IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 December 2022, accepted 13 January 2023, date of publication 19 January 2023, date of current version 26 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3238316

APPLIED RESEARCH

Dynamic Security Analysis Framework for Future Large Grids With High Renewable Penetrations

KISHAN PRUDHVI GUDDANTI, (Member, IEEE), BHARAT VYAKARANAM, (Senior Member, IEEE), KAVERI MAHAPATRA, (Member, IEEE), ZHANGSHUAN HOU, (Member, IEEE), PAVEL ETINGOV, (Senior Member, IEEE), NADER SAMAAN, (Senior Member, IEEE), TONY NGUYEN, (Member, IEEE), AND QUAN NGUYEN, (Member, IEEE)

Pacific Northwest National Laboratory, Richland, WA 99354, USA

Corresponding author: Kishan Prudhvi Guddanti (kishan.g@pnnl.gov)

This work was supported by the Advanced Grid Modeling (AGM) Project from the Department of Energy's Office of Electricity (DOE-OE).

ABSTRACT The 100% renewable energy targets from the policymakers for the future grids have drawn a significant amount of interest. The high renewable penetration made future large interconnected grids (LIGs) more volatile and harder to understand using historical observations. This led to the need to study future LIGs using a wide range of future-year operating conditions and contingencies. However, existing planning tools are not sufficient for the dynamic security assessment (DSA) of these future LIGs due to a lack of detailed modeling capabilities and computational limitations when processing a wide range of scenarios. This paper addresses these two challenges by proposing; 1) an efficient modeling framework that can generate large grid's dynamic data with cascade behaviors for a wide range of scenarios. This data is generated by respecting the constraints of production cost models and their respective AC power flow dynamic simulation models at an hourly resolution; 2) an unsupervised machine learning(ML)-based approach for fast scanning the DSA data. The proposed approach uses feature engineering techniques and fast Fourier transforms to transform the time series signals into visually distinguishable frequency domain signals for DSA. The proposed simulation framework is used to generate 1.485 terabytes of dynamic simulation data for the 2028 WECC system containing 4455 scenarios. The proposed ML framework used the 2028 WECC system to demonstrate its effectiveness and speed in identifying critical scenarios.

INDEX TERMS Dynamic security assessment, unsupervised learning, feature engineering, large grids.

I. INTRODUCTION

Dynamic security assessment (DSA) is one of the critical aspects of power system studies that are used to identify critical contingencies in the grid by analyzing their corresponding dynamic security constraint violations on the grid [1]. For lower computation burden, DSA planning studies for large interconnected grids (LIGs) are typically conducted only on a limited number of operating/loading conditions and well-known critical contingencies. These critical contingencies are selected using historical data, and the transmission operator's experience and judgment [2]. However, there is a shift in paradigm for DSA of power systems due to increasing penetrations of distributed energy resources (DERs), demand-side management technologies, home energy management technologies, and energy storage systems [3]. As a result, the grids of the future even at the bulk transmission level are turning out to be more volatile due to 100% renewable energy targets passed down by legislature at the state level [4].

Hence, the DSA of future grids must now consider various combinations of renewable (wind and solar) generation mixes that result in volatile operating conditions. This variable renewable generation mix with high penetrations is

The associate editor coordinating the review of this manuscript and approving it for publication was Youngjin Kim.

transforming the future grids, motivating a need to perform DSA for 1) a wide range of operating conditions considering variable renewable generation mix, and 2) a wide range of contingencies beyond the traditional priority list approach. To incorporate these two new constraints, many industries around the world focused on the development of scenarios for the future grids [5], [6]. However, these scenarios cannot be directly plugged and played in the grid without considering production cost models (PCMs) [7], [8]. PCMs for large interconnected grids generate chronological, hourly DC power flow cases for future scenarios considering economics, wind, photovoltaic, and load levels. PCMs use lossless, linear, DC power-flow solutions but DSA requires AC power-flow (ACPF) convergence for reliability planning studies [7], [8], [9], [10]. Hence, PCMs alone are not sufficient for reliability planning studies. Therefore, future grid DSA studies require PCM-embedded AC power-flow cases for a large number of scenarios and the existing planning tools are no longer suitable.

The development of future grid analysis tools is a growing area of interest for researchers and industries around the world [11]. Melbourne Energy Institute proposed a plan in [12] for a future Australian grid relying 100% on renewable resources. Reference [13] showed that balancing 100% renewable resources is possible in Australian National Electricity Market. The U.S RTO, PJM [14] showed that their grid can operate up to 99.9% of the time purely on renewable resources with cost estimates comparable to today's cost. However, these studies are based on balancing areas and are no longer accurate to understand the DSA of the system with very high DER penetration levels.

To effectively model and better understand the DSA of future grids, [15] showed on a large Irish power grid that chronological ''time series scanning'' is more effective than past experience-based methods since it can capture the varying operating conditions and inter-seasonal variations due to high wind penetrations. However, [15] selected the worst operating points from many years worth of data to reduce computational burden and the time-consuming nature of the time series scanning is not discussed. Reference [16] proposed a particle swarm optimization (PSO) based framework to address these future grid issues of outdated planning tools by using PCM embedded power-flow cases, and a machine learning-based approach [16] demonstrated the results on a simple IEEE 14-generator test system and addressed the issue of the local convergence behavior of PSO by increasing the size of swarm population which further increased the computational burden. However, it is not trivial to extend the same strategy from [16] to LIGs with 22,000+ buses as it further increases the computational burden compared to that of a simple 14-bus system. Hence, the original problem of needing a fast scanning framework to perform fast and efficient modeling-based DSA studies on future large interconnected grids (LIGs) remains unsolved. Specifically, this work addresses the problems associated with the planning of future LIGs like **1) efficient modeling framework for DSA studies, 2) time-consuming nature, and 3) big data issues of time series scanning.** After proposing our modeling framework to perform dynamic simulations of future LIGs in this paper, we also proposed a fast scanning tool (FAST) to address the time-consuming nature of the problem with the help of novel feature engineering capabilities of machine learning (ML) tailored for power system dynamic data. The main contributions of this paper are as follows:

1) The proposed simulation framework in this paper is demonstrated to generate terabyte-scale dynamic simulation data for better planning of future large interconnected grids considering a wide range of operating conditions and contingencies.
2) Proposed FAST methodology is demonstrated on WECC 22,000+ bus system and compared against the true critical cases of the system.
3) In this paper, strategically layered feature engineering techniques transform the raw data's high-dimensional time series signals into visually distinguishable signals from the perspective of total voltage and frequency limit violations. Hence, we use the K-mean clustering algorithm because it performs best when the inputs are visually distinguishable and gives an expected behavior.
4) In this paper, once the cluster centers are obtained using the K-means, the design property of feature-engineered inputs helped to quickly identify the cluster with critical scenarios. Using this, critical scenarios are identified accurately and hundreds of times faster than the time series scanning (brute force) method under a wide range of operating conditions and contingencies.

The paper is organized as follows. Section II presents the simulation framework to accurately model the future LIG simulations. Section III presents the proposed feature engineering approach to handle dynamic simulation's time series data. Section IV presents the overall proposed framework. Section V provides the demonstration of FAST and simulation results on the WECC system. Section VI concludes the paper and Section VII presents the future scope.

II. FUTURE GRID'S FRAMEWORK FOR DYNAMIC SIMULATION STUDIES

In this section, we present the framework that is used to generate the dynamic simulation data for planning future large interconnected grids (LIGs) under a wide range of operating conditions and contingencies. This framework addresses the problems that existing planning tools are facing to study future LIGs with high DER penetrations. Specifically, this framework uses various industry-focused Pacific Northwest National Laboratory's (PNNL's) tools [10], [17], [18] as the foundation and the process is illustrated in Fig. 1.

As shown in Fig. 1, the proposed framework for future grid planning studies has five main components. They are 1) market simulation-production cost models (PCM),

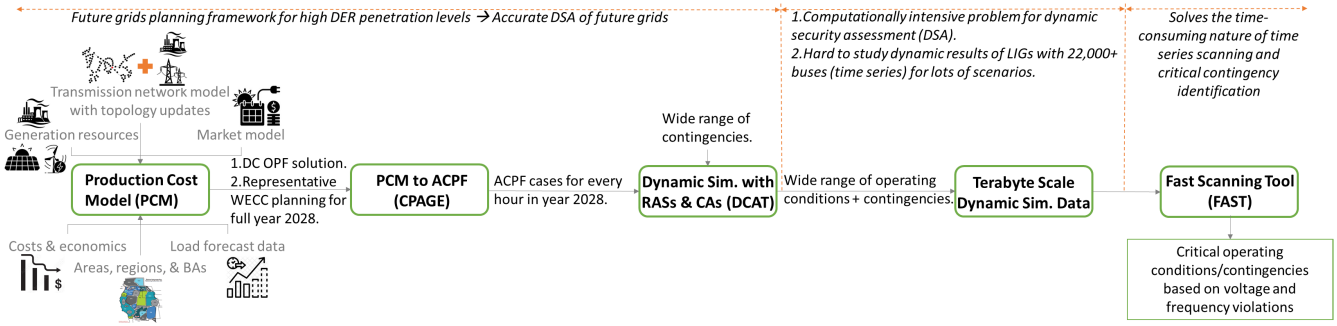

FIGURE 1. Framework for the planning of future large interconnected grids. .

2) Chronological AC Power Flow Automated Generation Tool (C-PAGE) [10] to translate data sets from PCMs and PFMs, 3) Dynamic Contingency Analysis Tool (DCAT) [18] to assess the impact of extreme contingencies and potential cascading events, 4) DSA metrics from terabyte-scale data, and 5) FAST.

A. FUTURE GRIDS PLANNING FRAMEWORK

In this work, we solve the resource adequacy challenge from an economic perspective by employing the PCM of the WECC power grid developed for the year 2028 at "hourly" resolution [17], hereafter referred to as 2028 Anchor Data Set (ADS). 2028 ADS represents the trajectory of recent Western Interconnection planning information, developments, and policies for the year 2028 [17]. As shown in Fig. 1, the highlights of this hourly chronological 2028 ADS (PCM) include the consideration of high DER penetrations for different WECC substations, planned transmission upgrades, decommission of existing generator units, commissioning of new generator units, WECC load forecast data for the year 2028, various types of generation units and their economic models of operation.

After PCM, as shown in Fig. 1, we resolve the DCPF to ACPF conversion challenge by adapting the PNNL's C-PAGE tool. C-PAGE automatically converts the WECC's PCM (2028 ADS) into corresponding converged ACPF cases. This combination of PCM [17] and C-PAGE [10] provides hourly ACPF cases of the WECC system for the year 2028. The availability of the ACPF cases at such a granular level for large grids of future years enables us to perform realistic and accurate DSA studies.

After CPAGE, as shown in Fig. 1, we resolve the problem of capturing cascading behaviors for LIGs like the WECC system by taking advantage of the PNNL's DCAT.

B. TERABYTE-SCALE DYNAMIC SIMULATION DATA FOR DSA

In this subsection, first, we describe the quality of dynamic simulation data. Second, we discuss the big data issues (time-consuming) of processing (time series scanning) the dynamic simulation data for DSA studies.

1) TERABYTE-SCALE DYNAMIC SIMULATION DATASET

Hereafter we use the term "scenario" to define a unique combination of an operating condition (hour of the year) and a contingency. For a given scenario, the shape of the output dynamic data from the DCAT module is given by $(\alpha \times N * S)$, where α represents the dynamic simulation time steps, S represents the total physical quantities of interest for DSA (e.g.: $S = 2$ if voltage and frequency magnitudes are of interest) and N represents total buses in the system ($N = 22883$ for the WECC system). We would like to highlight that the shape $(\alpha \times N * S)$ varies significantly depending on the values of α, S, and N. For example, α can change for different scenarios for DCAT when the cascading analysis is finished [18]. N can change depending on if only load buses or a mix of different buses are being considered for data analysis. For example, in our studies, the size of such a dynamic simulation dataset for 4455 unique scenarios on the WECC system resulted in 1.485 Terabytes of information to process for DSA studies. **The proposed framework combining 2028 ADS, PCM, CPAGE, and DCAT enabled the study of DSA of future large interconnected grids while the existing planning tools are outdated to capture the intricacies that the future grids experience. This is our first contribution.**

2) BIG DATA ISSUE OF TIME SERIES SCANNING

To process and visualize such a large volume of data, a MongoDB-based database management module was developed at PNNL [19]. However, it is observed that the database management module is slower to identify and retrieve critical scenarios based on total voltage and frequency limit violations. This slow processing time took many hours and even longer when the new scenario's dynamic simulation data was added to the existing scenarios' dataset. In this work, we addressed the time-consuming nature of the time series scanning approach for DSA studies in two steps. First, we propose a strategically layered set of feature engineering techniques to analyze terabyte-scale dynamic simulation data and obtain visually distinguishable transformed signals for K-means (discussed in Section III). Second, we input these visually distinguishable transformed signals

(from the perspective of voltage and frequency limit violations) into a clustering algorithm to obtain a novel and expected behavior out of the unsupervised ML technique (discussed in Section IV).

III. FEATURE ENGINEERING FOR DSA OF TERABYTE-SCALE DYNAMIC SIMULATION DATASETS

In this section, we discuss the challenges encountered during various stages of developing the proposed ML framework (FAST) and how we solved these challenges by strategically layering different feature engineering techniques together.

A. OBJECTIVE OF ML FRAMEWORK (FAST)

Before we discuss the methodology, first we present the objective of the proposed ML framework. Given different scenarios (unique combinations of operating conditions and contingencies), the objective is to identify the scenarios with the largest total voltage and frequency constraint violations. The total voltage constraint violations ($\mathcal{N}(v)_{\mathcal{T}_n}$) in a given scenario ($\mathcal{T}_n$) is given by (1).

$$\mathcal{N}(v)_{\mathcal{T}_n} = \sum_{i \in N} \sum_{t \in \{1,\cdots,\alpha_n\}} \beta_t^i,$$
$$\text{where } \beta_t^i = \begin{cases} 1; & \text{if } v_t^i > 1.05 \text{ p.u. OR if } v_t^i < 0.95 \text{ p.u.} \\ 0; & \text{if } 0.95 \leq v_t^i \leq 1.05, \end{cases} \tag{1}$$

N represents the total buses in the system, α_n represents the total dynamic simulation time steps for scenario $\mathcal{T}_n$, v_t^i represents the voltage magnitude value of bus i at time step t. A similar definition as equation (1) is used for total frequency violations with upper and lower thresholds of 60.005 Hz and 59.975 Hz respectively.

B. STAGE 1: DATA STORAGE

The output dynamic simulation data from DCAT (PSLF) in Fig. 1 is saved into "channel" (.chf) file format. This dynamic data is converted into parquet format and stored.

C. STAGE 2: OBTAINING REPRESENTATIVE DATA POINT OF A SCENARIO

In this subsection, we discuss the challenges of directly using the raw data as input for clustering algorithms and how these challenges are strategically addressed in the proposed framework to obtain good representative data points of different scenarios. After resolving the issues of memory and reading speed in Section III-B, one approach to solving the objective of this paper is by Obtaining representative data points corresponding to different scenarios and performing K-means clustering [20] on these data points.

1) REPRESENTATIVE DATA POINT OF A SCENARIO

For example, in the case of scenario $\mathcal{T}_k$, the shape of output dynamic simulation data for voltage magnitude signal is given by $(\alpha_k \times N)$ and as discussed above, the shape of its representative data point is given by $(\alpha_k * N \times 1)$ where α_k represents the dynamic simulation time steps for scenario $\mathcal{T}_k$; N represents the total buses in the system. However, we observed several challenges when the time-series signals are directly used as inputs to the K-means clustering algorithm. These challenges are discussed and addressed as follows

2) INTERPOLATION FOR MISSING DATA (ISSUE 2)

The lengths of the "representative data points" of each scenario are observed to be different; making it not possible to employ K-means clustering on the dataset with multiple scenarios (from issue 2 in Fig. 2). This is because the output dataset of different scenarios from DCAT has different lengths of simulation due to missing data and modeling of CAs as described in Section. II-B. For example, we observed in a few scenario datasets that some of the data corresponding to a given dynamic time step is repeated in two consecutive rows. These duplicates need to be removed and all the scenarios' datasets must be mapped with the same dynamic time step values.

3) REMEDY

For each scenario, We used linear interpolation to account for the missing data. It is shown by industry researchers in [21] that for high-frequency signals, linear interpolation yields reliable and accurate results. We also observed highly accurate interpolation results since the power system dynamic simulation data is very rich (80 samples available for every second).

Proposition 1: *The linear interpolation helps to adjust missing data in power system dynamic simulations and obtain equal length time-series signals when dynamic simulation datasets are not mapped for dynamic time step values.*

4) NORMALIZATION AT DIFFERENT kV LEVELS (ISSUE 3)

K-means clustering tries to cluster the different scenarios' data points (horizontally concatenated time-series signals of shape $(\alpha_k * N \times 1)$) using Euclidean distance. However, as illustrated in "issue 3" of stage 1 in Fig. 2, the magnitudes of time-series signals at different buses are different depending on the bus kV level. This contributes to biased weights/importance on certain features of the data points used for K-means clustering. Furthermore, these unnormalized values also contribute to the dataset being less compact. This "compactness issue of clustering" is further discussed separately below.

5) REMEDY

This challenge could be addressed by normalizing the signal values of buses using their base kV values. Additionally, this also contributes to a relatively more compact dataset (when compared to an unnormalized formulation) for K-means clustering.

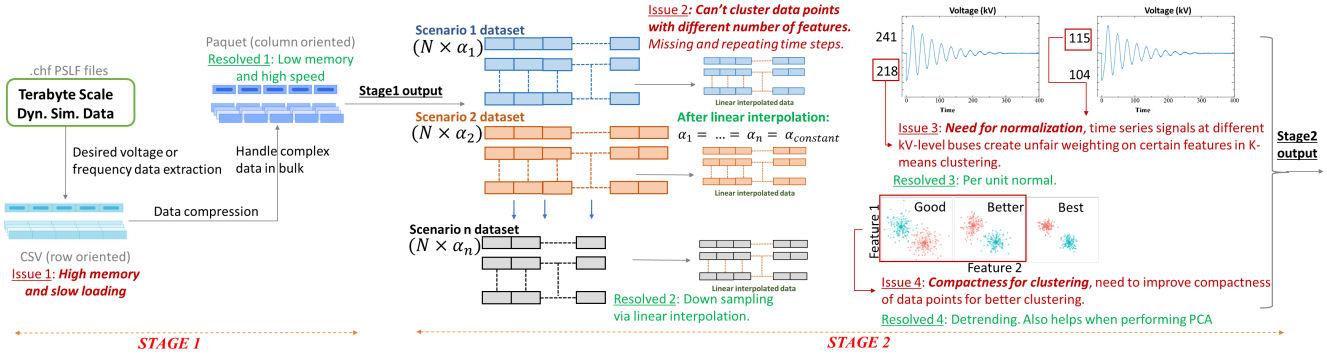

FIGURE 2. Stages 1 and 2 of the FAST for dynamic security assessment of future year large grids.

Proposition 2: *Normalization at bus kV levels helps to remove biased feature importance and improves the compactness of the dataset for K-means clustering*

6) COMPACTNESS ISSUE FOR CLUSTERING (ISSUE 4)

In this stage, we detrended (subtracting signals with 1 p.u. value) the time series signals for two reasons. They are 1) Principal component analysis is used in the later stages which requires a zero-mean dataset, and 2) detrending also helps to improve the K-means optimization process. We will show below how the objective of K-means clustering gets minimized as the compactness of the dataset is improved [22].

7) RELATION BETWEEN COMPACTNESS AND K-MEANS

The "issue 4" (stage 2) in Fig. 2 illustrates the compactness of a dataset and its visual impact when clustering. Formally, a dataset with different clusters is compact when it's total/expected conditional variance across all the clusters is smaller. The compactness of a dataset is given by $\mathbb{E}_p\left[var\left[Z|Z \in O_p\right]\right]$

$$= \sum_{p=1}^{k} var\left[Z|Z \in O_p\right] \cdot P\left(Z \in O_p\right), \tag{2}$$

where $\mathbb{E}_p$ is the expectation across all clusters, *var* represents the variance, Z is a random variable that represents the data points in cluster O_p, O_p is the p^{th} cluster, $p = \{1, 2, \cdots, k\}$ and k is the total clusters, and $P(\cdot)$ represents the probability function. Equation (2) can be further simplified as follows.

$$* = \sum_{p=1}^{k} \left(\frac{c_p}{c}\right) \cdot var\left[Z|Z \in O_p\right],$$

where c_p = count of data points in O_p, c = total data points

$$= \sum_{p=1}^{k} \left(\frac{c_p}{c}\right) \cdot \mathbb{E}\left[(Z-\mu)^2 |Z \in O_p\right], \tag{3}$$

$$\left(\because var\left[X\right] = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] = \mathbb{E}\left[\left(X-\mu_p\right)^2\right]\right)$$

$$= \sum_{p=1}^{k} \left(\frac{c_p}{c}\right) \cdot \left(\frac{1}{c_p} \cdot \sum_{y \in O_p} \left(y-\mu_p\right)^2\right), y \in Z \tag{4}$$

$$(\because \mathbb{E}\left[X\right] = \text{average of } X)$$

$$= \frac{1}{c} \cdot \sum_{p=1}^{k} \sum_{y \in O_p} \left(y-\mu_p\right)^2. \text{ (K-means objective)} \tag{5}$$

Therefore, from (2) and (5), we can see that minimizing the compactness of the dataset directly minimizes the K-means objective function's value even before the K-means is executed. Hence detrending helps to 1) obtain a zero mean dataset for PCA, and 2) improve the compactness of the dataset and thereby reduce the K-means objective value for better convergence.

Proposition 3: *Detrending of the time series signals reduces the magnitude of the variance in the dataset which makes the dataset more compact and thereby directly minimizes the K-means clustering objective beforehand.*

D. STAGE 3: PCA ON VERY LARGE MATRICES

Considering the objective of this paper from Section III-A, the application of stages 1 and 2 helped to reliably represent a scenario as a data point (a vector). This helped to apply K-means clustering on time-series datasets. However, we observed another challenge below when we input the representative data points of multiple scenarios as input to K-means clustering.

1) CURSE OF DIMENSIONALITY

We encountered the issue of longer training periods because the length of the representative data point is too large ("issue 6" from stage 3 in Fig. 3). For example, in the case of the WECC system, the typical length of a single data point representing a scenario is (18306400×1). Calculating the distances between such data points with more than 1 million features renders the K-means clustering algorithm to have poor convergence behavior and there is no guarantee of its performance.

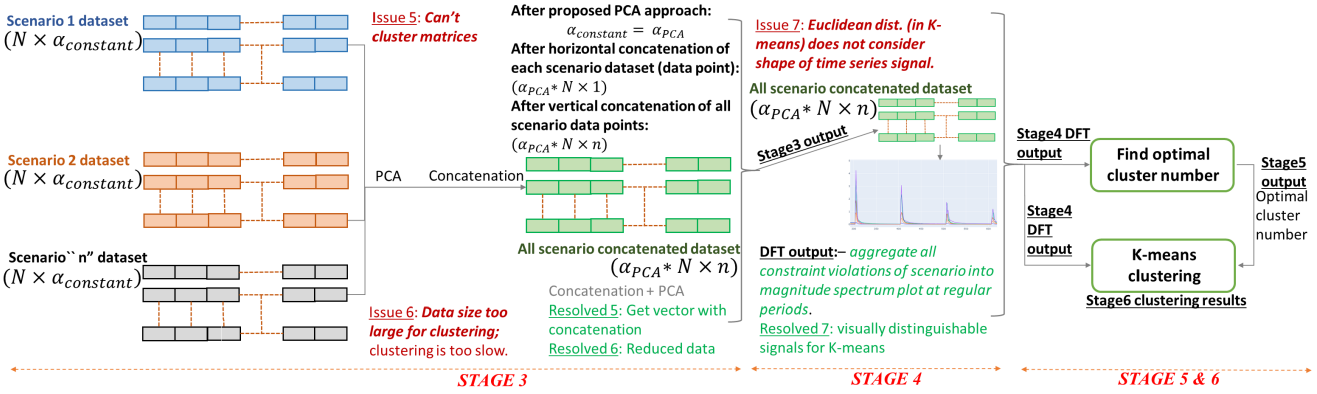

FIGURE 3. Stages 3, 4, 5, and 6 of the proposed FAST for dynamic security assessment of future large interconnected grids.

2) REMEDY FOR THE CURSE OF DIMENSIONALITY

One way to address the issue of the curse of dimensionality is by reshaping the dataset into a smaller shape by using standard principal component analysis (PCA). For example, as shown in stage 3 of Fig. 3, using PCA the shape of all scenarios' representative data points can be reduced from $\{(\alpha_1 * N \times 1), (\alpha_2 * N \times 1), \cdots, (\alpha_n * N \times 1)\}$ to $\{(\alpha_{PCA} * N \times 1), (\alpha_{PCA} * N \times 1), \cdots, (\alpha_{PCA} * N \times 1)\}$ where $\alpha_{PCA} < \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$. The standard approach to identify the value of α_{PCA} is by performing PCA on the very huge matrix $\mathbb{R}^{(\alpha_{PCA} * N \times n)}$. Unfortunately, applying PCA on such a matrix takes a lot of time since singular value decomposition of very large matrices (eg: from the WECC system) takes more computational power. In this paper, we solved this issue as follows.

3) COMPUTE γ_n

Identify the optimal number of principal components ($\gamma_n \in \mathbb{Z}^+$) that can retain 99% of the variance of the original dataset corresponding to one arbitrary scenario $\mathcal{T}_n \in \mathbb{R}^{(\alpha_n * N \times 1)}$. For scenario $\mathcal{T}_n$, its corresponding optimal number of principal components (γ_n) is calculated as follows. Let $\boldsymbol{\mathcal{B}_n}$ be the original dataset of scenario $\mathcal{T}_n$, $\boldsymbol{\mathcal{X}_n}$ be reduced/projected data given by $\boldsymbol{\mathcal{X}_n} = \boldsymbol{\mathcal{B}_n \mathcal{V}}$, $\boldsymbol{\mathcal{V}}$ be defined as the matrix with γ_n eigenvectors as columns i.e., $\boldsymbol{\mathcal{V}} := [\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_{\gamma_n}}]$. r be the number of samples in the dataset. λ_r represents the eigenvalues. The covariance of projected data ($\boldsymbol{\mathcal{C}_{\mathcal{X}_n}}$) is given by (6).

$$\begin{aligned}
\boldsymbol{\mathcal{C}_{\mathcal{X}_n}} &= \frac{1}{r} \cdot \left(\boldsymbol{\mathcal{X}_n^T \mathcal{X}_n}\right) = \frac{1}{r} \cdot \left(\left(\boldsymbol{\mathcal{B}_n \mathcal{V}}\right)^T \left(\boldsymbol{\mathcal{B}_n \mathcal{V}}\right)\right), \\
&\implies \frac{1}{r} \cdot \left(\boldsymbol{\mathcal{V}^T \mathcal{B}_n^T \mathcal{B}_n \mathcal{V}}\right) = \boldsymbol{\mathcal{V}^T} \left(\frac{1}{r} \cdot \boldsymbol{\mathcal{B}_n^T \mathcal{B}_n}\right) \boldsymbol{\mathcal{V}} \\
&= \boldsymbol{\mathcal{V}^T \mathcal{C}_{\mathcal{B}_n} \mathcal{V}} \\
&\implies \boldsymbol{\mathcal{V}^T \mathcal{C}_{\mathcal{B}_n} \mathcal{V}} = \sum_{\forall r} \lambda_r \sum_{\gamma_n} \left(\boldsymbol{v_{r,\gamma_n}}\right)^2, \\
\therefore \boldsymbol{\mathcal{C}_{\mathcal{X}_n}} &= \sum_{\forall r} \lambda_r \sum_{\gamma_n} \left(\boldsymbol{v_{r,\gamma_n}}\right)^2. \qquad (6)
\end{aligned}$$

Therefore, from (6), the optimal number of principal components (γ_n) for scenario $\mathcal{T}_n$ can be obtained by maximizing the variance of the projected dataset ($\boldsymbol{\mathcal{C}_{\mathcal{X}_n}}$) which is given by (7).

4) COMPUTE α_{PCA}

Obtain the set of optimal number of principal components $\{\gamma_1^*, \gamma_2^*, \cdots, \gamma_n^*\} \in \mathbb{Z}^+$ for all scenarios $\{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_n\}$. Now compute the final number of optimal principal components (α_{PCA}) where $\alpha_{PCA} = \max\left(\gamma_1^*, \gamma_2^*, \cdots, \gamma_n^*\right)$.

$$\max\left(\boldsymbol{\mathcal{C}_{\mathcal{X}_n}}\right) = \max_{\gamma_n^*} \left(\sum_{\forall r} \lambda_r \sum_{\gamma_n} \left(\boldsymbol{v_{r,\gamma_n^*}}\right)^2 \right). \qquad (7)$$

5) PCA FOR TRAINING PHASE

For each scenario in the dataset, we used α_{PCA} and the individual scenario's eigenvector matrix to perform dimensionality reduction. This max function ensures that the variance across all scenarios' datasets is within the desired threshold value while also making sure all the scenarios' data points have the same dimensions ($\alpha_{PCA} * N \times 1$) (stage 3 of Fig. 3).

Proposition 4: *Performing PCA on say 1252 dynamic scenarios' simulation data is equivalent to performing PCA on a matrix with dimensions* $\mathbb{R}^{(\alpha_{PCA} * N \times n)} = \mathbb{R}^{(18306400 \times 1252)}$. *In this work, we divided the problem into sub-problems, solved the sub-problems, and combined it for a final solution. This solved the large matrix issue and thereby the convergence issue of K-means for the power grid's time-series dynamic datasets.*

During testing phase, instead of performing PCA separately on every scenario in the testing dataset (similar to the training phase above). we randomly selected an eigenvector matrix that is calculated during the training phase and used it to transform the scenarios from the testing dataset. This helped to improve the speed of the testing phase. Furthermore, if the testing accuracy is worse then we recommend following a similar approach as that of the training phase above. However, interestingly, we did not observe any low accuracy situations in our studies on the WECC system with this approach.

E. STAGE 4: FREQUENCY DOMAIN CONVERSION FOR K-MEANS ON DYNAMIC TIME-SERIES DATA

Even though we addressed the challenges described in earlier subsections, the K-means algorithm with Euclidean distance is not equipped to identify critical scenarios defined by voltage or frequency limit violations using time series data. This is because euclidean distance does not compare the shapes of two time-series signals but only considers their overall magnitudes. Therefore, there is a need to aggregate the constraint violations (from the time-series domain) using a common reference point. In this subsection, we propose to solve this issue (stage 4 in Fig. 3) by using the real discrete Fourier transform (DFT) method. Specifically, the real DFT method is first used to convert all PCA-transformed time-series signals (from stage 3 of Fig. 3) into frequency domain at regular periods (stage 4 from Fig. 3). Finally, the amplitudes mapped with different frequencies of the DFT signals are used as inputs to the K-means clustering. The amplitude represents the aggregated constraint violation information whereas the frequency mappings helped to resolve the issue of a common reference point. **The illustration of the output transformed signal in the frequency domain is presented in Section V.**

Real DFT equations are given by

$$F_{real}^{i,\mathcal{T}_j}[k] = \frac{2}{\alpha_{PCA}} \cdot \sum_{n=1}^{\alpha_{PCA}-1} x_{i,\mathcal{T}_j}[n] \cos\left(\frac{2\pi kn}{\alpha_{PCA}}\right),$$

$$F_{img}^{i,\mathcal{T}_j}[k] = -\frac{2}{\alpha_{PCA}} \cdot \sum_{n=1}^{\alpha_{PCA}-1} x_{i,\mathcal{T}_j}[n] \sin\left(\frac{2\pi kn}{\alpha_{PCA}}\right), \quad (8)$$

where $x_{i,\mathcal{T}_j}$ ($x_{i,\mathcal{T}_j} \in \mathcal{X}_|, \forall i \in [1, N]$) is the time series signal with reduced dimensions (from PCA) at bus i for scenario $\mathcal{T}_j$, the time domain index $n \in [0, \alpha_{PCA}]$, frequency domain index $k \in [0, \alpha_{PCA}/2]$, $F_{real}^{i,\mathcal{T}_j}$ and $F_{img}^{i,\mathcal{T}_j}$ are the real and imaginary parts of the transformed frequency domain signal corresponding to the original time domain signal $x_{i,\mathcal{T}_j}$ respectively. The amplitudes (10) of the transformed signals are used as the input for the K-means clustering algorithm.

$$\|F^{i,\mathcal{T}_j}[k]\| = \sqrt{\left(F_{real}^{i,\mathcal{T}_j}[k]\right)^2 + \left(F_{img}^{i,\mathcal{T}_j}[k]\right)^2} \quad (9)$$

$$\forall j \in [0, n], \forall i \in [1, N], \forall k \in [0, \alpha_{PCA}/2] \quad (10)$$

Proposition 5: *The discrete Fourier transformation of the feature-engineered time-series signal when applied like in this paper could aggregate the constraint violation information into the magnitude spectrum of the transformed frequency domain signal.*

IV. FAST SCANNING FRAMEWORK

Section III explained how to obtain the input signals for the K-means clustering algorithm that is meaningful for the DSA of LIGs. In this section, first, we provide the methodology used to obtain the optimal cluster number. Second, we present the K-means clustering algorithm that is used. Third, we provide the pseudocode for the proposed fast-scanning framework.

A. OPTIMAL NUMBER OF CLUSTERS FOR K-MEANS

To identify the optimal cluster number, we used the average Sillhouette score of the entire dataset for different values of "K" (total clusters) defined in [23]. The range of the Sillhouette score of a data point is [−1,1]. A value closer to 1 indicates that the data point is compact within its cluster and far away from other clusters.

B. K-MEANS CLUSTERING AND CLUSTER WITH CRITICAL SCENARIOS

Upon identification of the optimal number of clusters "K" using Sillhouette score from [23], the frequency domain signals from (9) are used as inputs to the K-means clustering algorithm [20]. In this paper, K-means clusters the scenarios that are similar into a single cluster and thereby groups the critical scenarios with a large number of constraint violations into one cluster.

Once the K-means algorithm converges, it returns different clusters by grouping scenarios (their data points) with similar constraint limit violations into the same cluster. However, the original objective to identify the critical scenarios can be addressed by identifying the cluster containing these critical scenarios. We found a solution based on observation. It is to compare the L^2 norm values of centers of the K-means clusters and the cluster center with the smallest L^2 norm value belonging to the critical cluster. **This is further illustrated with results in Section V.**

Proposition 6: *The real discrete Fourier transform step and its magnitude spectrum (amplitude) capture the violation information. This is further illustrated in Section V-B.*

C. PROPOSED FAST SCANNING FRAMEWORK (FAST)

The complete framework for the proposed fast scanning is presented in Algorithm 1 highlighting its different stages and functionality as described in Section III and Section IV.

V. SIMULATION RESULTS

In this section, the proposed fast scanning ML framework (FAST) is used to demonstrate its novel ML design, performance, and speed for conducting dynamic security assessment (DSA) studies on the 2028 representation of the WECC system. This section is divided into the following subsections. First, we present the information related to the generation and quality of the terabyte-scale dataset to conduct the DSA of the 2028 WECC system. Second, we present the modified time-series signals that are output from various feature engineering stages proposed in this paper and how we obtain the "*visually distinguishable signals*" for final K-means clustering. Third, we evaluate the proposed method using a dataset with a wide range of contingencies. Finally, we also evaluate the proposed FAST using a dataset containing a wide range of contingencies and operating conditions.

Algorithm 1 Fast Scanning Framework

Input : Threshold for explained variance in PCA. Dynamic simulation datasets of "n" different scenarios ($\mathcal{T}_j \forall j \in [1, n]$) = $\mathbf{D}_{\mathcal{T}_j} \in \mathbb{R}^{N \times \alpha_j}$. Where N is the total buses in the system, and α_j is the total time steps of dynamic simulation for scenario $\mathcal{T}_j$. For a given scenario $\mathcal{T}_j$, dataset $\mathbf{D}_{\mathcal{T}_j} = \forall i \forall t \mathcal{X}^j_{i,t} = \forall i \forall t \mathcal{X}^j_i(t)$ where $i \in \{1, 2, \cdots, N\}$, $t \in \{1, 2, \cdots, \alpha_j\}$, and $\mathcal{X}^j_{i,t}$ is a time series signal at bus i for scenario $\mathcal{T}_j$.

Output: Critical scenarios with largest number of voltage or frequency limit violations.

Feature engineering:

for $j = \{1, 2, \cdots, n\}$ **do** ▷ For each scenario
  for $i = \{1, 2, \cdots, N\}$ **do** ▷ For each bus.
    Stage 1:= Convert storage of $\mathcal{X}^j_i(t) \forall t \in [1, \alpha_j]$ from row to columnar; ▷ Low RAM and improved speed.
    Stage 2a:= Linear interpolation; ▷ Make datasets symmetrical and improve overall time taken for DSA.
    Stage 2b:= Normalization based on bus kV level; ▷ Remove bias on features of $\mathcal{X}^j_i(t) \forall t \in [1, \alpha_j]$.
    Stage 2c:= Detrend the time series signal; ▷ Reduce magnitude of variance in dataset.
    ▷ Stages 2a,b,c: characterize limit violation information in the time series signals $\mathcal{X}^j_i(t) \forall t \in [1, \alpha_j]$.
  end for
  Stage 3a:= Using (7), compute γ_j for $\mathbf{D}_{\mathcal{T}_j}$; ▷ γ_j = principal components (PCs) required for 99% explained variance.
end for
Stage 3b:= Compute $\alpha_{PCA} = \max\left(\gamma_1, \gamma_2, \cdots, \gamma_j\right)$; ▷ PCs required for 99% variance retention on $\mathbf{D}_{\mathcal{T}_j} \forall j \in [1, n]$.
Stage 3c:= Using α_{PCA}, project dataset from $\boldsymbol{\mathcal{D}}_{\mathcal{T}_j}$ to $\overline{\boldsymbol{D}}_{\mathcal{T}_j} \forall j = [1, n]$; ▷ $\overline{\boldsymbol{D}}_{\mathcal{T}_j} \in \mathbb{R}^{(N \times \alpha_{PCA})}$.
Stage 3d:= Flatten the projected dataset $\overline{\boldsymbol{\mathcal{D}}}_{\mathcal{T}_j}$ to $\overline{\boldsymbol{\mathcal{D}}}^{flat}_{\mathcal{T}_j} \forall j \in [1, n]$; ▷ $\overline{\boldsymbol{D}}^{flat}_{\mathcal{T}_j} \in \mathbb{R}^{(N * \alpha_{PCA} \times 1)}$.
▷ Stages3a,b,c,d: Convert several datasets with different shapes into same shape. Retain variance of datasets. Formulate vector representations of scenarios i.e., data points for K-means clustering.
Stage 4:= Convert data points in time domain ($\overline{\boldsymbol{D}}^{flat}_{\mathcal{T}_j}$) to frequency domain ($\|F^{i,\mathcal{T}_j}[k]\| \forall i \in [1, N], \forall k \in [0, \alpha_{PCA}/2]$);
▷ Aggregate limit violation information into visually distinguishable signals using real DFT.

Clustering:

Stage 5: Using Silhouette score and all scenarios' data points ($\|F^{i,\mathcal{T}_j}[k]\|$), compute the optimal cluster number;
Stage 6: Perform K-means clustering;
Stage 7: Identify the critical cluster (i.e., cluster containing critical scenarios) by comparing the L^2 norm of all cluster centers and selecting the smallest valued one;
return Critical cluster number and critical scenarios in it;

A. TERABYTE-SCALE POWER SYSTEM DATASET FOR DSA OF FUTURE LARGE INTERCONNECTED GRIDS

Data Generation: The case studies in this paper are demonstrated on the 2028 representation of the WECC 22,883 bus system. The data required for the DSA of the 2028 WECC system is generated by using the modeling framework presented in Fig. 1. As described in Sections II-A, an hourly chronological model (with unit commitment, economics, and planned topology updates) of the 2028 WECC system is developed using PCM, this PCM information is converted into converged AC power flow (ACPF) cases using C-PAGE [10], and finally, these ACPF cases are used in conjunction with DCAT [18] to perform dynamic simulations that capture cascading behavior. **Data Information**: The WECC studies in this paper includes 4,455 unique scenarios containing a wide range of operating conditions (for the year 2028) and a wide range of contingencies. These 4,455 scenarios are presented in Tab. 1 and the second column in this table identifies whether the dataset contains varying operating conditions (datasets 1-7) or varying contingencies (dataset 8). The third column in Tab. 1 provides the size of the different datasets which sums to 1.484 Terabytes. *The operating conditions are of hourly resolution corresponding to the 2028 representation of the WECC system. The operating conditions have renewable (wind and solar) generation as high as 50% of the total generation in the system. Individually, the solar and wind penetrations were as high as 38% and 18% respectively. The (mean, standard deviation) of solar and wind generation during the year 2028 are (12934 MW, 15145 MW) and (7618 MW, 3009 MW) respectively.* In this paper, to demonstrate the performance and speed characteristics of the proposed FAST framework, datasets 1, 2, and 8 are selected. As shown in Tab. 2, dataset 8 is used for case study 1 (wide range of contingencies), and datasets 1,2,8 (wide range of operating conditions and contingencies) are used for case study 2. **This validation of the proposed ML framework on the WECC system is our second contribution.**

B. VARIOUS STAGES IN PROPOSED METHOD

In this subsection, we use the case study data from Tab. 2 and present how the dynamic time-series signals are transformed using the proposed strategically layered feature engineering framework. For illustration in this subsection, we select case

study 1 (dataset 8) with varying contingencies from Tab. 1. The objective is to identify the critical scenarios (from the 213 scenarios) that have the largest number of voltage violations (as defined in equation (1)) at different kV levels of the system. In this subsection, we present the results for the time series bus signals (voltage magnitude) whose nominal voltages are between 115 kV and 230 kV.

TABLE 1. Dataset information for DSA of 2028 WECC system. The total size of the dataset with the wide range of operating conditions (datasets 1-7) is 1.36 TB and the wide range of contingencies (dataset 8) is 124 GB.

Dataset ID	No. of unique scenarios (operating hour× contingency)	Size of raw data (GB) in .chf format
1	604 (604 × 1)	193
2	648 (648 × 1)	209
3	601 (601 × 1)	194
4	583 (583 × 1)	186
5	598 (598 × 1)	191
6	606 (606 × 1)	195
7	602 (602 × 1)	198
8	213 (1 × 213)	124

1) STAGE 1: STORAGE CONVERSION

As described in Section III-B, the issue of large RAM and slow reading speeds of the dataset is addressed by converting the storage format from row to columnar. Tab. 2 presents the reduction in datasets' sizes due to this operation and we observed that the reading speeds of the dataset have significantly improved.

TABLE 2. Case studies used for evaluating FAST framework and stage 1's impact on size (from Algorithm 1).

Case studies (dataset IDs)	# scenarios/samples	Raw data size (GB)	After stage 1, size in GB (% reduction)
Case study 1 (8)	213	191	133 (37.5 %)
Case study 2 (1,2)	1252	414	335 (19 %)

2) STAGE 2: INTERPOLATE, NORMALIZE, AND DETREND

Once the dataset is loaded into the compiler, as described in Section III-C, we performed stage 2 steps presented in Algorithm 1 for better clustering results that solve the objective of the problem. In the interest of space, these results are omitted in the initial submission.

3) STAGE 3: PRINCIPAL COMPONENT ANALYSIS

After stage 2, as described in Section III-D, the optimal number of principal components (PCs) required to retain 99% of the explained variance is computed using (7). The proposed methodology from stage 3c in Algorithm 1 is used to project the dataset into lower dimensions. This methodology helped to reduce the length of the time series signals significantly which allowed for practical implementation of K-means clustering and achieving convergence in the later stages of the FAST.

4) STAGE 4: DISCRETE FOURIER TRANSFORM AND FLATTENING

After stage 3, due to the problems highlighted in Section III-E, the low dimensional time domain datasets are transformed into low dimensional frequency domain datasets. The key advantages of this transformation are 1) aggregate the constraint violation information via the magnitude (amplitude) spectrum of DFT, and 2) obtain a common reference point at regular periods.

After frequency domain conversion, using the approach described in Section. III-C, a representative data point for each scenario is obtained by horizontally concatenating (flattening) all the frequency domain bus signals into a vector. Fig. 4 presents the transformed signals after application of DFT (8), magnitude spectrum (9), and flattening steps. From Fig. 4, it can be observed that the smaller L^2 norm-valued data points contain more voltage limit violations. **This visual difference between the representative data points (from voltage/frequency limit violations perspective) of the scenarios helped to identify critical scenarios using clustering. This is our third contribution.** *Section V-C further discusses how these magnitudes spikes in Fig. 4 contribute to the identification of the cluster with critical scenarios.*

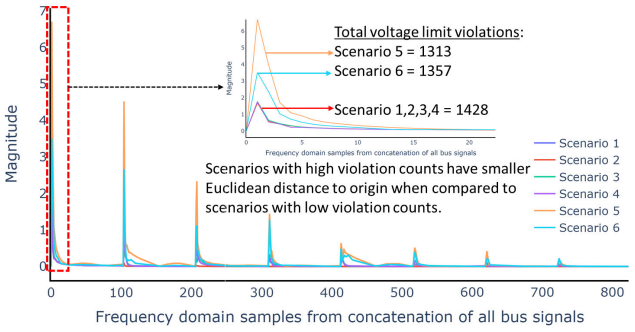

FIGURE 4. Visually distinguishable signals for different scenarios in the frequency domain.

5) STAGE 5: OPTIMAL CLUSTER NUMBER IDENTIFICATION

After stage 4, we obtain the data points (vectors) representing different scenarios that can be used as inputs to the K-means clustering algorithm. As discussed in Section IV-A, we used Silhouette score [23] to identify the optimal cluster number.

Using the representative scenarios' data points from stage 4 and the optimal clustering number from stage 5, we can implement the K-means clustering algorithm on case study 1's dataset to identify the critical scenarios with more total voltage limit violations. This is presented in Section V-C.

C. CLUSTER ANALYSIS

This subsection presents the K-means clustering results for case study 1 (wide range of contingencies) which aims to

identify the critical scenarios based on total voltage and frequency limit violations. First, we present the ground truth (brute force) results for the ranking of critical scenarios based on the total maximum voltage limit violations. Second, we present the results of identified critical scenarios.

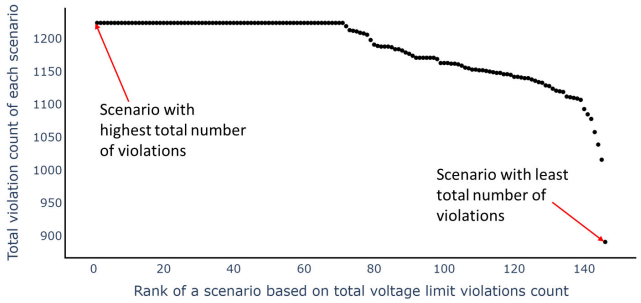

FIGURE 5. Time series scanning results: rank of the scenarios based on total violation counts versus their total violations.

1) CRITICAL SCENARIOS USING TIME SERIES SCANNING

The ground truth about the critical scenario is computed using the time series scanning (brute force) approach described in (1). From Fig. 5, the x-axis indicates the ranking of different scenarios (data points) and the y-axis represents their respective total violation counts. The scenarios' data points whose x-coordinate is close to zero are more critical when compared to other scenarios' data points with larger x-coordinate values.

2) CLUSTER ANALYSIS: CRITICAL SCENARIOS USING FAST

Here, first, we will present the proposed FAST results and compare them with the time series scanning approach. Second, we provide our explanation of how the FAST framework makes the input scenario data points visually distinguishable for optimal usage of K-means clustering. Third, we will discuss the identification of the critical cluster using the approach described in Section IV-B. Fourth, we present the efficacy of the proposed fast scanning method. Finally, we also compare the computational time taken for proposed and time series scanning methods.

1) Cluster results of FAST: Fig. 6 presents the K-means clustering results on dataset 8 with 213 scenarios. It can be observed from Fig. 6 that all the top critical scenarios are grouped into one cluster (cluster ID = 1).

2) Assigned clusters to data points (scenarios): Due to the proposed strategically layered feature engineering techniques, there is good explainable reasoning behind this clustering behavior. To illustrate this behavior, we present Fig. 7 which shows all the data points that are input to the K-means clustering algorithm during the training phase. Fig. 7 clearly shows a visual distinction (magnitude-wise) between different scenario data points of different clusters. This is a characteristic that aggregates the violation information as shown in Fig. 4.

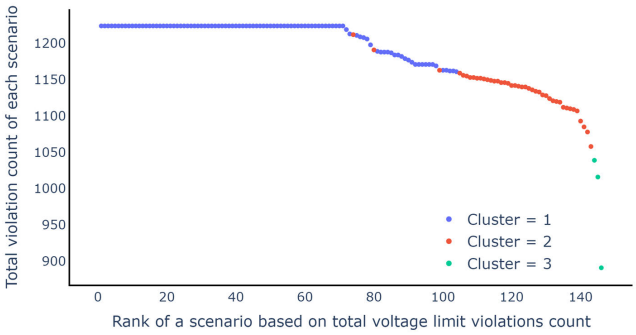

FIGURE 6. FAST results: clustering results for identification of critical scenarios. It can be observed that the top critical scenarios are grouped into one cluster (cluster ID = 1).

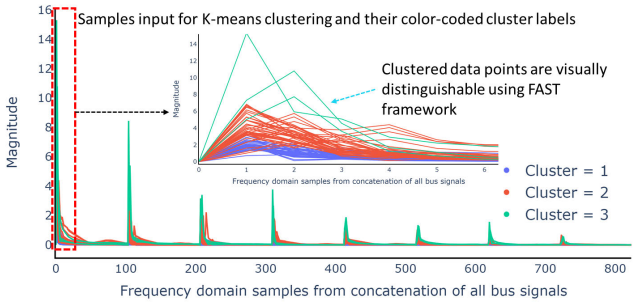

FIGURE 7. All scenarios' data points color-coded with their assigned cluster label. All clusters' data points are visually distinguishable and follow the trend explained in Fig. 4.

3) Identification of critical cluster We validated the proposed ML approach by using the time series scanning ranking (comparing Fig. 5 and Fig. 6). We identified that cluster 1 contains the critical scenarios. However, when the brute force (time series scanning ranking) results are not available then we do not know which cluster among the 3 clusters has the critical scenarios information. We solve this by using the methodology described in Section IV-B, Fig. 8 presents the centers of different clusters. It can be observed from Fig. 8 that the L^2 norm values of cluster centers as follows: $L^2(O_1) < L^2(O_2) < L^2(O_3)$. The Center of cluster 1 has the smallest L^2 norm value. We also observed from Fig. 4 that the smaller L^2 norm-valued data points contain more limit violations. Following the above observation, Fig. 7 shows all the data points from the training dataset and the data points with the most limit violation counts having smaller L^2 norm values and being grouped into cluster 1. Therefore, cluster 1 is the cluster containing the most critical scenarios, *this is also verified by manual comparison with brute force results.* Thus, the L^2 norm values of the cluster centers can help to identify the cluster with critical scenarios without needing any knowledge of the ground truth of the dataset.

D. PERFORMANCE AND SPEED FOR STUDIES 1 AND 2

In this subsection, we compare the performance and speed of the proposed framework and time series scanning (brute force) methods. The training testing datasets are split using

TABLE 3. Case studies 1 and 2: training and testing accuracy of the critical cluster for voltage and frequency limit violations criteria. The accuracy is the percentage of predicted critical scenarios that match with the top 40 critical scenarios.

Criteria	Case study ID	≤110kV		115 kV		115 to 230 kV		230 to 345 kV		345 to 500 kV	
		Train %	Test %	Train %	Test %	Train %	Test %	Train %	Test %	Train %	Test %
Total voltage limit violations	1	97.5	97.5	90	100	100	100	75(12/16)[1]	66(2/3)[1]	100(4/4)[1]	33(1/3)[1]
Total voltage limit violations	2	100	100	100	100	100	100	100	100	100	100
Total freq. limit violations	1	100	100	100	100	95	71(10/14)[1]	100	100	100	100
Total freq. limit violations	2	100	100	100	100	100	100	100	100	100	100

Cluster accuracy = (correctly predicted critical scenarios from top 40 actual critical scenarios)/40; except for cells with superscript [1]. For the cells with superscript [1], the dataset has very few to no constraint violations. Due to this class imbalance in the testing dataset, accuracy is impacted which is given by (correctly predicted critical scenarios)/(total scenarios with non-zero violations). This is explained further in Section V-D.

the 70-30 rule. The datasets of case studies 1 and 2 have 213 and 1252 scenarios/samples respectively. To evaluate the performance of the proposed method, we calculated cluster accuracy. The cluster accuracy is defined as the ratio between the number of scenarios from identified critical clusters that belong to the top 40 actual critical scenarios and 40. This means that an incorrect classification of one critical scenario from the top 40 critical scenarios introduces an error of 2.5% (1/40 * 100).

1) PERFORMANCE RESULTS

Case study 1: *case study 1 includes the dataset with scenarios containing a wide range of contingencies.* The original dataset (ID = 8) is separated into 5 sub-datasets by selecting the bus signals whose nominal kV matched the defined five kV levels i.e., 0-110kV, 115 kV, 115-230kV, 230-345kV, 345-500kV. This kV level aggregation helped to identify the critical scenarios for each kV level separately. Case study 2: *case study 2 includes the dataset with scenarios containing a wide range of contingencies and operating conditions.* Similar to case study 1 above, the original dataset (ID=1&2) is also separated into five datasets.

Tab. 3 presents the cluster accuracy of the identified critical cluster for different constraint violation criteria at different kV levels. From Tab. 3, the 100% cluster accuracy in case study 2 indicates that the proposed method can predict all the top 40 actual critical scenarios correctly (group them inside the critical cluster) in both training and testing datasets at different kV levels (this is our fourth contribution). It can also be observed that the performance in case study 2 is better than that of case study 1 for both voltage and frequency limit violation criteria. This is because, at certain kV levels, the corresponding time-series bus signals for the majority of the 213 scenarios have zero total violations which created a class imbalance in the training and testing datasets. Since there were not many scenarios/samples with non-zero violations, the accuracy calculation is heavily impacted by even a single incorrect prediction. Specifically, if there are only 10 data samples/scenarios with constraint violations then missing to identify one scenario results in a 10% error. For example, in Tab. 3, in case of voltage limit violation criteria and case study ID = 1 shows this phenomenon for the two kV levels 230-345 kV, and 345-500kV (data with superscript[1]). A similar observation is made for the frequency limit violation criteria (case study 1) at the 115-230 kV level. The above

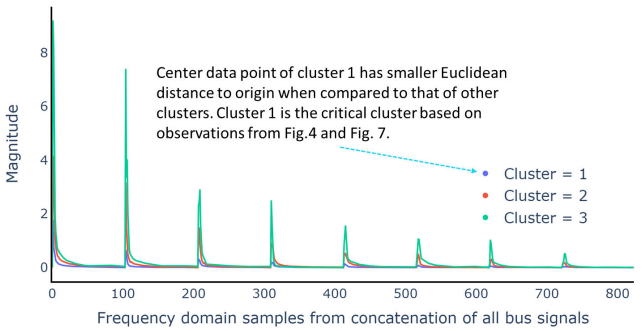

FIGURE 8. Centers of different clusters.

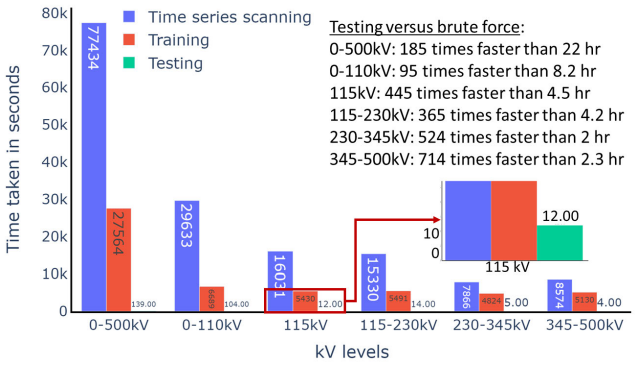

FIGURE 9. Time taken for time series scanning, and proposed method's training/testing for freq. limit violation criteria.

reason did not impact case study 2's result because it does not have a class imbalance issue like case study 1. This is because case study 2's dataset has 1252 scenarios whereas case study 1's dataset has only 213 scenarios.

2) SPEED RESULTS

Here, we present the time taken for the proposed method and the brute force approach (time series scanning) for the case of frequency limit violation criteria and case study 2. Similar times were observed for other simulations as well. The simulations are performed on a windows server machine with an Intel Xeon processor (24 cores) running at 2.4 GHz using 200 GB RAM. Fig. 9 shows the total time taken for the training and testing of the proposed method, and the brute force time series scanning method. It can be observed that the proposed method helps quickly scan through large volumes of power system dynamic data to identify the critical scenarios

based on voltage and frequency limit violation criteria. **The advantage in speed from Fig. 9 and accuracy from Tab. 3 are our fourth contribution.**

VI. CONCLUSION

This paper proposes a simulation framework for accurate planning (dynamic) studies of future large interconnected grids with high renewable penetrations where the existing planning tools are no longer sufficient. The proposed simulation framework performs dynamic simulation studies considering production cost models, AC power flow, and cascading behavior of the 2028 WECC system. This simulation framework is used to generate Terabyte-scale time-series power system dynamic simulation datasets. A new machine learning-based methodology (FAST) is proposed in this paper to perform a fast dynamic security assessment by identifying the critical scenarios based on total voltage and frequency limit violations. The proposed FAST methodology not only has great performance and speed advantages when compared to the traditional time series scanning approach but also has explainable characteristics behind its behavior. This paper uses the FAST framework to quickly identify the critical scenarios in the 2028 WECC system considering a wide range of operating conditions and contingencies.

VII. FUTURE WORK

The future scope of this work involves extending the proposed methodology to identify critical scenarios based on the rate of change of frequency, thermal line limits, and voltage recovery. The current methodology will be further improved to handle the exception case i.e. when the dataset has some scenarios with only maximum limit type violations and other scenarios that contain only minimum limit type violations.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ali Ghassemian from DOE-OE for his continuing support, help, and guidance.

REFERENCES

[1] K. Sun, S. Likhate, V. Vittal, V. S. Kolluri, and S. Mandal, "An online dynamic security assessment scheme using phasor measurements and decision trees," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1935–1943, Nov. 2007.

[2] J. Bebic, "Power system planning: Emerging practices suitable for evaluating the impact of high-penetration photovoltaics," Nat. Renew. Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/SR-581-42297, 2008. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=system+planning%3A+Emerging+practices+suitable+for+evaluating+the+impact+of+high-penetration+photovoltaics%2C%E2%80%99%E2%80%99&btnG=#d=gs_cit&t=1674467969419&u=%2Fscholar%3Fq%3Dinfo%3ArQhSFDdV4UUJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

[3] J. Cochran et al., "Flexibility in 21st century power systems," Nat. Renew. Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-6A20-61721, 2014. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=J.+Cochran%2C+M.+Miller%2C+O.+Zinaman%2C+M.+Milligan%2C+D.+Arent%2C+and+B.+Palmintier%2C+%E2%80%98%E2%80%98Flexibility+in+21st+century+power+systems%2C%E2%80%99%E2%80%99&btnG=#d=gs_cit&t=1674468240517&u=%2Fscholar%3Fq%3Dinfo%3AMSa_2WBGimQJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

[4] J. D. Jenkins and S. Thernstrom, "Deep decarbonization of the electric power sector insights from recent literature," Energy Innovation Reform Project, Fairfax, VA, USA, 2017. [Online]. Available: https://www.innovationreform.org/contact-us/ and https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=Deep+decarbonization+of+the+electric+power+sector+insights+from+recent+literatur&btnG=#d=gs_cit&t=1674468955033&u=%2Fscholar%3Fq%3Dinfo%3ARc9_Mm7gMj0J%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

[5] J. Foster et al., "Delivering a competitive Australian power system Part 2: The challenges, the scenarios," School Econ., Univ. Queensland, Brisbane, QLD, Australia, Tech. Rep. 1-2013, 2013. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=Delivering+a+competitive+Australian+power+system+part+2%3A+The+challenges%2C+the+scenarios&btnG=#d=gs_cit&t=1674468383944&u=%2Fscholar%3Fq%3Dinfo%3A4rj5HJWwk7wJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

[6] G. Sanchis, "E-highway2050: Europe's future secure and sustainable electricity infrastructure. project results," Eur. Union, Brussels, Belgium, Tech. Rep., 2015. [Online]. Available: http://www.pfbach.dk/firma_pfb/e_highway2050_booklet.pdf

[7] ISO New England Inc. (2019). *Transmission Planning Technical Guide Revision: 5.0*, System Planning. [Online]. Available: https://www.iso-ne.com/static-assets/documents/2019/10/transmission_plannings_techincal_guide_rev5.pdf

[8] WECC. (2021). *Data Preparation Manual: For Interconnection-Wide Cases*. [Online]. Available: https://www.wecc.org/Reliability/2021%20DPM.pdf

[9] K. P. Guddanti, Y. Weng, and B. Zhang, "A matrix-inversion-free fixed-point method for distributed power flow analysis," *IEEE Trans. Power Syst.*, vol. 37, no. 1, pp. 653–665, Jan. 2022.

[10] B. Vyakaranam, Q. H. Nguyen, T. B. Nguyen, N. A. Samaan, and R. Huang, "Automated tool to create chronological AC power flow cases for large interconnected systems," *IEEE Open Access J. Power Energy*, vol. 8, pp. 166–174, 2021.

[11] A. K. Bharati and V. Ajjarapu, "SMTD co-simulation framework with HELICS for future-grid analysis and synthetic measurement-data generation," *IEEE Trans. Ind. Appl.*, vol. 58, no. 1, pp. 131–141, Jan. 2022.

[12] P. Hearps, M. Wright, and B. Z. Emissions, "Australian sustainable energy: Zero carbon Australia stationary energy plan," Austral. Sustain. Energy, Zero Carbon Aust., Univ. Melbourne Energy Res. Inst., Melbourne, VIC, Australia, 2010. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=5%2C48&sciodt=0%2C48&cites=2418664146855236983&scipsc=&q=M.+Wright+and+P.+Hearps%2C+%E2%80%9CAustralian+sustainable+energy%3A+Zero+carbon+Australia+stationary+energy+plan%2C%E2%80%9D+Energy+Res.+Inst.%2C+Univ.+Melbourne%2C+Parkville%2C+Australia%2C+Tech.+Rep.%2C+2010&btnG=#d=gs_cit&t=1674469898177&u=%2Fscholar%3Fq%3Dinfo%3AMD_NNMLcGfoJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

[13] B. Elliston, I. Macgill, and M. Diesendorf, "Least cost 100% renewable electricity scenarios in the Australian national electricity market," *Energy Policy*, vol. 59, pp. 270–282, Aug. 2013.

[14] C. Budischak, D. Sewell, H. Thomson, L. Mach, D. E. Veron, and W. Kempton, "Cost-minimized combinations of wind power, solar power and electrochemical storage, powering the grid up to 99.9% of the time," *J. Power Sources*, vol. 225, pp. 60–74, Mar. 2013.

[15] E. Vittal, M. O'Malley, and A. Keane, "A steady-state voltage stability analysis of power systems with high penetrations of wind," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 433–442, Feb. 2010.

[16] R. Liu, G. Verbic, J. Ma, and D. J. Hill, "Fast stability scanning for future grid scenario analysis," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 514–524, Jan. 2018.

[17] WECC. (2021). *Anchor Data Set (ADS) Production Cost Model (PCM)*. [Online]. Available: https://www.wecc.org/SystemStabilityPlanning/Pages/AnchorDataSet.aspx

[18] B. G. Vyakaranam, A. N. Samaan, L. Xinya, H. Renke, Y. Chen, M. R. Vallem, T. B. Nguyen, A. Tbaileh, A. M. Elizondo, F. Xiaoyuan, and S. H. Davis, "Dynamic contingency analysis tool 2.0 user manual with test system examples," Pacific Northwest Nat. Lab. (PNNL), Richland, WA, USA, Tech. Rep. PNNL-29105, 2019.

[19] B. Vyakaranam, P. V. Etingov, H. Wang, X. Li, M. A. Elizondo, D. V. Zarzhitsky, N. A. Samaan, A. Tbaileh, and U. Agrawal, "Database management module framework for dynamic contingency analysis and visualization," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, 2020, pp. 1–5.

[20] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, p. 100, 1979.
[21] R. Mukerji, W. J. Burke, H. M. Merrill, and B. Lovell, "Creating data bases for power systems planning using high order linear interpolation," *IEEE Trans. Power Syst.*, vol. PS-3, no. 4, pp. 1699–1705, Nov. 1988.
[22] J. P. Theiler and G. Gisler, "Contiguity-enhanced K-means clustering algorithm for unsupervised multispectral image segmentation," in *Proc. SPIE*, Oct. 1997, pp. 108–118.
[23] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2020, pp. 747–748.

KISHAN PRUDHVI GUDDANTI (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2019 and 2021, respectively. His team is one of the winning teams of the L2RPN AI competition organized by RTE France, in 2019. He is currently working as a Power Systems Research Engineer with the Pacific Northwest National Laboratory, USA. His current research interests include the interdisciplinary area of AI applications in power systems in addition to voltage stability, and data-driven techniques for power system risk assessment and control.

BHARAT VYAKARANAM (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Cleveland State University, OH, USA, in 2011. He is currently a Senior Research Engineer with the Pacific Northwest National Laboratory, leading research activities in the area of power system modeling, dynamic simulation analysis and control for stability, and integration of renewable energy in power grid studies and machine learning applications to power systems. He is also a Registered Professional Engineer (PE) in Washington.

KAVERI MAHAPATRA (Member, IEEE) received the B.Tech. degree in electrical engineering from KIIT University, in 2011, the M.Tech. degree in electrical engineering from SOA University, India, in 2013, and the Ph.D. degree in electrical engineering from Pennsylvania State University, USA, in 2020. In 2019, she worked with the General Electric Global Research Center, Niskayuna, NY, USA, as a Research Fellow Intern. She is currently working as a Power Systems Research Engineer with the Pacific Northwest National Laboratory, Richland, WA, USA. Her current research interests include wide area monitoring, protection and control, power system dynamics and resilience, cyber physical security, machine learning, data analytics, and optimization.

ZHANGSHUAN HOU (Member, IEEE) is currently the Chief Data Scientist and the Team Lead of the Earth System Data Science Team, PNNL. He is also a pioneer and the leader in developing and applying advanced machine learning (ML), uncertainty quantification (UQ), and extreme event analysis approaches. He has blended his data sciences expertise with an educational background in science and engineering to uniquely act as the data scientist and a domain expert. As such, he and his team have developed new ML, UQ, and big data analytics approaches and methods that have significantly advanced fields of research in earth systems, energy systems, and environmental systems. These cross-cutting advances have been broadly applied to land-atmosphere modeling, stochastic operation and planning of power systems, risk assessment and extreme event analysis in subsurface and surface flow phenomena, and carbon sequestration, energy exploration, and environmental remediation.

PAVEL ETINGOV (Senior Member, IEEE) received the degree (Hons.) in electrical engineering from Irkutsk State Technical University, in 1997, and the Ph.D. degree from the Energy Systems Institute of the Russian Academy of Sciences, Irkutsk, Russia, in 2003. He is currently a Staff Research Engineer with the Pacific Northwest National Laboratory (PNNL), U.S. Department of Energy, Richland, WA, USA. He joined PNNL, in 2008, where he serves as the Project Manager, a Principal Investigator (PI)/the Co-PI, and a Key Technical Contributor in multiple projects. His research interests include stability analysis of electric power systems, power system operation, modeling and control, phasor measurement units (PMUs) application, wind and solar power generation, application of artificial intelligence to power systems, and software development.

NADER SAMAAN (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Texas A&M University, College Station, in 2004. He has been a Chief Power Systems Research Engineer and the Team Lead with the Pacific Northwest National Laboratory, Richland, WA, USA, since 2009. Prior to that, he was a Power Systems Engineer with EnerNex Corporation for four years. He was a Visiting Assistant Professor with the Department of Electrical and Computer Engineering, Kansas State University, from 2004 to 2005. His research interests include renewables integration, transmission planning, and cascading outage analysis. He is also a Registered Professional Engineer in the State of Ohio.

TONY NGUYEN (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana–Champaign, in 1998, 1999, and 2002, respectively. He joined the Pacific Northwest National Laboratory, in 2002. He is the coauthor of more than 60 publications, including journal articles, conference proceedings, book chapters, and technical reports. His research interests include power and energy systems, such as operation and control, dynamics and stability, renewable energy, system modeling and simulation, distributed energy resources, application software development, plug-in hybrid electric vehicles, and energy storage.

QUAN NGUYEN (Member, IEEE) received the B.E. degree in electrical engineering from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2012, and the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Austin, Austin, TX, USA, in 2016 and 2019, respectively. He is currently a Power System Engineer with the Pacific Northwest National Laboratory. His research interests include transmission and distribution planning and operation, such as production cost modeling, control, optimization, and simulation, renewable energy integration, power quality, and applications of power electronics in power systems.