification and predictive analysis of a multi-area wecc power system model using synchrophasors," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 1977-1986, 2017.