






# Data-Driven Event Identification in the U.S. Power Systems Based on 2D-OLPP and RUSBoosted Trees

Shengyuan Liu , Shutang You , *Member, IEEE*, Zhenzhi Lin , *Member, IEEE*, Chujie Zeng, Hongyu Li, Weikang Wang , Xuetao Hu, and Yilu Liu , *Fellow, IEEE*

## I. INTRODUCTION

**Abstract**—Accurate event identification is an essential part of situation awareness ability for power system operators. Therefore, this work proposes an integrated event identification algorithm for power systems. First, to obtain and filter suitable inputs for event identification, an event detection trigger based on the rate of change of frequency (RoCoF) is presented. Then, the wave arrival time difference-based triangulation method considering the anisotropy of wave propagation speed is utilized to estimate the location of the detected event. Next, the two-dimensional orthogonal locality preserving projection (2D-OLPP)-based method, which is suitable for multiple types of measured data, is employed to achieve higher effectiveness in extracting the event features compared with traditional one-dimensional projection and principle component analysis (PCA). Finally, the random undersampling boosted (RUS-Boosted) trees-based classifier, which can mitigate the data sample imbalance issue, is utilized to identify the type of the detected event. The proposed approach is demonstrated using the actual measurement data of U.S. power systems from FNET/GridEye. Comparison results show that the proposed event identification algorithm can achieve better performance than existing approaches.

**Index Terms**—Event identification, triangulation, two-dimensional orthogonal locality preserving projection (2D-OLPP), random undersampling boosted (RUSBoosted) trees, FNET/GridEye.

vv

Manuscript received August 30, 2020; revised January 17, 2021 and May 5, 2021; accepted June 20, 2021. Date of publication June 24, 2021; date of current version December 23, 2021. This work was supported in part by the Engineering Research Center Program of the National Science Foundation and the Department of Energy under NSF Award no. EEC-1041877, the CURENT Industry Partnership Program and the NSF Cyber-Physical Systems (CPS) Program under Award no. 1931975, in part by the National Natural Science Foundation of China under Grant 52077195, and in part by the Zhejiang University Academic Award for Outstanding Doctoral Candidates under Grant 202022. Paper no. TPWRS-01481-2020. (*Corresponding Author: Zhenzhi Lin.*)

Shengyuan Liu is with the School of Electrical Engineering, Zhejiang University, Hangzhou 310027, China, and also with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA (e-mail: eelsy@zju.edu.cn).

Shutang You, Chujie Zeng, Hongyu Li, and Weikang Wang are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA (e-mail: syou3@utk.edu; czeng8@vols.utk.edu; hli90@utk.edu; wwang72@vols.utk.edu).

Zhenzhi Lin is with the School of Electrical Engineering, Zhejiang University, Hangzhou 310027, China, and also with School of Electrical Engineering, Shandong University, Jinan 250061, China (e-mail: linzhenzhi@zju.edu.cn).

Xuetao Hu is with the School of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: hxt\_hult@yeah.net).

Yilu Liu is with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA, and also with Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (e-mail: liu@utk.edu).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TPWRS.2021.3092037>.

Digital Object Identifier 10.1109/TPWRS.2021.3092037

THE dynamic response of power systems after disturbances is more and more complex due to the ever-increasing power system scale and penetration of renewable energy sources (RES) [1]. Hence, it becomes more difficult to detect, locate, and classify the events in power systems accurately and timely. In the past, power system operators relied on field crews to identify grid events according to operation records and conduct the corresponding post-analysis thereafter. However, event detection and identification following in this way have certain limitations, since real-time alerts cannot be sent to system operators and control measures cannot be carried out in a short time. On the one hand, the rapid detection and accurate localization and identification of power system events can greatly enhance the situation awareness capability for operators and help them take timely correction measures. On the other hand, the wide-area measurement system (WAMS) and advanced metering infrastructure (AMI) have been deployed widely in modern power systems and smart grids to enhance the situation awareness ability of power system operators [2], [3]. Besides, the synchrophasor measurement devices (SMDs) such as phasor measurement units (PMUs) [4], frequency disturbance recorders (FDRs) and universe grid analyzers (UGAs) [5] have also been installed in many countries, which enable detecting and identifying power system events in practice. Therefore, it is feasible and also of great significance to study event identification for actual power systems based on wide-area measurements.

Recently, several studies have focused on this issue. They can be divided into two categories: i) event detection that aims to discover and locate disturbances in power systems; and ii) event classification targeting at classifying what a disturbance is. In [6], voltage amplitude data is analyzed by the discrete wavelet transform (DWT) and Kalman filter to detect on-line voltage events. In [7], voltage imbalances are detected by PMU data with a downsampled negative sequence. In [8], offline coherency identification of power systems is performed first and the event is detected and located by monitoring the change of cluster centers of voltage and frequency information. In [9], geometric template matching and quality threshold clustering are utilized to measure the similarly for dynamic responses of frequency in different locations, and radial basis function artificial neural network (RBF-ANN) is employed to estimate other inaccessible measurements. In [10], DWT is employed for PMU data and normalized wavelet energy (NWE) is defined to

detect events and indicate their locations. To address the problem of huge communication burden within event detection, several data reduction methods are proposed to compress data first to decrease the time delay of event detection on the data collection stage. In [11], the least square-based curve fitting method is used to compress PMU data. In [12], principle component analysis (PCA) is employed to reduce the scale of PMU data. Then a real-time event indicator whose threshold is given in advance is used to determine whether there is an event in power systems. In [13], dynamic programming is embedded in swing door trending (SDT) to reduce the amount of streaming PMU data while magnitude and slope rules are employed for event detection. In [14], an unequal-interval reduction and reconstruction method that can obtain higher accuracy for active and reactive power data are proposed, and local outlier factor (LOF) is utilized to detect and locate events in power systems. In [15], the  $T^2$  and  $Q$  statistics are calculated in the process of PCA to detect load generation mismatch and islanding events, respectively. On this basis, moving window PCA is utilized in [16] for multiple event detection. Besides, the  $k$ -nearest neighbor (KNN) method is used for  $T^2$  and  $Q$  statistics in [17] to acquire two new monitoring statistics for faster event detection. In [18], the dynamic response of power systems is quantified by recurrence quantification analysis (RQA), and two quantitative measures of recurrence (i.e., recurrence rate and determinism) are used for data noise reduction and event detection and localization.

In [19], minimum-volume-enclosing ellipsoid (MVEE) is constructed based on voltage, frequency and power flow data, and clustering methods are used for event identification. In [20], the distances between the events to be studied and the pre-defined event dictionary is calculated by the angle between their subspaces, which are used for further classification analysis. In [21], time-frequency analysis and fast discrete S-transform (FDST) are utilized to select several features, and extreme machine learning is employed for classifying events. In [22], a diffusion type kernel density estimator (DKDE) and deep neural network are used for real-time event identification while the impact on identification accuracy due to the fluctuation and penetration level of RES are also studied. However, the aforementioned algorithms are suitable for single event identification only, which limits their application in practical situations. Therefore, some latest studies aim to achieve multiple event detection in power systems. In [23], multiple events are regarded as the combination of several typical “root events” recorded in frequency monitoring network (FNET) [24] and a sparse unmixing method is proposed to distinguish them. Furthermore, this method is boosted by cluster-based sparse coding in [25] to achieve better performance. In [26], the dominant eigenvalues and singular values are extracted as input features fed into the convolutional neural network (CNN) and the previous events are predicted and subtracted from the real-time measured data so as to identify events accurately. In [27], a Teager–Kaiser energy operator (TKEO)-based method is introduced for event detection and the 1-nearest neighbor (1NN) classifier with energy similarity measure (ESM) that has superposition property is employed together to identify multiple events in power systems.

It can be seen that almost all existing studies do not consider the time and location of events in the process of event identification. In fact, accurate time and location can help filter suitable data for event identification so as to improve its performance. Therefore, this work aims to propose an event identification with the event time and location considered, and to achieve higher event identification accuracy by using advanced machine learning algorithms. The major contributions of this work can be summarized as follows.

- 1) To the best of our knowledge, this work for the first time, proposes an event identification algorithm with the event time and location considered to improve identification accuracy for actual power systems. Furthermore, with the help of FNET/GridEye, the proposed event identification algorithm can be deployed easily and achieve real-time situational awareness for power systems.
- 2) The event time determination algorithm based on the rate of change of frequency (RoCoF) and the event location determination algorithm based on triangulation using the differences of wave arrival times among SMDs are presented. Compared with the traditional triangulation method, the proposed algorithm further considers the anisotropy of wave propagation speed, which can obtain more accurate event locations.
- 3) The two-dimensional orthogonal locality preserving projection (2D-OLPP) is utilized to extract the inherent features of different types of events that occur in power systems. It can extract event features from the 2D perspective directly rather than unfolding the 2D-features into a vector and then using one-dimensional analysis methods. In fact, the measured data are with multiple types (i.e., frequency, voltage phase angle and amplitude), which means the original measured data are in 2D form. Therefore, the presented 2D-OLPP method is more suitable and can achieve better effectiveness.
- 4) On this basis of feature extraction by 2D-OLPP, the random undersampling boosted (RUSBoosted) trees, which can mitigate the issue of sample imbalance in confirmed event cases, are employed for event identification. Compared with other event identification algorithms, it can achieve the highest recall rate, accuracy and F1-score, which means it achieves better performance than existing algorithms.

The rest of this paper is organized as follows. The principles for event time and location determinations based on triangulation are presented in Section II; the 2D-OLPP-based event feature extraction method is proposed in Section III; the event identification classifier based on RUSBoosted trees is proposed in Section IV; case studies on actual U.S. power systems and corresponding comparisons with other algorithms are performed in Section V; some discussions and illustrations are given in Section VI; finally, the conclusions and further works are given in Section VII.

## II. EVENT TIME AND LOCATION DETERMINATIONS BASED ON TRIANGULATION UTILIZING ROCoF AND CONSIDERING CHARACTERISTICS OF WAVE PROPAGATION

Accurate event time and location determinations are the basis of event identification. Concretely, the result of event time determination is the prerequisite and trigger of event identification, and the results of event location determination can provide quite specific geographical information, which can also enhance the performance of event identification. Hence, the latest developments of event time and location determinations based on triangulation are presented in this section. It is worth mentioning that compared with the previous triangulation algorithm, the latest one replaces the voltage phase angle curves with RoCoF curves as the wave to be studied and considers the anisotropy of the speed of wave propagation so as to obtain more accurate triangulation results. Assume the frequency curve of the  $m^{\text{th}}$  SMD can be denoted as  $f_m = [f_{m,1}, f_{m,2}, \dots, f_{m,N}]$ , ( $m = 1, 2, \dots, M$ ), where  $M$  and  $N$  are the number of SMDs and the number of sample points, respectively. In fact, there will be large frequency fluctuations when events happen in power systems. Thus, the RoCoF can be used for event time determination and defined as

$$f_{m,n}^{\text{RoCoF}} = \frac{f_{m,n-r\Delta t}^{\text{RoCoF}} - f_{m,n}^{\text{RoCoF}}}{\Delta t} \quad (1)$$

where  $\Delta t$  is the time interval and  $r$  is the sampling rate of measured data. In practice, an event will be detected if

$$\sum_{m=1}^M N_m^\mu \geq \kappa M \quad (2)$$

where

$$N_m^\mu = \begin{cases} 1, & \text{if } \max_n \{|f_{m,n}^{\text{RoCoF}}|\} \geq \mu \\ 0, & \text{if } \max_n \{|f_{m,n}^{\text{RoCoF}}|\} < \mu \end{cases} \quad m = 1, 2, \dots, M \quad (3)$$

where  $\mu$  is the threshold of RoCoF for event time determination,  $N_m^\mu$  is the 0-1 variable that indicates whether the  $m^{\text{th}}$  SMD detects an event.  $\kappa$  is a predefined ratio and equation (2) means an event will be detected and declared if  $\kappa M$  SMDs from  $M$  SMDs received the frequency fluctuations. After an event is detected, more detailed analyses [28], [29] will be performed for RoCoF and phase angle curves to determine the wave-front arriving time  $t_m^{\text{wave}}$  of the  $m^{\text{th}}$  SMDs ( $m = 1, 2, \dots, M$ ). Hence, the following equation holds if the differences between electrical distance and physical distance are neglected.

$$d_m = v_m(t_m^{\text{wave}} - t^{\text{E}}) \quad m = 1, 2, \dots, M \quad (4)$$

where

$$d_m = R_E \cos^{-1} [\sin \alpha_m \sin \beta^{\text{E}} + \cos(\beta^{\text{E}} - \beta_m) \sin \alpha^{\text{E}} \sin \alpha_m] \quad (5)$$

where  $d_m$  is the distance between the  $m^{\text{th}}$  SMD and the actual location of the detected event,  $R_E$  is the radius of the earth.  $v_m$  is the wave propagation speed from the event location to the  $m^{\text{th}}$  SMD.  $\alpha_m$  and  $\beta_m$  are the latitude and longitude of the  $m^{\text{th}}$  SMD, respectively.  $\alpha^{\text{E}}$  and  $\beta^{\text{E}}$  are the latitude and longitude of the event location, and  $t^{\text{E}}$  is the time when the event happens,

respectively. In (4) and (5),  $\alpha_m$ ,  $\beta_m$  and  $t^{\text{E}}$  are the variables to be solved, and  $v_m$  and  $t_m^{\text{wave}}$  can be viewed as the known variables. Therefore, the objective function for finding the least square fitting error of event location determination can be represented as

$$\min L(\alpha^{\text{E}}, \beta^{\text{E}}, t^{\text{E}}) = \sum_{m=1}^{M'} \varepsilon_m^2 = \sum_{m=1}^{M'} [d_m - v_m(t_m - t^{\text{E}})]^2 \quad (6)$$

where  $M'$  is the number of SMDs used for event location determination. To solve (6), the partial derivative of  $L$  should equal 0 as

$$\frac{\partial L}{\partial \mathbf{x}} = 2\mathbf{J}^T \boldsymbol{\varepsilon} = \mathbf{0} \quad (7)$$

where  $\mathbf{x} = [\alpha^{\text{E}}, \beta^{\text{E}}, t^{\text{E}}]$  and  $\mathbf{J}$  is the Jacobian matrix as

$$\mathbf{J} = \begin{bmatrix} \left. \frac{\partial \varepsilon_1}{\partial \alpha^{\text{E}}} \right|_{\mathbf{x}} & \left. \frac{\partial \varepsilon_2}{\partial \alpha^{\text{E}}} \right|_{\mathbf{x}} & \dots & \left. \frac{\partial \varepsilon_{M'}}{\partial \alpha^{\text{E}}} \right|_{\mathbf{x}} \\ \left. \frac{\partial \varepsilon_1}{\partial \beta^{\text{E}}} \right|_{\mathbf{x}} & \left. \frac{\partial \varepsilon_2}{\partial \beta^{\text{E}}} \right|_{\mathbf{x}} & \dots & \left. \frac{\partial \varepsilon_{M'}}{\partial \beta^{\text{E}}} \right|_{\mathbf{x}} \\ v_1 & v_2 & \dots & v_{M'} \end{bmatrix}^T \quad (8)$$

where

$$\begin{aligned} \left. \frac{\partial \varepsilon_m}{\partial \alpha^{\text{E}}} \right|_{\mathbf{x}} &= \frac{-R[\cos \alpha^{\text{E}} \sin \alpha_m - \cos \alpha_m \cos(\beta^{\text{E}} - \beta_m) \sin \alpha^{\text{E}}]}{\sqrt{1 - [\sin \alpha^{\text{E}} \sin \alpha_m + \cos \alpha_m \cos(\beta^{\text{E}} - \beta_m) \cos \alpha^{\text{E}}]}} \quad (9) \end{aligned}$$

$$\begin{aligned} \left. \frac{\partial \varepsilon_m}{\partial \beta^{\text{E}}} \right|_{\mathbf{x}} &= \frac{-R \cos \alpha^{\text{E}} \cos \alpha_m \cos(\beta^{\text{E}} - \beta_m)}{\sqrt{1 - [\sin \alpha^{\text{E}} \sin \alpha_m + \cos \alpha_m \cos(\beta^{\text{E}} - \beta_m) \cos \alpha^{\text{E}}]}} \quad (10) \end{aligned}$$

Therefore, the  $l^{\text{th}}$  iteration can be denoted as

$$\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} + [-(\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \boldsymbol{\varepsilon}] \quad (11)$$

It is noted that the values of  $\mathbf{J}$  and  $\boldsymbol{\varepsilon}$  in (11) will change in each iteration and the iteration will be terminated when  $|(\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \boldsymbol{\varepsilon}|$  is smaller than a given threshold (e.g.,  $10^{-5}$ ). It should be noted that (11) can be calculated only the column of the  $\mathbf{J}$  is full rank, in other words, the columns of  $\mathbf{J}$  should be linearly independent and  $\mathbf{J}^T \mathbf{J}$  should be invertible. If the column of the  $\mathbf{J}$  is not full rank, on the one hand, the other alternative SMDs can be selected to ensure  $\mathbf{J}^T \mathbf{J}$  is invertible; on the other hand, Moore-Penrose pseudoinverse [30] can be utilized for solving the irreversible problem during iteration. It should be mentioned that the most important improvement of the presented event location determination algorithm is that the differences in wave propagation speeds in various directions are considered. In fact, the frequency wave is electromechanical and its speed is different for various directions. However, in the past, the values of  $v_m$  ( $m = 1, 2, \dots, M$ ) are assumed as the same one and denoted to one variable  $v$  as shown in Fig. 1(a), while now they are assigned as different values as shown in Fig. 1(b) so as

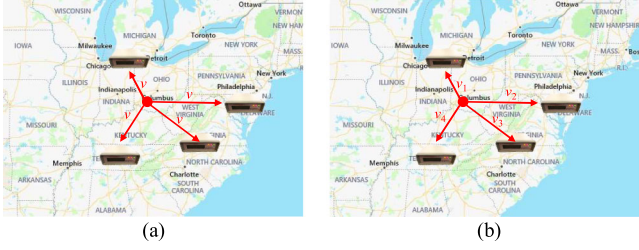


Fig. 1. Illustration of wave propagation speed. (a) Without the anisotropy of wave propagation speed considered; (b) With the anisotropy of wave propagation speed considered.

to achieve better performance for event location determination. Indeed, it is not easy to determine the accurate values of  $v_m$  ( $m = 1, 2, \dots, M$ ), but this problem can be solved by studying the past confirmed cases for estimating the wave propagation speed associated with each SMD. Although these values are not always very accurate for all cases, diversifying them would have a better performance compared with assuming that they all have the same value.

The algorithms for event time and location determinations have been deployed in one of the famous WAMSs in U.S. power systems, i.e., FNET/GridEye [31], and more illustrations can be found in Section IV.

### III. EVENT FEATURE EXTRACTION BASED ON 2D-OLPP

It is worth mentioning that feature extraction is a very essential process before using classifiers to identify the types of events that happened. Therefore, the 2D-OLPP-based method is presented in this section for extracting the event features so as to enhance the effectiveness of event identification. Locality preserving projection (LPP) is one of the manifold methods, whose main idea is to learn the local neighborhood structure of samples in high-dimensional spaces and find a space that can retain the manifold structure so that the samples can get a better local neighbor relationship after being projected to a low-dimensional space when compared with principal component analysis (PCA) or linear discriminant analysis (LDA) [32]. In detail, LPP is a linear mapping for dealing with a variational problem that aims to preserve the neighborhood structure of a given dataset optimally. LPP is a choice for the classical linear technique (i.e., PCA), which projects the data onto the directions with the largest variance. The LPP can be determined by searching the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold if high dimensional data lies on a low dimensional manifold [33]. Therefore, LPP has several data representation properties of nonlinear methods such as Locally Linear Embedding (LLE) and Laplacian Eigenmaps [33]. As for the 2D-OLPP method used in this work, it is an extended method from LPP and has the same basic principles and similar solving process with LPP. The differences between them are that 2D-OLPP can handle the two-dimensional data directly and it uses orthogonal bases when performing projection so as to achieve less computation burden.

In this work, the frequency, voltage angle and amplitude data are available and utilized for feature extraction. Assume the

measured data of the  $i^{\text{th}}$  event can be denoted as

$$\mathbf{A}_i = [\mathbf{A}_i^{\text{Fre}}, \mathbf{A}_i^{\text{Ang}}, \mathbf{A}_i^{\text{Amp}}] \quad (12)$$

where  $\mathbf{A}_i^{\text{Fre}}$ ,  $\mathbf{A}_i^{\text{Ang}}$  and  $\mathbf{A}_i^{\text{Amp}}$  are respectively the frequency, voltage angle and amplitude data of the  $i^{\text{th}}$  event measured by SMDs located in different places during the detected time period. Generally, the data of several SMDs nearest to the event location determined in Section II would be selected for feature extraction. It can be seen from (12) that the  $\mathbf{A}_i$  is a matrix rather than a vector, so the traditional LPP method is no longer feasible for extracting the event features in this case. To solve this problem, there are two solutions: i) unfold the matrix into a vector and then use the traditional LPP method; ii) use 2D-LPP method directly for the matrix. In fact, it has been illustrated in [32] that using the 2D-LPP method directly has two advantages: one is that it can reduce the computation time and achieve higher classification accuracy compared with using the traditional LPP method; another is that the computational efficiency can be further improved if using orthogonal bases in the space transformation. Therefore, the 2D-OLPP-based method is suitable for extracting the event feature in this work.

The presented 2D-OLPP-based event feature extraction method aims to minimize the weighted sum of squares of the distances between measured data samples in the low dimensional space [32] and it can be denoted as

$$\min \sum_{i=1}^M \sum_{j=1}^N (\mathbf{B}_i - \mathbf{B}_j) S_{ij} \quad (13)$$

where  $\mathbf{B}_i = \mathbf{A}_i \boldsymbol{\zeta}$  is the projection of  $\mathbf{A}_i$  using the orthogonal basis matrix  $\boldsymbol{\zeta} = [\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \dots, \boldsymbol{\zeta}_k]$ , and  $k$  is the number of orthogonal vectors that can span the new low-dimension space.  $S_{ij}$  is the element of similarity weight matrix  $\mathbf{S}$  and can be determined by Euclidean distance or Gaussian kernel function [32]. The objective function (13) can be further converted [32], [34] as

$$\begin{aligned} & \min \boldsymbol{\zeta}^T \mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \boldsymbol{\zeta} \\ & \text{s.t. } \boldsymbol{\zeta}^T \mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A} \boldsymbol{\zeta} = 1 \quad \forall i = 1, 2, \dots, k-1; k \geq 2 \\ & \quad \boldsymbol{\zeta}_i^T \boldsymbol{\zeta}_k = 0 \end{aligned} \quad (14)$$

where  $\mathbf{Q}$  is a diagonal matrix whose diagonal elements equal to the sum of corresponding columns of  $\mathbf{S}$ , i.e.,  $Q_{ii} = \sum_{j=1}^N S_{ij}$  ( $i = 1, 2, \dots, M$ ).  $\mathbf{E}$  is the identity matrix,  $\mathbf{L} = \mathbf{Q} - \mathbf{S}$ , and  $\otimes$  is the Kronecker product operator, respectively. It can be seen from (14) that the constraints guarantee the orthogonality of the bases  $\boldsymbol{\zeta}$ .

The problem (14) can be solved separately in two situations, i.e.,  $k = 1$  and  $k \geq 2$ . For the first situation (i.e.,  $k = 1$ ), the second constraint in (14) can be neglected and  $\boldsymbol{\zeta}_1$  can be determined by solving the eigenvector of (15) associated with the smallest non-zero eigenvalue  $\lambda$  [32], [34] as:

$$\boldsymbol{\zeta}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \boldsymbol{\zeta} = \lambda \mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A} \boldsymbol{\zeta} \quad (15)$$

To solve the situations of  $k \geq 2$ , i.e., to consider the second constraint in (14), Lagrangian multipliers can be involved, and

the  $\zeta_k$  can be obtained through minimizing the following Lagrangian function  $\psi^{(k)}$  [32], [34] as

$$\begin{aligned} \psi^{(k)} = & \zeta_k^T \mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \zeta_k - \sum_{i=1}^{k-1} \mu_i \zeta_k^T \zeta_i \\ & - \lambda [\zeta_k^T \mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A} \zeta_k - 1] \end{aligned} \quad (16)$$

where  $\mu_1, \mu_2, \dots, \mu_{k-1}$  are Lagrangian multipliers corresponding to the second constraint in (14). Equation (16) can be optimized by letting its partial derivatives equal to 0 as

$$\frac{\partial \psi^{(k)}}{\partial \lambda} = \zeta_k^T \mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A} \zeta_k - 1 = 0 \quad (17)$$

$$\frac{\partial \psi^{(k)}}{\partial \mu_i} = \zeta_k^T \zeta_i = 0 \quad (18)$$

$$\begin{aligned} \frac{\partial \psi^{(k)}}{\partial \zeta_k} = & 2\mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \zeta_k \\ & - 2\lambda \mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A} \zeta_k - \sum_{i=1}^{k-1} \mu_i \zeta_i = 0 \end{aligned} \quad (19)$$

It is noted that  $\zeta_k^T \zeta_i = 0$  ( $i < k$ ) since they are orthogonal bases. Therefore, utilize  $\zeta_k^T$  to multiply (19) and the  $\lambda$  can be obtained as

$$\lambda = \frac{\mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \zeta_k}{\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A} \zeta_k} \quad (20)$$

It is demonstrated in [34] that  $\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}$  is non-singular and positive definite, which means it is invertible. A series of  $k-1$  equations can be obtained from multiplying (19) by  $\zeta_i^T \mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}$  ( $i = 1, 2, \dots, k-1$ ) continuously and their compact form can be reorganized to obtain the elements of  $\mu^{(k-1)} = [\mu_1, \mu_2, \dots, \mu_{k-1}]$  as

$$\mu^{(k-1)} = 2[\sigma^{(k-1)}]^{-1} [\zeta^{(k-1)}]^T [\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \zeta_k \quad (21)$$

where

$$\zeta^{(k-1)} = [\zeta_1, \zeta_2, \dots, \zeta_{k-1}] \quad (22)$$

$$\sigma^{(k-1)} = [\zeta^{(k-1)}]^T [\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}]^{-1} \zeta^{(k-1)} \quad (23)$$

Multiply (19) by  $[\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}]^{-1}$  and replace  $\mu^{(k-1)}$  using (21), then the equation can be obtained as

$$(\mathbf{E} - \mathbf{Z}^{(k)}) [\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A} \zeta_k = \lambda \zeta_k \quad (24)$$

where

$$\mathbf{Z}^{(k)} = [\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}]^{-1} \zeta^{(k-1)} [\sigma^{(k-1)}]^{-1} [\zeta^{(k-1)}]^T \quad (25)$$

Thus, (24) can be converted into the problem of solving the eigenvector of (26) associated with the smallest non-zero eigenvalue [34] as

$$\mathbf{X}^{(k)} = (\mathbf{E} - \mathbf{Z}^{(k)}) [\mathbf{A}^T (\mathbf{Q} \otimes \mathbf{E}) \mathbf{A}]^{-1} [\mathbf{A}^T (\mathbf{L} \otimes \mathbf{E}) \mathbf{A}] \quad (26)$$

Thus, the  $\zeta = [\zeta_1, \zeta_2, \dots, \zeta_k]$  can be determined iteratively and served as the input features for event identification. It is

noted that the aforementioned  $\zeta$  is a generalized denotation, and the features of the  $j^{\text{th}}$  event can be denoted as  $\zeta^j$  in Section IV.

#### IV. EVENT IDENTIFICATION BASED ON RUSBOOSTED TREES

Once the event features are extracted in Section III, they can be fed into the classifiers to identify the types of detected events. In this work, four types of events are considered for event identification, i.e., short-circuit fault (SCF), generation trip (GT), load shedding (LS), and line trip (LT). Due to the nature of the different occurrence possibilities for these events, the recorded events that happened in actual power systems are mainly the GT, LS and LT while the number of SCF in the transmission system is limited. Therefore, an essential issue for the event identification in this work is the sample imbalance of training data [35], [36]. In such a situation, it is difficult to establish an effective classifier since the traditional data mining method would tend to predict most of the samples into the majority class. As a simple example, if there are 998 negative samples, but only 2 positive samples, one learning method can achieve a 99.8% precision rate by returning a classifier that always predicts the new sample as a negative one. Such a classifier, however, is often worthless since it does not predict any positive cases. In other words, it can achieve a high precision rate but a terrible recall rate [37]. To solve this issue in this work, on the one hand, simulated data in the simplified actual power system are also utilized in combination with measured data for classifier training; on the other hand, an advanced machine learning method (i.e., RUSBoosted trees) that can mitigate the sample imbalance problem [38] is employed in this work. This section will introduce the RUSBoosted trees-based event identification method and the details of simulated data will be introduced in Section V.

In fact, several techniques are presented to deal with the class imbalance problems, such as data sampling and boosting. For data sampling, there are two types, i.e., undersampling and oversampling. Undersampling means to remove some samples in the majority class to maintain the class balance in the training set while oversampling means to duplicate the samples in the minority class to maintain the class balance in the training set. The typical undersampling method is random undersampling (RUS) and the typical oversampling methods are synthetic minority oversampling technique (SMOTE) [39] and adaptive synthetic (ADASYN) sampling method [40]. It should be noted that either RUS, SMOTE and ADASYN has its strength and weakness. For RUS, removing samples in the majority class may cause the information loss problem; however, it greatly reduces the training time and multiple RUS can be performed to mitigate the information loss problem. For SMOTE and ADASYN, although there is no information loss problem, the over-fitting problem would be caused when using the oversampling data. Furthermore, the computation burden increases significantly since a larger training set is used [41]. For boosting, it is a technique to boost the effectiveness of classification by revising and updating the weight of each sample in the training process. If the samples with minority class are classified incorrectly, then they would be assigned with larger weights in the following training rounds.

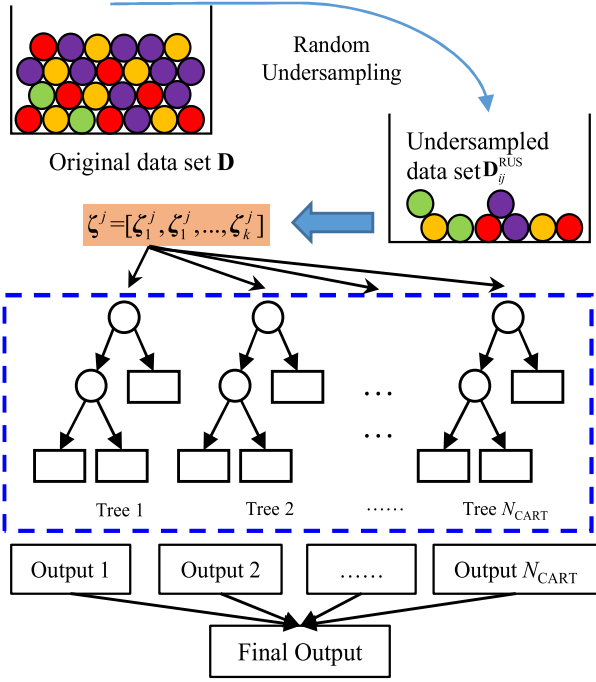


Fig. 2. Illustration of the RUSBoosted tree-based method.

Therefore, the RUSBoosted-trees presented in this work combine the techniques of data sampling and boosting to mitigate the class imbalance problem. It is worth mentioning that the tests in different datasets [38] demonstrate that the RUSboost and SMOTEBoosting can achieve similar performance (RUSboost is often better than SMOTEBoosting in fact) while RUSboost is much faster. Therefore, the authors of Ref. [38] highly recommend using the RUSBoost for improving the classification effectiveness.

The basic principle of RUSBoosted trees is two-fold: i) utilize random undersampling to mitigate sample imbalance; and ii) combine the results of several classification and regression trees (CARTs) to boost the performance of the final result. An illustration is shown in Fig. 2 and the steps of the RUSBoosted trees-based event identification method can be summarized as follows.

*Step 1:* Initiate the weight of the  $i^{\text{th}}$  CART with the  $j^{\text{th}}$  sample as  $w_{ij} = 1/N_E$  ( $i = 1, 2, \dots, N_{\text{CART}}; j = 1, 2, \dots, N_E$ ), where  $N_{\text{CART}}$  is the number of CARTs used for event identification and  $N_E$  is the number of events in original data set, respectively. Let  $i = 1$ .

*Step 2:* Utilize the random undersampling method as shown in Fig. 2 to generate  $N_{\text{set}}$  data set  $\mathbf{D}_{ij}^{\text{RUS}}$  ( $j = 1, 2, \dots, N_{\text{set}}$ ) for the  $i^{\text{th}}$  CART from the original data set  $\mathbf{D}$  whose elements are the event features extracted in Section III. Besides, obtain the weights  $w_{ij}^{\text{RUS}}$  associated with the data set  $\mathbf{D}_{ij}^{\text{RUS}}$  ( $j = 1, 2, \dots, N_{\text{set}}$ ).

*Step 3:* Train the  $i^{\text{th}}$  CART based on  $\mathbf{D}_{ij}^{\text{RUS}}$  and  $w_{ij}^{\text{RUS}}$  to obtain the  $i^{\text{th}}$  CART classifier for event identification as  $y'_{ij} = \chi_i(\zeta^j)$ ,

where  $y'_{ij}$  is the predicted type label for the  $j^{\text{th}}$  event and  $\chi_i(\cdot)$  is the function of the  $i^{\text{th}}$  classifier, respectively.

*Step 4:* Determine the prediction loss for the  $i^{\text{th}}$  CART classifier as

$$\varepsilon_i = \sum_{j=1}^M [1 - I(y_j, y'_{ij})] \quad (27)$$

where  $y_j$  is the true type label of the  $j^{\text{th}}$  event and  $I(\cdot, \cdot)$  is the indicative function.  $I(y_j, y'_{ij}) = 1$  if  $y_j = y'_{ij}$ , otherwise  $I(y_j, y'_{ij}) = 0$ .

*Step 5:* Obtain the weight update factor for the  $i^{\text{th}}$  CART classifier as (28) and update the weight for the  $(i+1)^{\text{th}}$  CART classifier and the  $j^{\text{th}}$  event sample as (29).

$$\eta_i = \varepsilon_i / (1 - \varepsilon_i) \quad (28)$$

$$w_{i+1,j} = w_{ij} \eta_i^{0.5[1-I(y_j, y'_{ij})]} / \sum_{j=1}^{N_E} w_{ij} \eta_i^{0.5[1-I(y_j, y'_{ij})]} \quad (29)$$

*Step 6:* If  $i = N_{\text{CART}}$ , go to Step 7; otherwise,  $i = i+1$  and go to Step 2.

*Step 7:* Determine the final classifier based on RUSBoosted trees by weighted vote from the  $N_{\text{CART}}$  classifiers as

$$\chi_{\text{final}}(\zeta^{\text{new}}) = \arg \max_k \sum_{i=1}^{N_{\text{CART}}} I[\chi_i(\zeta^{\text{new}}) = u] \log \frac{1}{\eta_i} \quad (30)$$

where  $u$  is the enumerable type label.

For better illustration, the flowchart of event identification is given in Fig. 3. It can be seen that there are offline and online stages in the proposed approach. In the offline stage, data preparation, parameter tuning and classifier training are completed in advance, as these steps relatively time-consuming and largely rely on historical data and confirmed events. However, in the online stage, the rapid monitoring for power system events can be achieved and corresponding information can be sent to system operators in almost real time and saved in the database for further analysis.

## V. CASE STUDIES

To demonstrate the performance of the proposed event identification algorithm, the actual measured data of the U.S. power system from FNET/GridEye are employed for case studies. Besides, the comparisons between the proposed and other event identification algorithms are given for detailed analysis.

### A. Case Studies of Event Identification in Actual U.S. Power Systems

To better illustrate the effectiveness of the proposed event identification algorithm, a confirmed generation trip event that happened in U.S. Eastern Interconnection (EI) system is taken as an example. In 2020-04-04 16:56:32 UTC, several generation units are tripped in Dominion Power-Brunswick County Power Station with a total amount of 578MW. The measured frequency,

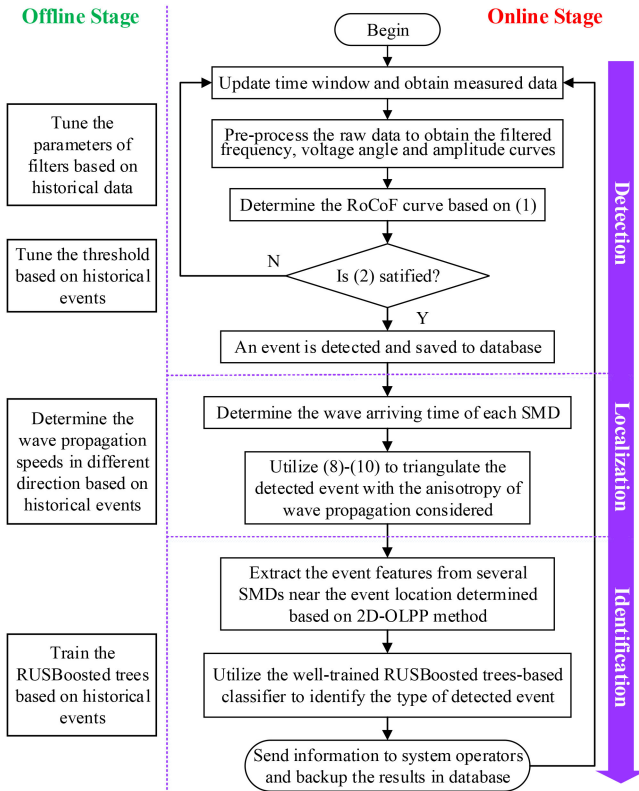


Fig. 3. Flowchart of the proposed event identification algorithm.

TABLE I  
INFORMATION OF THE FDRs WITH THE EARLIEST WAVE ARRIVAL TIME OF THE STUDIED GENERATOR TRIP EVENT

Order	FDR Name	Wave Arrival Time	Wave Propagation Speed
1	UsVaNewportnews847	12.88079s	1042.20km/s
2	UsNcClayton1060	13.03242s	1031.05km/s
3	UsNcRaleigh891	13.03520s	981.18km/s
4	UsVaRichmond601	13.07382s	959.90km/s
5	UsSCCharleston762	13.40013s	996.79km/s
6	UsScAiken931	13.40068s	888.68km/s
7	UsNcCharlotte963	13.49149s	742.90km/s
8	UsGaNorcross984	13.51401s	831.90km/s
9	UsAlBirmingham873	13.53389s	886.95km/s
10	UsNjTomsriver774	13.57967s	1102.86km/s
...	...	...	...

RoCoF, voltage angle and amplitude curves are shown in the four sub-graphs in Fig. 4, respectively. In Fig. 4(b), the red dash line denotes the event time determination threshold. It can be seen that there are large fluctuations from about the 13<sup>th</sup> second, and the event time determination trigger is activated right after that. Therefore, the generation trip event is successfully detected and the event location determination will be further executed.

In this actual case, SMDs located in different places detected the propagation waves at different time as shown in Table I, and

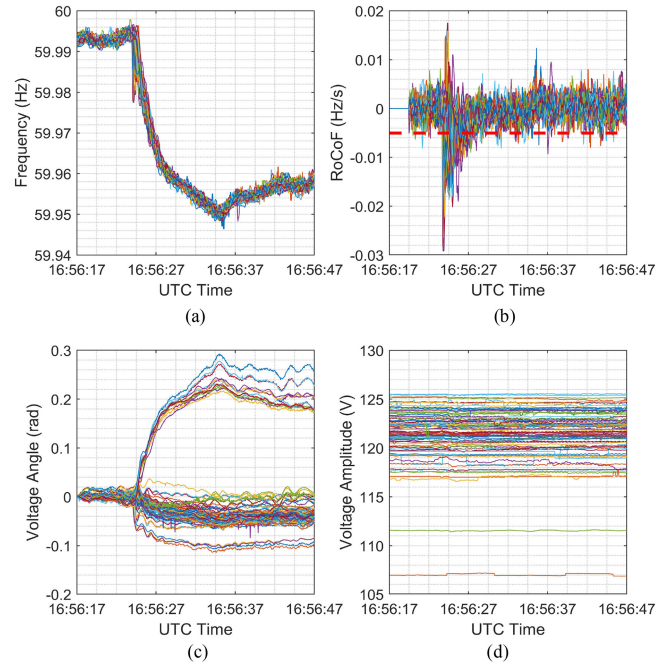


Fig. 4. Measured data curves for the event happens in 2020-04-04 16:56:32 UTC. (a) Frequency; (b) RoCoF; (c) Voltage angle; (d) Voltage amplitude.

the wave propagation speeds associated with different SMDs (i.e., FDRs in FNET/GridEye) estimated by historical confirmed events are also given in Table I. It is observed that wave propagation speeds associated with SMDs located at grid edges are slightly larger compared with those located in the middle of the grid. This is primarily because grid edges have less inertia. It should be mentioned that the farther an SMD is from the true event location, the less accurate the wave arrival time determination will be. Therefore, 5~10 SMDs with the earliest wave arrival time are enough and suitable for triangulation according to practical experience. Based on this principle, the event location is estimated at (37.8208, -77.4462). It can be seen that the distance between the true event location and the estimated one is only 119.6 km, which is very close (especially considering the sparse distribution of current FNET/GridEye sensors) and enough for choosing suitable data curves of SMDs to identify the type of event later.

It is worth mentioning that the proposed algorithm can be deployed for real-time applications and large-scale event cases. In fact, for each detected event, an event report including the estimated event location and the FDRs that receive the wave propagation earliest can be generated as shown in Fig. 5, which can support the following event identification. For event identification, the frequency, voltage angle and amplitude curves of the FDRs with the earliest wave arrival time are utilized for extracting corresponding features based on 2D-OLLP method and then the features are input into the well-trained RUSBoosted trees-based classifier. In this case,  $N_{CART}$  is set as 20 according to training results and all 20 CART classifiers vote this event as a generation trip event. Therefore, this event is correctly identified as a generation trip.

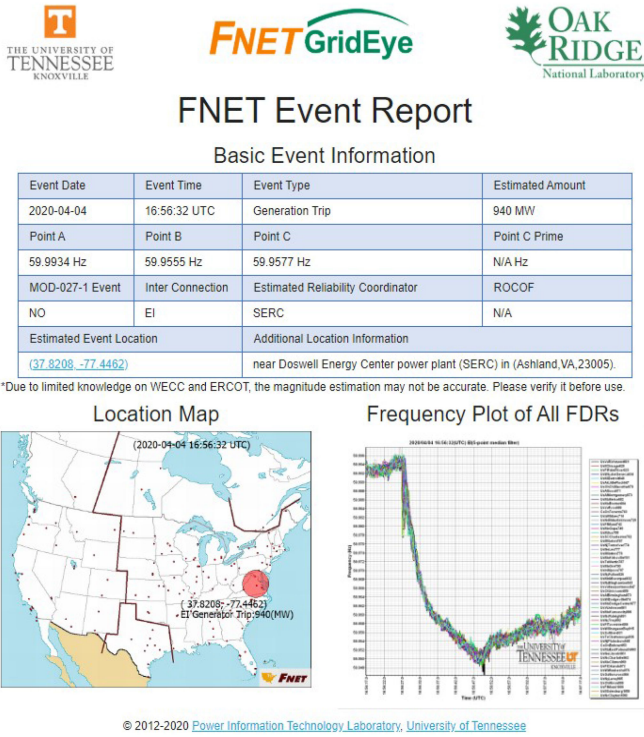


Fig. 5. Report generated for the studied generator trip event happened in EI power system on April 4, 2020, 16:56 (UTC).

TABLE II  
NUMBERS AND TYPES OF EVENT IDENTIFICATION CASES

Type	Number	Actual	Simulated
Short-Circuit Fault	384	43	341
Line Trip	742	162	580
Load Shedding	1135	674	461
Generation Trip	821	368	453
Total	3082	1247	1835

*B. Comparisons with Other Event Identification Algorithms Using Numerous Actual Measured and Simulated Cases*

Since the number and types of confirmed event cases in power systems are limited, batch simulations in simplified actual power systems (e.g., Inter-connected New England test system (NETS) and New York power system (NYPS), and WECC-179 bus power system) are performed to supplement the data for event identification demonstration. There are a total of 3082 actual measured or simulated event identification cases with four different types, and the numbers and typical curves of different types of events are given in Table II and Figs. 6–9, respectively. The hold-out method and cross-validation method are two commonly used methods for performance evaluation of different algorithms [42]. In general, the hold-out method, which divides the whole data set into the training set and the test set, is suitable for large scale data set since it can be carried out fast while the result largely depends on how to divide the training set and the test set. In this work, the number of cases is not very large, so the cross-validation method is more suitable since the result is more reliable although a larger computation burden is caused.

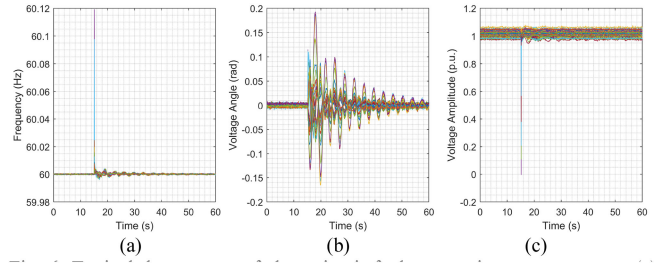


Fig. 6. Typical data curves of short-circuit fault events in power systems. (a) Frequency; (b) Voltage angle; (c) Voltage amplitude.

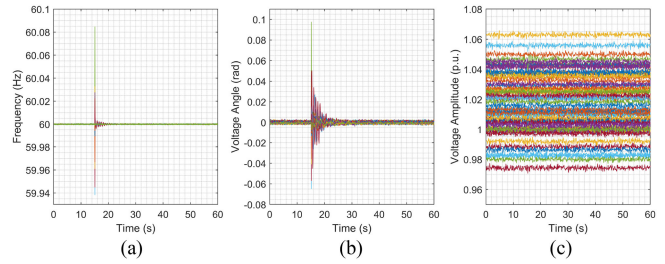


Fig. 7. Typical data curves of line trip events in power systems. (a) Frequency; (b) Voltage angle; (c) Voltage amplitude.

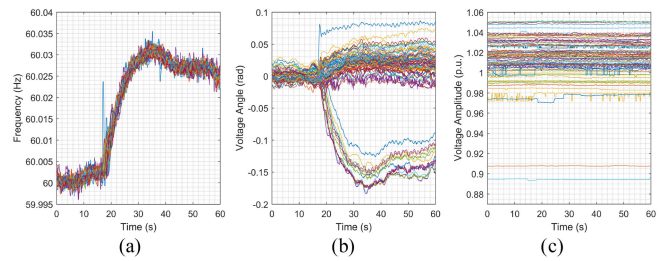


Fig. 8. Typical data curves of load shedding events in power systems. (a) Frequency; (b) Voltage angle; (c) Voltage amplitude.

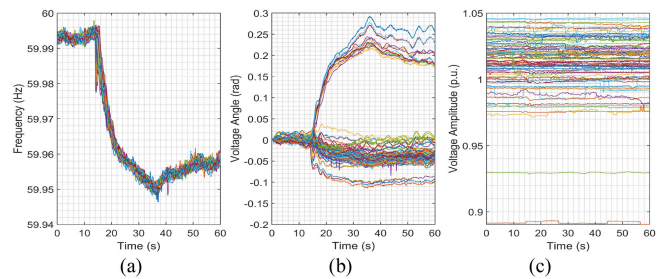


Fig. 9. Typical data curves of generation trip events in power systems. (a) Frequency; (b) Voltage angle; (c) Voltage amplitude.

Concretely, the whole data set is divided into  $k$  equal-sized subsets randomly and the cross-validation process is repeated  $k$  times. In each time, a single subset is used for testing the model, and the rest  $k-1$  subsets are used as the training set. Obviously, the cross-validation is more reasonable than the hold-out method and suitable for data set with a relatively small scale. Therefore, the 10-fold cross-validation ( $k = 10$  is commonly used) [42] is utilized to test the proposed algorithm, and the training set and test set are not explicitly represented in this work.



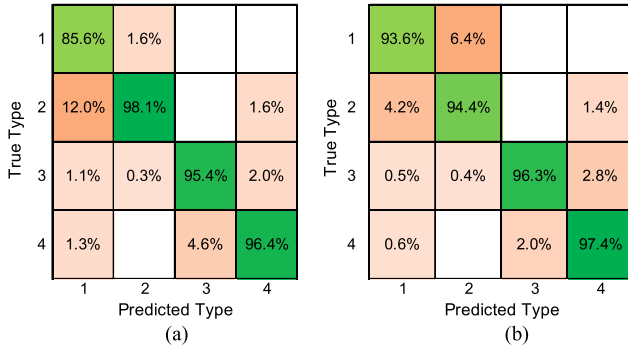


Fig. 10. Confusion matrices of the proposed event identification algorithm. (a) TPR and FNR; (b) PPV and FDR.

TABLE III  
EVALUATION INDEXES FOR THE PROPOSED 2D-OLPP-RUSBOOSTED  
TREES-BASED EVENT IDENTIFICATION ALGORITHM

Type	Precision	Recall	Accuracy	F1
Short-Circuit Fault	85.6%	93.6%	90.2%	89.4%
Line Trip	98.1%	94.4%	97.2%	96.2%
Load Shedding	95.4%	96.3%	96.5%	95.8%
Generation Trip	96.4%	97.6%	97.4%	97.0%

It can be seen from Figs. 6–9 that different types of events have different data curves, so the proposed 2D-OLPP method is utilized to extract the event features of each event type. It should be clarified that not all measured data curves but those from FDRs that are closest to the estimated event location are used for feature extraction and event identification since they have less wave arrival time errors in practical applications. It can also be seen from Table II that the numbers of different types of events are quite unbalanced. Thus, the proposed method based on RUSBoosted trees is employed to mitigate this issue and to identify the types of various event cases, and the results are shown in Fig. 10 and Table III. On the one hand, the confusion matrix associated with the true positive rate (TPR) and false-negative rate (FNR) is given in Fig. 10(a) and the sum of each column is equal to 100%. The TPR is also called as precision and is defined as  $R_P = N_{TP}/(N_{TP} + N_{FP})$ , where  $N_{TP}$  and  $N_{FP}$  are the numbers of true-positive (TP) event samples and false-negative (FP) event samples from the prediction results of the algorithm, respectively. For example, in the set of event samples that are predicted as the 1<sup>st</sup> type (i.e., short-circuit fault), there are 82.6% of them are true short-circuit fault events while 14.5%, 1.4% and 1.4% of them are respectively line trip, generation trip and load shedding events in fact but incorrectly predicted as short-circuit fault event.

On the other hand, the confusion matrix associated with positive predictive values (PPV) and false discovery rate is given in Fig. 10(b) and the sum of values in each row is equal to 100%. The PPV is also called as the recall rate and is defined as  $R_R = N_{TP}/(N_{TP} + N_{FN})$ , where  $N_{FN}$  is the number of false-negative (FN) event samples. For example, for the true 1<sup>st</sup> type (i.e., short-circuit fault) of event samples, 93.6% of them are correctly predicted while 6.4% of them are incorrectly predicted as line trip events.

It can be seen that the precision and recall rate can only reflect the performance of the event identification algorithm from one aspect, so two more comprehensive evaluation indexes are also employed for comparison, i.e., accuracy  $R_A = (N_{TP} + N_{TN})/(N_{TP} + N_{TN} + N_{FP} + N_{FN})$  and F1-score  $R_{F1} = 2R_P R_R / (R_P + R_R)$ ; where  $N_{TN}$  is the number of true negative event samples. It can be seen from Table III that the proposed 2D-OLPP-RUSBoosted trees-based algorithm achieves good performance for event identification. It is noted that the evaluation indexes of the short-circuit fault events are relatively small, and the reason is that the number of short-circuit fault events for training is the least as shown in Table II.

To show the improvement of the proposed algorithm, the results of event identification obtained by other state-of-the-art algorithms are given in Table IV for comparisons. It should be mentioned that the determination of the hyperparameters for machine learning algorithms is crucial to the performance of event identification. For the proposed 2D-OLPP-RUSBoosted trees-based algorithm, the hyperparameters include *maximum number of splits*, *number of learners* and *learning rate*, which can be tuned by grid search with cross-validation (GridSearchCV) method [44] according to actual situations. In this work, the *maximum number of splits* is tuned as 30, the *number of learners* (i.e.,  $N_{CART}$ ) is tuned as 20, and *learning rate* is tuned as 0.1, respectively. In general, the under-fitting problem may be caused if  $N_{CART}$  is set too small, while the over-fitting problem may be caused if  $N_{CART}$  is set too large. Therefore, the value of  $N_{CART}$  is tuned by the GridSearchCV method in practice. Furthermore, the setting of  $N_{CART}$  has little relationship with the scale of samples (i.e., the number of events for training) since the number of features extracted by 2D-OLPP is unchanged. For the random undersampling, a larger  $N_{CART}$  may help to reduce the risk of information loss problem while a larger computation burden and the over-fitting problem will be caused in the meantime.

It can be seen that the proposed algorithm outperforms other algorithms for most evaluation indexes (i.e., 95.9% for the recall rate, 96.1% for accuracy and 95.5% for F1-score). Indeed, the cluster-based sparse coding algorithm obtains higher precision (i.e., 98.1%) than the proposed algorithm (i.e., 95.1%), but its recall rate is much lower (i.e., 88.6% v.s. 95.9%) and the accuracy and F1-score are lower (i.e., 92.6% v.s. 96.1%; 93.1% v.s. 95.5%) as well. It should be clarified that the precision and recall rate indexes in Table IV denote the weighted average values for all types of events, i.e.,  $R_P = (N_{SCF}R_{P,SCF} + N_{LT}R_{P,LT} + N_{LS}R_{P,LS} + N_{GT}R_{P,GT}) / (N_{SCF} + N_{LT} + N_{LS} + N_{GT})$ , where  $N_{SCF}$ ,  $N_{LT}$ ,  $N_{LS}$ , and  $N_{GT}$  denote the numbers of the short-circuit fault, line trip, load shedding, and generation trip events, respectively. Thus, the precision index can be determined. Similarly, the recall rate, accuracy and F1-value in Table IV can be determined in this way as well, which can measure the performances of different algorithms comprehensively.

For the LSTM-based algorithm that has advantages on processing time-series data, it can obtain better performance than CNN indeed since more time-series information is considered. However, it is still inferior to the proposed

TABLE IV  
COMPARISONS OF EVENT IDENTIFICATION AMONG DIFFERENT ALGORITHMS

Algorithm	Precision	Recall	Accuracy	F1	Computation Time
1-Nearest Neighbors [27]	86.9%	91.5%	90.2%	89.1%	13.2s
Support Vector Machine [43]	84.5%	80.5%	84.6%	82.5%	14.7s
Sparse Unmixing [23]	93.5%	80.6%	86.9%	86.6%	1.53s
Cluster-Based Sparse Coding [25]	98.1%	88.6%	92.6%	93.1%	1.65s
Convolutional Neural Network [26]	92.1%	74.3%	75.3%	82.2%	546.82s
Long Short-Term Memory Network [45]	93.4%	80.2%	86.6%	86.3%	963.31s
2D-OLPP-Random Forest	93.9%	94.0%	93.2%	94.0%	35.44s
2D-PCA-RUSBoosted Trees	88.6%	87.5%	90.7%	88.1%	47.37s
2D-OLPP-ADASYNBoosted Trees	95.2%	94.6%	96.3%	94.9%	326.91s
2D-OLPP-SMOTEBoosted Trees	94.6%	95.2%	95.6%	94.9%	345.46s
Proposed 2D-OLPP-RUSBoosted Trees Algorithm	95.1%	95.9%	96.1%	95.5%	53.06s

2D-OLPP-RUSBoosted trees-based algorithm. The reason is the proposed algorithm considers the imbalance of event samples and use random undersampling to mitigate this issue. Therefore, the proposed algorithm can achieve a much higher recall rate for event identification when compared with other algorithms. It is worth mentioning that the 2D-OLPP-random forest and the 2D-PCA-RUSBoosted trees-based algorithms are also tested, and the comparisons show the advantages of the random undersampling and OLPP separately by control the feature extraction method or event classifier fixed.

To demonstrate the effectiveness of RUS, the comparisons among ADASYNBoosted trees, SMOTEBoosted trees and RUSBoosted trees with the same 2D-OLPP feature extraction method are also given in Table IV. It can be seen that the values of the four indexes are similar while the computation time of ADASYNBoosted trees and SMOTEBoosted trees are much longer. The reasons are that: i) Several CARTs are utilized in the proposed 2D-OLPP-RUSBoosted trees-based algorithm, so the random undersampling would be performed several times independently, which means the risk of information loss problem can be greatly reduced. ii) Essentially, both the undersampling and oversampling methods cannot create new information, which means they can obtain similar performance theoretically. iii) The ADASYN and SMOTE methods resample massive samples during training, which means that the scale of the dataset would be enlarged and the computation burden would be increased significantly.

The computation times of other algorithms are also given in Table IV and all tests are performed on Windows 10 platform with Intel i5-7400 CPU and 16GB RAM. Although the computation times of the first four algorithms are much shorter, their identification performances are inferior to the proposed 2D-OLPP-RUSBoosted trees-based algorithm. Therefore, it can be concluded that the proposed 2D-OLPP-RUSBoosted trees-based algorithm outperforms other algorithms when comprehensively considering the performance and computation for event identification.

### C. Testing of the Effects of Noise and Data Loss

It should be mentioned that the employed data set consists of actual measured and simulated cases, therefore, it is worth

TABLE V  
EVALUATION INDEXES FOR THE PROPOSED 2D-OLPP-RUSBOOSTED TREES-BASED EVENT IDENTIFICATION ALGORITHM CONSIDERING THE EFFECTS OF LOSS AND NOISE IN DATA

Type	Precision	Recall	Accuracy	F1
Short-Circuit Fault	83.2%	92.4%	89.6%	88.9%
Line Trip	97.1%	93.5%	96.4%	94.9%
Load Shedding	94.5%	95.4%	95.2%	95.2%
Generation Trip	94.0%	96.7%	97.0%	96.2%

testing the robustness of the proposed algorithm considering the effects of noise and data loss. To do this testing, Gaussian noise with zero mean and 0.1mHz standard deviation is added in the simulated data and it is assumed that 1% of data in simulated cases are randomly lost, which are similar to the actual situation of measurement data in FNET/GridEye.

Before using the proposed algorithm, data pre-processing should be performed in advance for both the measured and simulated data, which can mitigate the effects of loss and noise data. The principles of pre-processing are: i) if an FDR does not have available measurements in a given time window or most of the time, then it will be abandoned; ii) using linear interpolation method to fill the missing data; iii) using 5-point moving median filter to denoise and avoid outlier data.

After the data pre-processing, the proposed 2D-OLPP-RUSBoosted trees-based algorithm is tested and the results are shown in Table V. Compared with Table III, it can be seen that the performance decreases slightly, which is still acceptable in most situations. Therefore, it can be concluded that the noise and data loss do affect the performance of the proposed algorithm, while the effects are acceptable and the robustness of the proposed algorithm is good.

It also should be clarified that the topology change would influence the performance of the proposed algorithm. However, this effect has been considered already in this work. It can be seen from Section V.B that the data used for case studies are constituted by the measured data of U.S. power systems from FNET/GridEye, the simulated data of NETS-NYPS, and the simulated data of the WECC-179 bus power system. The topologies of these three power systems are different from each other, in

other words, the topologies can be regarded as changeable in the testing stage. In the meantime, the final testing results in Table III and Table V show that the proposed 2D-OLPP-RUSBoosted trees-based algorithm can achieve good performance, which means that the performance of the 2D-OLPP-RUSBoosted trees-based algorithm can be ensured when topology changes.

## VI. DISCUSSIONS AND ADDITIONAL ILLUSTRATIONS

The measurement system configuration is important for implementing the proposed event identification algorithm successfully. FNET/GridEye is a wide-area monitoring system that covers worldwide power systems. As the essential elements, FDRs can measure the frequency, voltage angle and amplitude information. Besides, FDRs have a GPS module to provide accurate timestamps to measurements. The data collected by the FDR are sent through the Internet to the data center, and then processed in the data server and then digested in web servers, real-time application servers, post-event analysis and storage servers, and backup servers [24]. By 2020, it has around 300 sensor units deployed around the world and most of them are deployed in the U.S., which provides strong supports for the event detection and identification in this work.

It is noted that besides the four types of event aforementioned (i.e., short-circuit fault, line trip, load shedding and generation trip), there are oscillation events in power systems as well, such as low-frequency oscillation (LFO) with a frequency around 0.1~2.5Hz, sub-synchronous oscillation (SSO) with a frequency slightly below 50Hz or 60Hz, and high-frequency oscillation (HFO) with a frequency of hundreds or thousands of Hertz [46]. Due to the limited sampling rates (i.e., 10Hz) of current widely deployed FDRs, the SSO and HFO cannot be monitored and identified by the proposed algorithm in this work indeed. According to the Shannon's sampling theorem, the data sampling rate of each FDR should reach more than two times the frequency of oscillation theoretically, therefore, it is unrealistic to monitor and identify SSO and HFO events indeed. For the same reason, the high-frequency faults caused by the high integration of power electronic equipment are also hard to be monitored by the proposed algorithm. If the advanced version of FDRs (i.e., UGAs), which can achieve up to 1.44kHz sampling rate [5], can be widely deployed in the future, the SSO could be detected and identified theoretically while it is still hard to detect HFO events. It should be acknowledged that the proposed algorithm aims to detect and identify the events associated with the electromechanical transient processes and has limitations for the electromagnetic transient processes. For the LFO events, the FNET/GridEye can monitor and analyze them successfully by the matrix pencil method [47]. It should be mentioned that it is found that most LFO events are caused by generator trip and load shedding events in fact [48], which means that they are generator trip or load shedding events in essence. Therefore, they are classified as either generator trip or load shedding events in this work.

In the future, more and more RES will be connected to power systems, which will involve more uncertainties and fluctuations in power systems. Thus, it is more important to monitor and

identify the events that occur in power systems timely and accurately. It can be seen that the proposed 2D-OLPP-RUSBoosted trees-based algorithm is only based on measurement data and does not require the detailed model of power systems. Therefore, theoretically, the proposed algorithm can adapt to detect and identify the ever-increasing events in power systems as long as the Shannon's sampling theorem is satisfied.

## VII. CONCLUSION AND FUTURE WORKS

In this work, an event identification algorithm is proposed with event time and location determination considered for real power systems. First, the RoCoF is utilized to detect events that occur in power systems; then, the differences among wave arrival times for SMDs located in different places are employed for event localization with the preliminary but first-time consideration for the anisotropy of wave propagation speed; finally, on the basis of event time and location, a 2D-OLPP-RUSBoosted trees-based algorithm is proposed for accurate event identification, which can achieve better performance than existing algorithms. Although the identification for multiple events is not discussed in this work, they were already achieved in our previous work [24], [25], and can be applied for the proposed algorithm when necessary. Nevertheless, it should be mentioned that we found lots of identified load shedding events are actually pump storage units turning off the motors, because they have quite similar system responses. Therefore, further work is to distinguish pump storage events from load shedding events in actual power systems. A possible solution is that the location information of the pump storage stations can be further involved in the process of event identification. For example, if an event is detected near the location of a pump storage station and is identified as a load shedding event at first, then it can be inferred that it is a pump storage event actually. In such a situation, a threshold or a weight may need to be tuned to distinguish these two types of events, which is worth further studies in the future.

## REFERENCES

- [1] C. Lu, T. Bi, M. Ilic, K. Tomsovic, and X. Wang, "Guest Editorial: New trends in wide-area monitoring and control of power systems with large scale renewables," *IET Gener. Transmiss. Distrib.*, vol. 11, no. 18, pp. 4403–4405, Dec. 2017.
- [2] S. Liu *et al.*, "Robust system separation strategy considering online wide-area coherency identification and uncertainties of renewable energy sources," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 3574–3587, Sep. 2020.
- [3] A. Ghasempour, "Advanced metering infrastructure in smart grid: Requirements, challenges, architectures, technologies, and optimizations," in *Smart Grids: Emerging Technologies, Challenges and Future Directions*, ch. 3, J. Lou, Ed., New York, NY, USA: Nova Science Publishers, 2017.
- [4] T. Jiang, L. Bai, G. Li, H. Jia, Q. Hu, and H. Yuan, "Estimating inter-area dominant oscillation mode in bulk power grid using multi-channel continuous wavelet transform," *J. Modern Power Syst. Clean Energy*, vol. 4, no. 3, pp. 394–405, Jun. 2016.
- [5] S. You *et al.*, "Comparative assessment of tactics to improve primary frequency response without curtailing solar output in high photovoltaic interconnection grids," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 718–728, Apr. 2019.
- [6] E. Perez and J. Barros, "A proposal for on-line detection and classification of voltage events in power systems," *IEEE Trans. Power Del.*, vol. 23, no. 4, pp. 2132–2138, Oct. 2008.

- [7] T. Routtenberg, R. Concepcion, and L. Tong, "PMU-based detection of voltage imbalances with tolerance constraints," *IEEE Trans. Power Del.*, vol. 32, no. 1, pp. 484–494, Feb. 2017.
- [8] K. Mei, S. M. Rovnyak, and C. Ong, "Clustering-based dynamic event location using wide-area phasor measurements," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 673–679, May 2008.
- [9] F. Bai, X. Wang, Y. Liu, X. Liu, Y. Xiang, and Y. Liu, "Measurement-based frequency dynamic response estimation using geometric template matching and recurrent artificial neural network," *CSEE J. Power Energy Syst.*, vol. 2, no. 3, pp. 10–18, Sep. 2016.
- [10] D. Kim, T. Y. Chun, S. Yoon, G. Lee, and Y. Shin, "Wavelet-based event detection method using PMU data," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1154–1162, May 2017.
- [11] Y. Ge, A. J. Flueck, D. K. Kim, J. B. Ahn, J. D. Lee, and D. Y. Kwon, "Power system real-time event detection and associated data archival reduction based on synchrophasors," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2088–2097, Jul. 2015.
- [12] L. Xie, Y. Chen, and P. R. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2784–2794, Nov. 2014.
- [13] M. Cui, J. Wang, J. Tan, A. R. Florita, and Y. Zhang, "A novel event detection method using PMU data with high precision," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 454–466, Jan. 2019.
- [14] S. Liu *et al.*, "Data-driven event detection of power systems based on unequal-interval reduction of PMU data and local outlier factor," *IEEE Trans. Smart Grid*, vol. 10, no. 11, pp. 1630–1643, Mar. 2020.
- [15] X. Liu, D. M. Laverty, R. J. Best, K. Li, D. J. Morrow, and S. McLoone, "Principal component analysis of wide-area phasor measurements for islanding detection-A geometric view," *IEEE Trans. Power Del.*, vol. 30, no. 2, pp. 976–985, Apr. 2015.
- [16] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window PCA," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2537–2548, Sep. 2016.
- [17] L. Cai, N. F. Thornhill, S. Kuenzel, and B. C. Pal, "Wide-area monitoring of power systems using principal component analysis and  $k$ -nearest neighbor analysis," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4913–4923, Sep. 2018.
- [18] P. Bhui and N. Senroy, "Application of recurrence quantification analysis to power system dynamic studies," *IEEE Trans. Power Syst.*, vol. 31, no. 1, pp. 581–591, Jan. 2016.
- [19] J. Ma, Y. V. Makarov, R. Diao, P. V. Etingov, J. E. Dagle, and E. De Tuglie, "The characteristic ellipsoid methodology and its application in power systems," *IEEE Trans. Power Syst.*, vol. 4, no. 27, pp. 2206–2214, May 2012.
- [20] W. Li, M. Wang, and J. H. Chow, "Real-time event identification through low-dimensional subspace characterization of high-dimensional synchrophasor data," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4937–4947, Sep. 2018.
- [21] M. Biswal, S. M. Brahma, and H. Cao, "Supervisory protection and automated event diagnosis using PMU data," *IEEE Trans. Power Del.*, vol. 31, no. 4, pp. 1855–1863, Aug. 2016.
- [22] R. Yadav, S. Raj, and A. K. Pradhan, "Real-time event classification in power system with renewables using Kernel density estimation and deep neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6849–6859, Nov. 2019.
- [23] W. Wang *et al.*, "Multiple event detection and recognition through sparse unmixing for high-resolution situational awareness in power grid," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 1654–1664, Jul. 2014.
- [24] Y. Liu *et al.*, "Recent developments of FNET/GridEye - A situational awareness tool for smart grid," *CSEE J. Power Energy Syst.*, vol. 2, no. 3, pp. 19–27, Sep. 2016.
- [25] Y. Song, W. Wang, Z. Zhang, H. Qi, and Y. Liu, "Multiple event detection and recognition for large-scale power systems through cluster-based sparse coding," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4199–4210, Nov. 2017.
- [26] W. Li and M. Wang, "Identifying overlapping successive events using a shallow convolutional neural network," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4762–4772, Nov. 2019.
- [27] R. Yadav, A. K. Pradhan, and I. Kamwa, "Real-time multiple event detection and classification in power system using signal energy transformations," *IEEE Trans. Ind. Inform.*, vol. 15, no. 3, pp. 1521–1531, Mar. 2019.
- [28] W. Li, J. Tang, J. Ma, and Y. Liu, "Online detection of start time and location for hypocenter in North America power grid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 253–260, Dec. 2010.
- [29] T. Xia *et al.*, "Wide-area frequency based event location estimation," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, 2007, pp. 1–7.
- [30] P. Roger, "A generalized inverse for matrices," *Math. Proc. Cambridge Philos. Soc.*, vol. 51, no. 3, pp. 406–413, Jun. 1955.
- [31] S. Liu *et al.*, "Model-free data authentication for cyber security in power systems," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4565–4568, Sep. 2020.
- [32] S. Chen, H. Zhao, M. Kong, and B. Luo, "2D-LPP: A two-dimensional extension of locality preserving projections," *Neurocomputing*, vol. 70, no. 4–6, pp. 912–921, Jan. 2007.
- [33] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.
- [34] G. Shikkenawis and S. K. Mitra, "2D orthogonal locality preserving projection for image denoising," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 262–273, Jan. 2016.
- [35] N. F. Avila, G. Figueroa, and C. Chu, "NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7171–7180, Nov. 2018.
- [36] S. Liu *et al.*, "Practical method for mitigating three-phase unbalance based on data-driven user phase identification," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1653–1657, Mar. 2020.
- [37] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.
- [38] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUS-Boost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst. Man Cybern. A Syst. Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.
- [39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jan. 2002.
- [40] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, 2008, pp. 1322–1328.
- [41] M. Jagelid and M. Movin, "A comparison of resampling techniques to handle the class imbalance problem in machine learning." Accessed: Dec. 16, 2020. [Online]. Available: <http://www.diva-portal.org/smash/get/diva2:1108465/FULLTEXT02.pdf>
- [42] M. Stone, "Cross-validated choice and assessment of statistical predictions," *J. Roy. Statist. Soc.: Ser. B (Methodological)*, vol. 36, no. 2, pp. 111–147, Jan. 1974.
- [43] M. Biswal, Y. Hao, P. Chen, S. Brahma, H. Cao, and P. De Leon, "Evaluating performance of classifiers for supervisory protection using disturbance data from phasor measurement unit," in *Proc. 5th IEEE PES Innov. Smart Grid Technol. Eur.*, 2014, pp. 1–6.
- [44] "Parameter estimation using grid search with cross-validation," [Online]. Available: [https://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_grid\\_search\\_digits.html](https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html)
- [45] Q. Wang and S. Bu, "Deep learning enhanced situation awareness for high renewable-penetrated power systems with multiple data corruptions," *IET Renewable Power Gener.*, vol. 14, no. 7, pp. 1134–1142, May 2020.
- [46] Y. Chi *et al.*, "Overview of mechanism and mitigation measures on multi-frequency oscillation caused by large-scale integration of wind power," *CSEE J. Power Energy Syst.*, vol. 4, no. 5, pp. 433–443, Dec. 2019.
- [47] M. L. Crow and A. Singh, "The matrix pencil for power system modal extraction," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 501–502, Feb. 2005.
- [48] "FNET oscillation report," Accessed: Dec. 16, 2020. [Online]. Available: <http://powerit2.eecs.utk.edu/fnet-oscillation-report/oscilreport.php?eventid=35719&AuthCode=5cy84A8h>