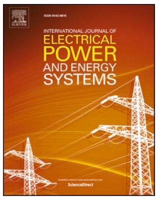
Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes



Data source authentication for wide-area synchrophasor measurements based on spatial signature extraction and quadratic kernel SVM

Shengyuan Liu [a,b], Shutang You [b], He Yin [b], Zhenzhi Lin [a,c,*], Yilu Liu [b,d], Yi Cui [e], Wenxuan Yao [f], Lakshmi Sundaresh [b]

[a] *Department of Electrical Engineering, Zhejiang University, Hangzhou 310027, China*
[b] *Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA*
[c] *Department of Electrical Engineering, Shandong University, Jinan 250061, China*
[d] *Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA*
[e] *School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia*
[f] *College of Electrical and Information Engineering, Hunan University, Changsha 410082, China*

ARTICLE INFO

Keywords:
Data source authentication
Spatial signature extraction
quadratic kernel support vector machine (QKSVM)
Synchrophasor measurement
wide-area measurement system (WAMS)

ABSTRACT

As essential components of the wide-area measurement system (WAMS), phasor measurement units (PMUs), frequency disturbance recorders (FDRs) and universal grid analyzers (UGAs) collect valuable data continuously to reveal the dynamic variations of power systems and to enhance the operators' situational awareness ability. However, these devices are vulnerable to multiple types of data exception emerging in recent years, such as data source ID mix exception spoofing, substantially threatening system security. To ensure the cyber security of WAMS, this work proposes a new spatial signature extraction method, followed by the quadratic kernel support vector machine (QKSVM)-based algorithm, to authenticate data source in WAMS. First, the load–frequency characteristic (LFC), which can represent the impacts of load variations on frequency, is utilized to extract the spatial signatures of FDRs located in different regions. Then, the quadratic kernel function is employed in the QKSVM-based algorithm to map the signatures into Hilbert space to authenticate the data source more accurately. Finally, case studies in the U.S. Western and Eastern power systems show that the proposed model-free algorithm is less sensitive to system sizes, and can achieve a higher authentication accuracy in a much shorter window length compared with other algorithms.

1. Introduction

1.1. Backgrounds and motivations

Driven by the ever-increasing demand for wide-area monitoring and advanced controls, modern power systems will be more and more dependent on wide-area measurement systems (WAMS) [1,2]. However, some data exception issues have already emerged towards the WAMS, in which malicious attackers inject fake data to compromise measurement data from synchrophasor measurement devices (SMDs) such as phasor measurement units (PMUs), frequency disturbance recorders (FDRs) and universal grid analyzers (UGAs), to deceive power system operators and mislead them to make a wrong decision. Data spoofing issues occur when the SMDs in WAMS are hacked and measurement data are manipulated maliciously by hackers. Data spoofing uses false measurement data or data with a false source ID number to inject into WAMS networks and devices, pretending to be measurements from legitimate sensors [3–6]. Data spoofing aiming at mixing the data source of SMDs can be performed in many ways and it can be concluded from Refs. [5] and [7] that i) Phasor data concentrators (PDCs) indeed can check the PMU data quality and align the timestamps of PMU data, however, it cannot handle the intentional change and the mix of PMU data. Even though the PDCs are configurated correctly, there is nothing PDCs can do if the data source ID is changed or mixed during the communication from PMUs to PDCs. Since the attacks are carried out artificially, they could happen at any time. ii) The data source ID can also be changed or mixed during the communication from PDCs to data storage servers, or even be changed or mixed in the data storage servers if security mechanisms are not perfect. iii) The IEEE standard alerts that users need to be aware of the risks of unsecured communications and

[*] Corresponding author at: Department of Electrical Engineering, Zhejiang University, Hangzhou 310027, China.
E-mail address: linzhenzhi@zju.edu.cn (Z. Lin).

https://doi.org/10.1016/j.ijepes.2022.108083
Received 23 May 2021; Received in revised form 23 July 2021; Accepted 18 February 2022
Available online 15 March 2022

should consider adopting more secure methods, which means there are always potential threats during the communication process. Most WAMS-based systems are quite vulnerable to data spoofing [8,9] since they rely on unencrypted communication protocols and use only ID numbers to verify the identities of SMDs. Therefore, it has significant value to develop data source authentication algorithms.

1.2. Literature review

Existing research on this issue can be divided into three categories: i) advanced state estimation combined with statistical algorithms to detect direct false data injection attacks (FDIAs) in PMUs [10–14]; ii) defense mechanism against GPS signal spoofing, which leads to fake PMU in a round-about way [6,15]; and iii) machine learning-based algorithms to detect data exception [5,16–18].

In [10], the Kullback-Leibler (KL) distance is employed to represent the difference between the probability distribution determined by historical data and the one that deviated from new measurement data through state estimation. The value of KL distance will be very large when there is a false data injection. In [11], Kalman filter (KF) is utilized to estimate the state variables in power systems, and the chi-square detector and Euclidean detectors are utilized together to detect system abnormalities. In [12], a covariance selection-based data spoofing detection algorithm for smart grids is proposed. This approach decomposes the state estimation problem into several subproblems of maximum likelihood estimation (MLE) so as to achieve a decentralized and fast detection. In [13], the Markov graph of bus phase angles, which is consistent with the power grid graph in normal conditions, is employed to detect the data exception. This method leverages the fact that the Markov graph would change if the power system is under data spoofing. In [14], load forecast information, generation schedules and measurement data are synthesized together and a predictive state estimation method is used to detect anomalies for measured data. In [15], the spoofing-matched algorithm is proposed to correct the measured data under GPS data exception. In [6], a cross-layer defense mechanism is proposed to detect simultaneous GPS data spoofing that aims to mix the sources of PMUs. The carrier-to-noise ratio and trustworthiness evaluation are respectively employed in two layers to cross authenticate the data source. In [16], support vector machine (SVM) technique is combined with state estimation to achieve the detection of stealthy false data injection. In [17], several machine learning algorithms such as SVM, k-nearest neighbor (k-NN) and sparse logistic regression (SLR) as well as ensemble learning algorithms including adaptive boosting (AdaBoost) and multiple kernel learning (MKL) are further studied for state vector estimation (SVE) to detect data exception. The results show that these machine learning algorithms are more sensitive to the system size while the AdaBoost and MKL are more robust despite the longer computational time. In [5] and [18], mathematical morphology (MM)-based decomposition is used to extract the features of different PMUs/FDRs, then multi-grained cascade forest (gcForest) and random forest classifications are respectively employed to determine the source of measurement data. The aforementioned algorithms are beneficial to data exception detection but still have some limitations. For [10–14,16,17], the whole and detailed mathematical models and electrical parameters are required for performing state estimation. For [6,15], the physical characteristics of GPS devices are further needed. However, these models, parameters and characteristics may be inaccessible due to proprietary issues, or outdated due to system changes. For the model-free algorithms in [5] and [18], their window length for detection is quite long, making them difficult to detect data exceptions timely in practical applications.

1.3. Contributions and organizations

In light of the above challenges, this work proposes a quadratic kernel SVM (QKSVM)-based data source authentication algorithm for wide-area synchrophasor measurements in power systems and aims to authenticate the data source of a large number of SMDs in bulk power systems using low-reporting rate measurement data with a relatively short time window. More specifically, the SMDs mentioned in this work mainly refer to FDRs. The major contributions of this work can be summarized as follows.

i) Inspired by the load–frequency characteristics, a novel feature extraction method for measured data is first proposed. Compared with discrete wavelet transformation (DWT) or MM in existing methods, the proposed LFC-based method has more physical meaning. In other words, more effective spatial signatures of FDR in different places can be extracted, thus more effective input features can be utilized for the source authentication algorithm.

ii) QKSVM-based algorithm has cooperated with the LFC-based spatial signature extraction method for data source authentication. Compared with other data source authentication algorithms, the proposed algorithm can achieve better performances including:

- The proposed LFC-QKSVM algorithm does not require detailed models and parameters of power systems, making it model-free and more practical in actual applications;
- The proposed LFC-QKSVM algorithm is less sensitive to the system size, which can be used for a large number of SMDs (e.g. FDRs) in bulk power systems;
- The proposed LFC-QKSVM algorithm can achieve a much higher data source authentication accuracy with a shorter window length using low-reporting measurement data, which means data exceptions can be detected more accurately and timely in practice.

The rest of this work is organized as follows. Section 2 gives the problem descriptions for the data source authentication. Section 3 briefly introduces the concept of load–frequency characteristics (LFC) and then presents the LFC-based spatial signature extraction method for the measured data of FDRs. Section 4 introduces the QKSVM classifier and proposes the corresponding data source authentication algorithm. Section 5 performs case studies and comparisons on two actual U.S. power systems, and examples of practical applications are also given. Section 6 gives some meaningful discussions for the proposed algorithm. Finally, several conclusions are given in Section 7.

2. Problem descriptions for the data source authentication of SMDs

The aim of data source authentication is to detect whether there are changes in the data source information for the SMDs. It should be clarified that SMDs include PMUs, FDRs, UGAs, etc., and this work mainly focuses on the FDRs deployed in FNET/GridEye [19]. FNET/GridEye is a GPS-synchronized wide-area frequency measurement network established by our research groups in the past few decades, which can measure the frequency, phase angle and voltage amplitude, and the measured data are transmitted via the Internet to the servers in real time [20]. Therefore, if each FDR can be identified successfully, i.e., each FDR can be distinguished from other FDRs successfully, then the problem of data source authentication can be solved. In other words, “data source authentication” can be also regarded as the “data source identification” (i.e., to know which FDR data come from) in this work. Hence, the problem of data source authentication can be converted into a classification problem. More specifically, since the historical measurement data of FDRs can be obtained and utilized for training, the problem of data source authentication is a supervised learning classification problem, which can be solved by several feature extraction and machine learning algorithms.

To evaluate the performance of different data source authentication algorithms, several evaluation indexes are given here first.

i) Data authentication accuracy R_{acc}.

$$R_{\mathrm{acc}} = (N_{\mathrm{TP}} + N_{\mathrm{TN}})/(N_{\mathrm{TP}} + N_{\mathrm{TN}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}) \tag{1}$$

ii) Recall rate of data authentication R_{recall}.

$$R_{\mathrm{recall}} = N_{\mathrm{TP}}/(N_{\mathrm{TP}} + N_{\mathrm{FN}}) \tag{2}$$

iii) F-1 score of data authentication R_{F1}.

$$R_{\mathrm{F1}} = 2N_{\mathrm{TP}}^2/(2N_{\mathrm{TP}}^2 + N_{\mathrm{TP}}N_{\mathrm{FN}} + N_{\mathrm{TP}}N_{\mathrm{FP}}) \tag{3}$$

where N_{FN}, N_{FP}, N_{TN} and N_{TP} respectively denote the total numbers of false negative, false positive, true negative and true positive samples of data source authentication. Therefore, all the R_{acc}, R_{F1} and R_{recall} indexes are the average and overall values in the total dataset. Generally, the larger the R_{acc}, R_{F1} and R_{recall} are, the better the authentication algorithm is [21].

3. LFC-Based spatial signature extraction method for FDR data

In practical applications, feature extraction should be performed before employing the machine learning algorithm. In fact, the quality of feature extraction has a large impact on the final result of a machine-learning algorithm. Therefore, an LFC-based method is first proposed in this section to extract the spatial signatures from the measured data for each FDR. The load variations in power systems cause frequency variation as well. In that case, the inertial effect is involved and the primary (and secondary frequency) control will also be triggered for generators to change the steam intake (or water intake) of the prime mover and adjust the input power of the generator to meet the load demand. These effects can be called as the load–frequency characteristics [22–24]. Assume that there are M FDRs located in different places and the frequency measured by the m^{th} FDR at time t is denoted as $f_m(t)$. The frequency measured by the m^{th} FDR would be influenced by the load variations in the region around the m^{th} FDR as well as the regions around the rest $M-1$ FDRs. However, the load variations in different regions would have different influences on the m^{th} FDR, thus the features of data measured by different FDRs could be extracted. Inspired by load–frequency characteristics, it is assumed the following relationship holds.

$$\begin{bmatrix} f_1(t+L) \\ f_2(t+L) \\ \vdots \\ f_M(t+L) \end{bmatrix} = \boldsymbol{K}_1 \begin{bmatrix} f_1(t+L-1) \\ f_2(t+L-1) \\ \vdots \\ f_M(t+L-1) \end{bmatrix} + \boldsymbol{K}_2 \begin{bmatrix} f_1(t+L-2) \\ f_2(t+L-2) \\ \vdots \\ f_M(t+L-2) \end{bmatrix} + \cdots + \boldsymbol{K}_{L-1} \begin{bmatrix} f_1(t-1) \\ f_2(t-1) \\ \vdots \\ f_M(t-1) \end{bmatrix} + \boldsymbol{K}_L \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_M(t) \end{bmatrix} \tag{4}$$

where M is the number of FDRs and L is the number of steps for the LFC-based extraction method. $\mathbf{K}_1$, $\mathbf{K}_2$, …, $\mathbf{K}_L$ are all M*M matrices, and the elements of them represent the LFC among different regions. In other words, the spatial signatures of FDRs in different regions are included in the matrices $\mathbf{K}_1$, $\mathbf{K}_2$, …, $\mathbf{K}_L$.

It should be mentioned that power systems are complex nonlinear systems and it may not be perfect to express the measurement frequency of an FDR linearly by all measured frequencies of FDR in the region in the previous L time. The motivations of using linear expression here are as follows: i) According to the load–frequency characteristics [22–24], the relationship between load and frequency can be linearized in the steady-state operating point [23]. In a similar way, the frequency variations caused by load fluctuations among different regions can also be regarded as linear approximately as long as the power systems are in the steady-state operating point [23]. ii) It is also hard to determine which higher-degree polynomial or nonlinear relationship should be utilized to describe the LFC if extreme accuracy is required since the detailed load model is quite complex. In fact, it can achieve acceptable accuracy by using the linear expression for LFC in the steady-state operating point [23]. iii) More importantly, more parameters would be involved if a more complex model is used. Thus, it would be much harder to solve these parameters since more parameters required to be solved means more equations are needed and the robustness of fitting results would decrease too. Furthermore, nonlinear equations would cause the multiple-solution problem, which should be avoided in practice. Hence, the linear expression is utilized in this work for practical applications.

It can be seen that there are M*M*L unknowns to be determined in $\mathbf{K}_1$, $\mathbf{K}_2$, …, $\mathbf{K}_L$ but only M equations in (4). To solve this issue, the data measured within the time range from $t+1$ to $t+2ML$ are used together to form the following equation as.

$$\boldsymbol{B}_{ML\times M} = \boldsymbol{A}_{ML\times ML}\boldsymbol{X}_{ML\times M} \tag{5}$$

where.

$$\boldsymbol{X}_{ML\times M} = \begin{bmatrix} K_1^{1,1} & K_1^{2,1} & \cdots & K_1^{M,1} \\ K_2^{1,1} & K_2^{2,1} & \cdots & K_2^{M,1} \\ \vdots & \vdots & \ddots & \vdots \\ K_L^{1,1} & K_L^{2,1} & \cdots & K_L^{M,1} \\ \vdots & \vdots & \ddots & \vdots \\ K_1^{1,M} & K_1^{2,M} & \cdots & K_1^{M,M} \\ K_2^{1,M} & K_2^{2,M} & \cdots & K_2^{M,M} \\ \vdots & \vdots & \ddots & \vdots \\ K_L^{1,M} & K_L^{2,M} & \cdots & K_L^{M,M} \end{bmatrix} \tag{6}$$

$$\boldsymbol{B}_{ML\times M} = \begin{bmatrix} f_1(t+ML+1) & f_2(t+ML+1) & \cdots & f_M(t+ML+1) \\ f_1(t+ML+2) & f_2(t+ML+2) & \cdots & f_M(t+ML+2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(t+2ML) & f_2(t+2ML) & \dots & f_M(t+2ML) \end{bmatrix} \tag{7}$$

$$\boldsymbol{A}_{ML\times ML} = \left[\boldsymbol{F}_1^{\mathrm{A}}, \boldsymbol{F}_2^{\mathrm{A}}, \dots, \boldsymbol{F}_M^{\mathrm{A}}\right] \tag{8}$$

$$\boldsymbol{F}_m^{\mathrm{A}} = \begin{bmatrix} f_m(t+L) & f_m(t+L-1) & \cdots & f_m(t+1) \\ f_m(t+L+1) & f_m(t+L) & \cdots & f_m(t+2) \\ \vdots & \vdots & \ddots & \vdots \\ f_m(t+L+ML-1) & f_m(t+L+ML-2) & \cdots & f_m(t+ML) \end{bmatrix} \tag{9}$$

where the elements of $\mathbf{X}_{ML\times M}$ are the M*M*L unknowns in $\mathbf{K}_1$, $\mathbf{K}_2$, …, $\mathbf{K}_L$. Thus, M*M*L equations are obtained in (5) and $\mathbf{K}_1$, $\mathbf{K}_2$, …, $\mathbf{K}_L$ can be determined by solving these equations together. The i^{th} column ($i = 1, 2, \dots, M$) of $\mathbf{X}_{ML\times M}$ has M*L elements. For example, $K_1^{1,1}$ represents the impact of LFC in the 1st region on the frequency data measured by the 1st FDR at time $t+1$; $K_L^{1,M}$ represents the impact of LFC in the M^{th} region on the frequency data measured by the 1st FDR at time $t+L$. As for the linear equation (5), it can be solved successfully as long as the square matrix $\mathbf{A}_{ML\times ML}$ is a full rank matrix. On the one hand, it can be seen from (9) that both the column and row elements of $\mathbf{A}_{ML\times ML}$ are linear independent, which means the $\mathbf{A}_{ML\times ML}$ is a full rank matrix theoretically. On the other hand, the FDRs can obtain high-precision measurements, which means the numerical problems can be avoided in actual situations. Therefore, the linear equation (5) can be solved successfully for the most time. Besides, the pseudo-inverse method can be used in case $\mathbf{A}_{ML\times ML}$ is a singular matrix.

It is also worth mentioning that although some advanced machine learning algorithms such as the convolutional neural network (CNN) can extract the features automatically, they might not obtain good results for this specific work. Details of this point can be found in the comparisons among different data source authentication algorithms in Section 5.

4. QKSVM-Based data source authentication algorithm

Once the features are extracted, the QKSVM-based algorithm is employed to authenticate the data source. Assume that the spatial signatures extracted from the i^{th} data sample can be denoted as.

$$\boldsymbol{x}_i = \left(K_1^{i,1}, K_2^{i,1}, \ldots, K_L^{i,1}, \ldots, \ldots, K_1^{i,M}, K_2^{i,M}, \ldots, K_L^{i,M}\right)^{\mathrm{T}} \tag{10}$$

The basic idea of the SVM technique is to classify several samples into different classes based on their distances to the separating hyperplane. Further, the QKSVM algorithm utilizes the quadratic kernel function to achieve better performance of the data source authentication considering the problem of linear inseparability. For better understanding, SVM is introduced briefly first, then the quadratic kernel function is involved to illustrate the process of the QKSVM algorithm.

Basically, the data source authentication problem can be converted into a linear inseparable problem as.

$$\begin{aligned} &\underset{\boldsymbol{w},b,\gamma}{min}\ \frac{1}{2}\|\boldsymbol{w}\|^2 + \eta\sum_{i=1}^{N_s}\gamma_i \\ &\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \gamma_i,\ \gamma_i \geqslant 0,\ i = 1, 2, \ldots, N_s \end{aligned} \tag{11}$$

where $\boldsymbol{w}$ and b are respectively the normal vector and intercept of the separating hyperplane. It can be seen from (10) that the values of data sample $\boldsymbol{x}_i$ ($i = 1, 2, \ldots, N_s$) are determined by the matrix $\mathbf{X}$, and the matrix $\mathbf{X}$ is determined by the matrices $\mathbf{A}$ and $\mathbf{B}$ according to (5). y_i is the class of the i^{th} data sample and N_s is the number of data samples in total; γ_i is a slack variable for the i^{th} data sample and η is the corresponding penalty coefficient associated with misclassification. In practical applications, η is usually set to 1.0 by default and can be tuned by grid search with cross-validation (GridSearchCV) method [25] according to actual situations. To solve the problem in (11), its dual problem is involved and can be further converted [26] as.

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{min}\ \frac{1}{2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j) - \sum_{i=1}^{N_s}\alpha_i \\ &\text{s.t. } \sum_{i=1}^{N_s}\alpha_i y_i = 0, 0 \leqslant \alpha_i \leqslant \eta,\ i = 1, 2, \ldots, N_s \end{aligned} \tag{12}$$

where $\alpha_i \geqslant 0$, $\mu_i \geqslant 0$ and $i = 1, 2, \ldots, N_s$. Due to the problem of linear inseparability, it is hard for a linear model to obtain good performance for data source authentication as illustrated in Fig. 1a. Therefore, the kernel function is employed in this work to improve the accuracy of data source authentication. The basic idea of kernel function $\Phi(\boldsymbol{x}, z) = \varphi(\boldsymbol{x})\varphi(z)$ is to utilize a non-linear transformation $\varphi(\boldsymbol{x})$ to map the input space (i.e., Euclidean space shown in Fig. 1a) onto a characteristic space (i.e., Hilbert space shown in Fig. 1b). Then, the problem of searching nonlinear separating boundary in Euclidean space is converted into the problem of searching a linear separating boundary in Hilbert space.

In this work, the quadratic kernel function $\Phi(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + 1)^2$ is involved to convert the data samples in Euclidean space into Hilbert

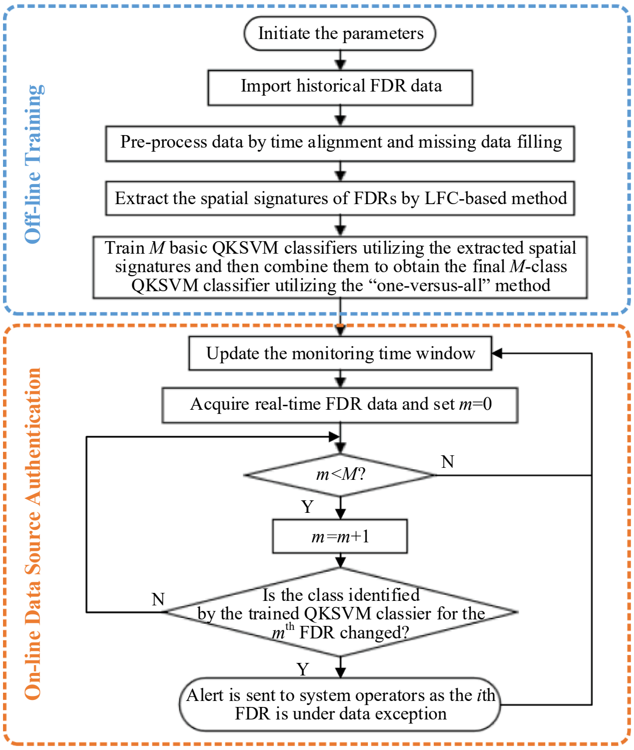

Off-line Training
Initiate the parameters
Import historical FDR data
Pre-process data by time alignment and missing data filling
Extract the spatial signatures of FDRs by LFC-based method
Train M basic QKSVM classifiers utilizing the extracted spatial signatures and then combine them to obtain the final M-class QKSVM classifier utilizing the "one-versus-all" method
On-line Data Source Authentication
Update the monitoring time window
Acquire real-time FDR data and set m=0
m<M?
N
Y
m=m+1
N
Is the class identified by the trained QKSVM classier for the m^th FDR changed?
Y
Alert is sent to system operators as the ith FDR is under data exception

Fig. 2. Flow chart of the proposed LFC-based spatial signature extraction method and QKSVM-based data source authentication algorithm.

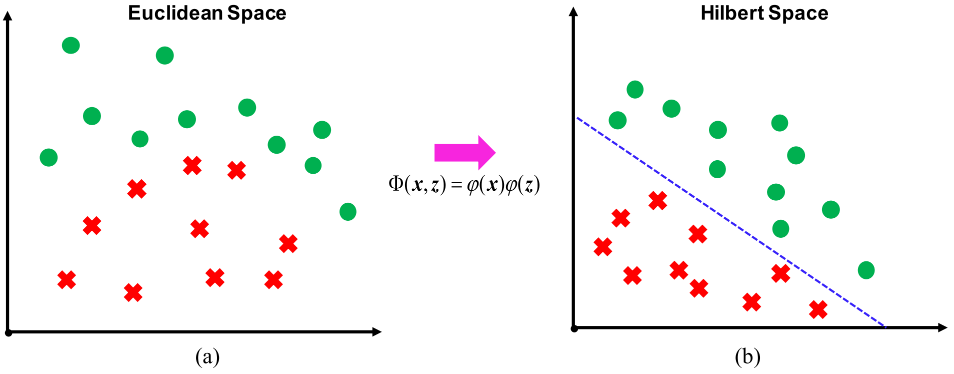


Fig. 1. Schematic diagram for SVM with a kernel function. (a) Euclidean space before the transformation. (b) Hilbert space after the transformation.

space. It can be noted from problem (12) that the variables associated with data samples are only the inner product$\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$. In fact, the $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$ can be viewed as a linear kernel function (i.e., $\Phi(\boldsymbol{x}_i,\boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$) for the traditional SVM, and it can be replaced by the quadratic kernel function $\Phi(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + 1)^2$ directly [26] to obtain the QKSVM format as follows.

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\alpha_i\alpha_j y_i y_j \Phi(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{i=1}^{N_s}\alpha_i \\ &\text{s.t.} \sum_{i=1}^{N_s}\alpha_i y_i = 0,\ 0\leqslant\alpha_i\leqslant\eta,\ i = 1,2,\dots,N_s \end{aligned} \tag{13}$$

It should be clarified that: data source authentication is a multi-class classification problem in this work although the equations (11)-(13) are the derivations for a basic QKSVM that can only perform binary classification. In fact, the derivations of the basic QKSVM classifier are the key points for the classification problem, and once the basic QKSVM classifiers can be obtained, they can be easily combined by the "one-versus-all" method [27] to solve the multi-class classification problem.

The integrated flow chart of the proposed LFC-based spatial signature extraction method and QKSVM-based data source authentication algorithm is given in Fig. 2. In the off-line stage, historical data are pre-processed and employed for extracting spatial signatures, then the QKSVM-based detector is trained based on a large number of cases. In the on-line stage, the well-trained QKSVM detector will be utilized for data source monitoring. In practical applications, an alert will be triggered and corresponding information can be sent to system operators once the data exception is detected.

5. Case studies

The proposed LFC-based spatial signature extraction method and QKSVM-based data source authentication algorithm are tested by actual measured data in two actual U.S. power systems and examples for practical applications are also given. In addition to that, comprehensive discussions associated with different kernel functions of SVM, the combinations of feature extraction methods and machine learning algorithms are given in detail, which demonstrate the advantages of the proposed LFC-QKSVM-based data source authentication algorithm. In Section 5.1, the data descriptions and testing environment are introduced. In Sections 5.2 and 5.3, the studies on data authentication of every FDR pair are performed, which means that the data exception faults are tested in different locations. In Section 5.4, two scenarios for two different types of data exceptions (i.e., only two data sources of FDRs are mixed and multiple sources of FDRs are mixed) are analyzed for demonstrating the effectiveness of the proposed algorithm.

5.1. Descriptions of the measurement data used for case studies

To demonstrate the effectiveness of the proposed LFC-based spatial signature extraction method and the QKSVM-based data source authentication algorithm, two cases with actual data collected from FDRs in FNET/GridEye are employed. In this work, the data are collected in the period from 2019/09/06 00:00:00.000 to 2019/09/06 23:59:59.900 with a 10 Hz reporting rate. Besides, these data are collected from 15 FDRs in the U.S. Western Interconnection Grid and 54 FDRs in the U.S. Eastern Interconnection Grid, respectively.

According to common practice, 80% and 20% of the datasets are assigned as the training and test sets and all cases are performed on MATLAB 2020a installed on Windows 10 platform with Core i7-9700 CPU and 16 GB RAM. It is worth mentioning that the spatial signatures of FDRs in different regional systems are significantly different, so it is not difficult to detect the data source ID mix exceptions among different regional systems. Hence, this work focuses on the more difficult one, i.e., data source ID mix exceptions within the same regional systems. Thus, two cases for the U.S. Western and Eastern Interconnection

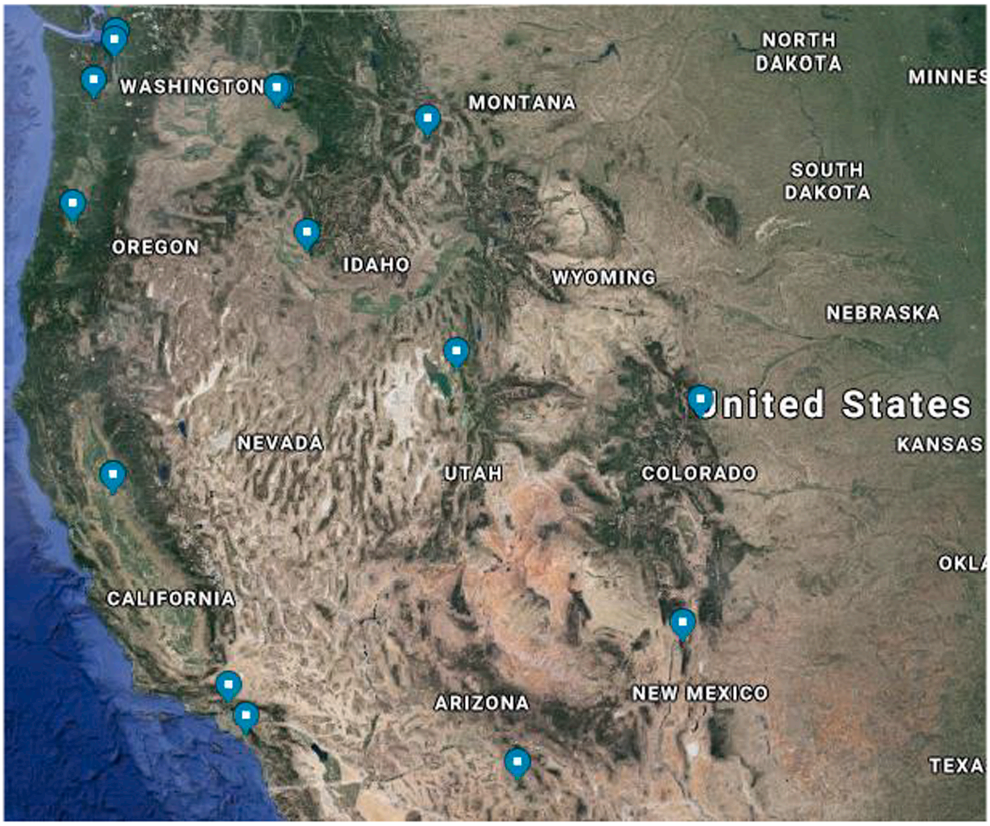


Fig. 3. Locations of 15 FDRs in the U.S. Western Interconnection Grid.

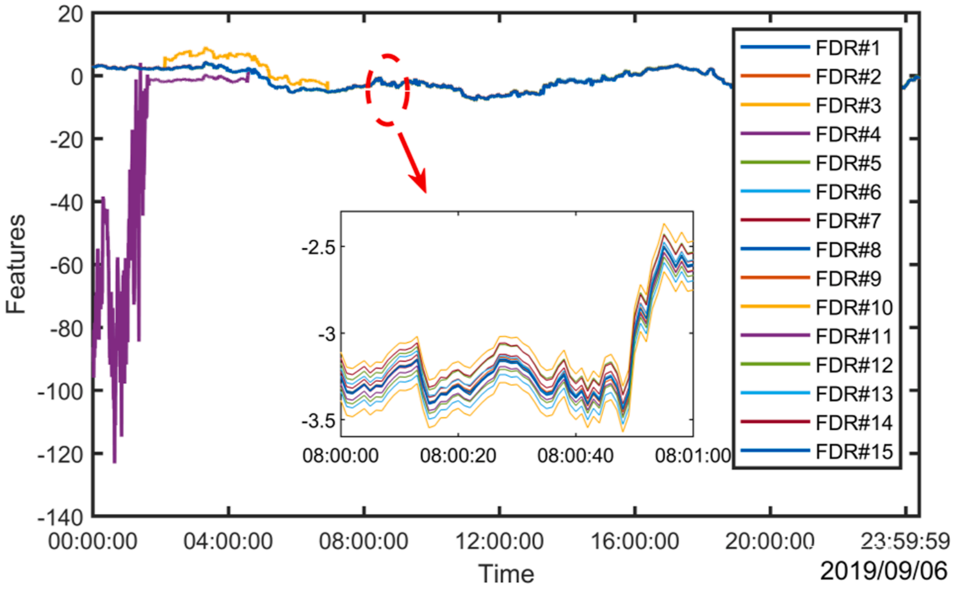


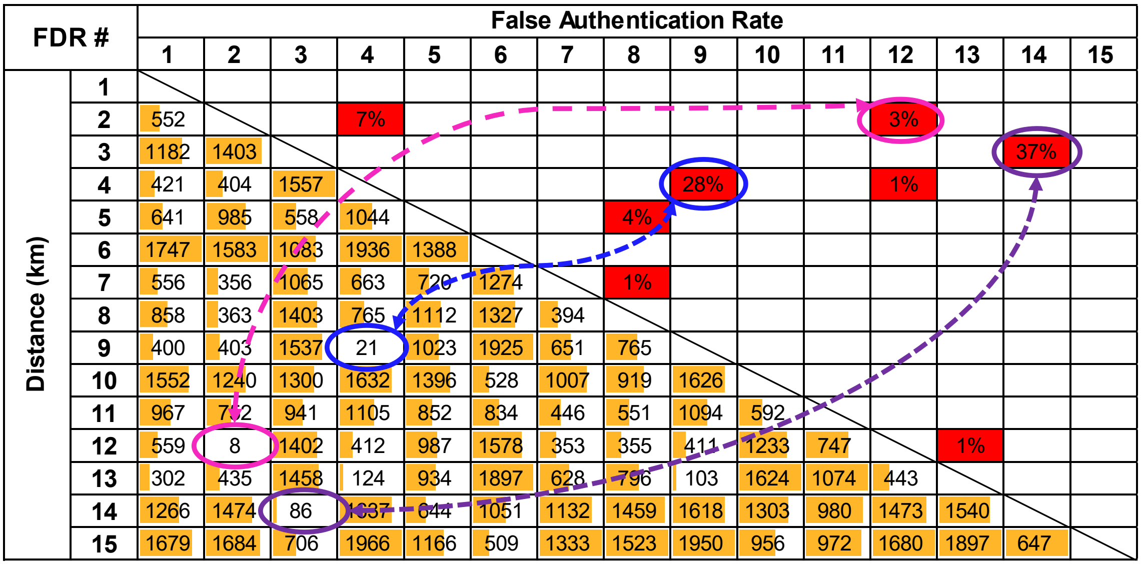
Fig. 4. Extracted spatial signatures of different FDRs in the U.S. Western Interconnection Grid.

Grids are studied separately as follows.

5.2. *Case studies and comparisons between the proposed algorithm and existing algorithms for the U.S. Western Interconnection Grid*

After pre-processing, the data of 15 FDRs on 2019/09/06 are available in the U.S. Western Interconnection Grid, and their locations are shown in Fig. 3.

First, the LFC-based spatial signature extraction method is performed with the parameter $L = 10$. For better illustration, only the 1st spatial signature of FDRs #1–15 is plotted in Fig. 4. Although their variation tendencies are similar at the first sight, there still are some differences among these FDRs in the zoom-in view of the feature figures, which can help to identify each FDR. For example, their relative ranks are unchanged for the 1st feature. Thus, the relative ranks of feature values can be utilized as an index to detect the data exceptions. For example, once the data source ID of FDRs is mixed or changed, the relative ranks of FDRs will be changed and a corresponding alert can be triggered in this case. It should be mentioned that 150 features are extracted by the proposed LFC-based method for each FDR, and they are considered comprehensively and utilized as the input data for the QKSVM-based data source authentication algorithm to determine whether there is a data source ID mix exception. It should be clarified that it is not trying to emphasize the similarity of the characteristics of FDRs in different locations. On the contrary, although their variation tendencies are similar at the first sight, there still are some differences among these FDRs in the zoom-in view of the feature figures, which can help to identify each FDR. For this case, the penalty coefficient η is tuned as 0.37714 by the GridSearchCV method [25] and historical data for the implementation of data source authentication, which can achieve relatively robust and good results. The results of QKSVM-based data source authentication are shown in Fig. 5. In Fig. 5, the numbers in the lower-left corner denote the distances between FDR pairs and the lengths of yellow bars describe the relative values of the distances. For example, the first number “552” means the distance between FDR pair 1&2 is 552 km. The numbers in the upper-right corner denote the false authentication rates associated with FDR pairs, and the blanks mean the false authentication rates are smaller than 1%. It is noted that the false authentication rate for the m^{th}



FDR #		False Authentication Rate														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Distance (km)	1															
	2	552			7%								3%			
	3	1182	1403												37%	
	4	421	404	1557						28%			1%			
	5	641	985	558	1044				4%							
	6	1747	1583	1083	1936	1388										
	7	556	356	1065	663	720	1274		1%							
	8	858	363	1403	765	1112	1327	394								
	9	400	403	1537	21	1023	1925	651	765							
	10	1552	1240	1300	1632	1396	528	1007	919	1626						
	11	967	752	941	1105	852	834	446	551	1094	592					
	12	559	8	1402	412	987	1578	353	355	411	1233	747		1%		
	13	302	435	1458	124	934	1897	628	796	103	1624	1074	443			
	14	1266	1474	86	1637	644	1051	1132	1459	1618	1303	980	1473	1540		
	15	1679	1684	706	1966	1166	509	1333	1523	1950	956	972	1680	1897	647	

Fig. 5. Results of data source authentication in the U.S. Western Interconnection Grid.

FDR is defined as $R_{\mathrm{false},m} = N_{\mathrm{FP},m}/(N_{\mathrm{FP},m} + N_{\mathrm{TN},m})$, where $N_{\mathrm{FP},m}$ and $N_{\mathrm{TN},m}$ are the numbers of false-positive and true-negative samples of the m^{th} FDR, respectively. Generally, the smaller the $R_{\mathrm{false},m}$ is, the better the authentication algorithm is. It can be seen that the data source authentication by the proposed algorithm achieves quite high authentication accuracy. For most FDR pairs, they can be authenticated correctly with more than 99% probability. The worst result happens between the FDRs #3 and #14, and they could be authenticated incorrectly from each other with 37% probability. In summary, the proposed LFC-based spatial signature extraction method and QKSVM-based data source authentication algorithm for FDRs in the U.S. Western Interconnection Grid can achieve 89.60% accuracy.

It can also be seen from Fig. 5 that the distances between different FDRs do influence the authentication accuracy. For instance, the data source authentication result mistakes data sources for FDR pair 3&14 with a 37% probability (see the purple circles in Fig. 5); and mistakes data sources for FDR pair 4&9 with a 28% probability (see the blue circles). Correspondingly, the distance between FDR pair 3&14 is 86 km and the distance between FDR pair 4&9 is 21 km, which are very close. However, the distance between FDR pair 2&12 is 8 km while the corresponding false authentication rate is only 3% (see the pink circles in Fig. 5), which is quite low. It is noted that the tests are performed several times and the results are similar, i.e., the distance between FDRs does influence the authentication accuracy while other impact factors also influence the accuracy. These factors include electrical distance and inherent signatures of devices, which deserve further studies.

Specifically, the inherent signatures of devices indicate the unique characteristics of each FDR, such as the measurement noise and measurement variance. For example, even two FDRs located in the same place would have a slight difference in the measurement data since these two FDRs cannot have the completely same measurement characteristics. Due to the different electronic components and product batches, the measurement noise would be different and their variances would be different too. These characteristics are only associated with each certain FDR, so they are called inherent signatures of devices here and can be utilized for the data authentication of FDRs. Similarly, in modern WAMSs, the apparatuses are different and from different vendors. If the measurement devices are heterogeneous or have different accuracy levels, then these differences can be regarded as more distinct features and can be more useful for authenticating the data sources. In other words, data source authentication can be achieved earlier in such situations. Therefore, this work mainly forces on the more difficult situations, i.e., studying the data source authentication algorithm for homogeneous measurement devices with the similar reporting rate and accuracy level in WAMS.

To demonstrate the effectiveness and superiority of the proposed algorithm, several existing data source authentication algorithms (i.e., algorithms based on MM-gcForest [5], MM-RFC [18] and DWT-BP [28]) are employed for comparisons, whose results are given in Table 1. It should be noted that the values listed in Table 1 are the average values of the data source authentication for different FDR pairs (i.e., different locations in the U.S. Western Interconnection Grid). It can be seen that the accuracy obtained by other algorithms is much lower and cannot be used in practical applications. It should be clarified that most of the existing algorithms have a relatively good performance if the number of FDRs is not large. For example, if only three or five FDRs need to be authenticated, then the algorithms based on DWT-BP [28] or MM-RFC [18] can achieve about 75%~85% accuracy. In the meantime, the MM-gcForest [5] algorithm is greatly improved from [18] and can achieve relatively high accuracy even for ten FDRs. To compare the proposed algorithm with the advanced machine learning algorithm that can extract features automatically, the result of the CNN-based algorithm is also given in Table 1. It can be seen that the CNN-based algorithm performs better than the existing ones while can only achieve 69.2% accuracy either. In this work, the LFC-based feature extraction method is utilized for raw data before performing QKSVM algorithm. For CNN, however, the raw data are directly utilized as the input data of the network. Although CNN also has the ability for feature extraction, it is more suitable for feature extraction for pictures. In addition, the recall rate and F-1 score determined by different algorithms are further given in Table 1 to evaluate the performances of different algorithms for such a class imbalance classification problem. It can be seen that the proposed LFC-QKSVM algorithm can achieve the best results with regard to different evaluation indexes. Hence, it can be concluded that the proposed QKSVM-based algorithm combined with the LFC-based spatial signature extraction method outperforms the other four algorithms.

Table 1
Evaluation Indexes of Data Source Authentication in the U.S. Western Interconnection Grid by Using Different Algorithms.

Index	MM-gcForest [5]	MM-RFC [18]	DWT-BP [28]	CNN	Proposed LFC-QKSVM
Accuracy	40.62%	43.19%	34.35%	69.20%	89.60%
Recall	6.54%	44.44%	37.34%	42.86%	76.92%
F-1	10.55%	58.48%	47.58%	49.34%	79.37%

DWT: Discrete Wavelet Transform MM: Mathematical Morphology BP: Back Propagation.
gcForest: multi-grained cascade Forest CNN: Convolutional Neural Network.

5.3. Case studies and comparisons between the proposed algorithm and existing algorithms for the U.S. Eastern Interconnection Grid

To show the effectiveness of the proposed LFC-based spatial signature extraction method and QKSVM-based data source authentication algorithm for a large number of FDRs, the data of the U.S. Eastern Interconnection Grid are utilized. After pre-processing, the data from 54 FDRs are available in the U.S. Eastern Interconnection Grid on 2019/09/06, and their locations are shown in Fig. 6. Similar to Fig. 4, the LFC-based spatial signature extraction method is performed with the parameter $L = 3$ first, then the QKSVM-based algorithm is used. For this case in the U.S. Eastern Interconnection Grid, the penalty coefficient η is tuned as 3.736 by GridSearchCV method [25] and historical data for the implementation of data source authentication, which can achieve relatively robust and good results. The average accuracy of the proposed algorithm for data source authentication in the U.S. Eastern Interconnection Grid is 80.12%, and the results obtained by other existing algorithms are also given in Table 2 for comparisons.

It can be seen that the accuracies of all algorithms decrease compared with the results obtained in the U.S. Western Interconnection Grid due to the increase in the number of FDRs. For the proposed LFC-QKSVM algorithm, the accuracy only decreases from 89.6% to 80.12% while the accuracies of the other three existing algorithms decrease much more. The CNN-based algorithm works better than the existing ones while can only achieve 64.3% accuracy too. In addition, the recall rate and F-1 score determined by different algorithms are further given in Table 2. It should be noted that the values listed in Table 2 are the average values of the data source authentication for different FDR pairs (i.e., different locations in the U.S. Eastern Interconnection Grid). It can be seen that the proposed LFC-QKSVM algorithm can achieve the best results with regard to different evaluation indexes. Therefore, it can be concluded that the proposed LFC-QKSVM algorithm earns a great improvement compared with other algorithms especially when a large number of data of FDRs are required to be authenticated.

5.4. Examples of practical applications

In order to illustrate the proposed algorithm in practical applications, two scenarios in the U.S. Western and Eastern Interconnection Grids are studied as examples. Scenario 1 is utilized to demonstrate the effectiveness of the proposed data source authentication algorithm when only two data sources of FDRs are mixed. Scenario 2 is utilized to

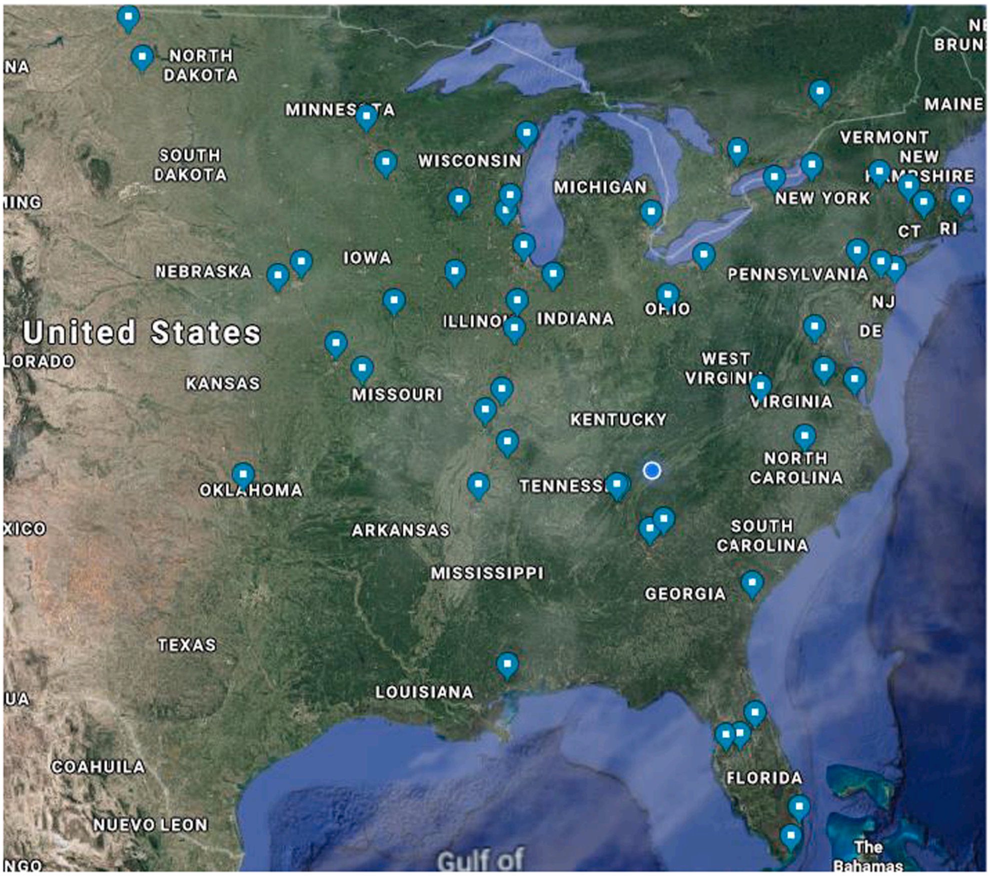


Fig. 6. Locations of 54 FDRs in the U.S. Eastern Interconnection Grid.

Table 2
Evaluation Indexes of Data Source Authentication in The U.S. Eastern Interconnection Grid by Using Different Algorithms.

Index	MM-gcForest [5]	MM-RFC [18]	DWT-BP [28]	CNN	Proposed LFC-QKSVM
Accuracy	23.50%	17.33%	19.61%	64.3%	80.12%
Recall	29.17%	57.69%	63.11%	74.35%	86.96%
F-1	8.38%	42.06%	39.28%	52.84%	75.11%

demonstrate the effectiveness of the proposed data source authentication algorithm when multiple sources of FDRs are mixed in large-scale power systems.

5.4.1. *Scenario 1: Data sources of FDRs #3 and #9 in the U.S. Western Interconnection Grid are mixed from 22:00:00 to 22:09:59*

In this scenario, the data from 00:00:00 to 19:59:59 are used as the training set and the data from 20:00:00 to 23:59:59 are used as the testing set. The performances of practical applications of the proposed data source authentication are shown in Figs. 7 and 8, and the off-line training time and on-line authentication time for this scenario are given in Table 3.

In the period of 22:00:00 to 22:09:59, the data sources of FDRs #3 and #9 are mixed intentionally. Due to the space limitation, only the authentication results of FDRs #3 and #9 are given in Figs. 7 and 8. The reporting rate of FDRs is 10 Hz, and the window length of the proposed algorithm is 1 s (i.e., every 10 samples of frequency data are used to authenticate FDR sources once). Therefore, there are 4*3600 = 144,000 points in Fig. 7 or 8. The blue circles and green rhombuses denote the data source authentication results for FDRs #3 and #9, respectively. Besides, the detailed views for the beginning and ending periods of data exception are shown in the sub-boxes in Figs. 7 and 8, respectively. It can be seen that data from FDRs #3 and #9 are correctly authenticated as the data from sources #3 and #9 for most of the time from 20:00:00 to 21:59:59 and 22:10:00 to 23:59:59 (see the left and right parts of Figs. 7 and 8). In the period of data exception (i.e., 22:00:00 ~ 22:09:59), it can be seen from the monitoring graph that the mixed data from FDRs #3 and #9 are correctly identified as the data from sources #9 and #3, indicating a data exception that exchanges the data sources of FDRs #3 and #9 is ongoing (see the middle parts of Figs. 7 and 8). It is noted that although there could be some authentication disturbances at the beginning and end of the data exception, the authentication result would be stabilized later on, as can be seen from the detailed views of Figs. 7 and 8. Therefore, it can be concluded that the proposed algorithm is effective in practical applications.

As for the computation time, it can be seen from Table 3 that the proposed LFC-QKSVM spends 322.63 s for training while 0.0651 s for authentication. Although the off-line training stage requires a relatively long time, the on-line authentication stage just requires a quite short time. Therefore, the proposed LFC-QKSVM can meet the real-time requirement.

5.4.2. *Scenario 2: Data sources of FDRs #8, #30 and #50 in the U.S. Eastern Interconnection Grid are mixed from 22:01:00 to 22:02:59*

In this scenario, the data from 00:00:00 to 19:59:59 are used as the training set. The data from 20:00:00 to 23:59:59 are used as the testing set and the performances of practical applications are shown in Fig. 9. In Fig. 9, the blue stars, orange triangles and yellow crosses circles denote the data source authentication results for FDRs #8, #30 and #50, respectively. In the period of 23:01:00 to 23:02:59, the data sources of FDRs #8, #30 and #50 are mixed intentionally. To show the results clearly, only the identified sources of FDRs #8, #30 and #50 are given

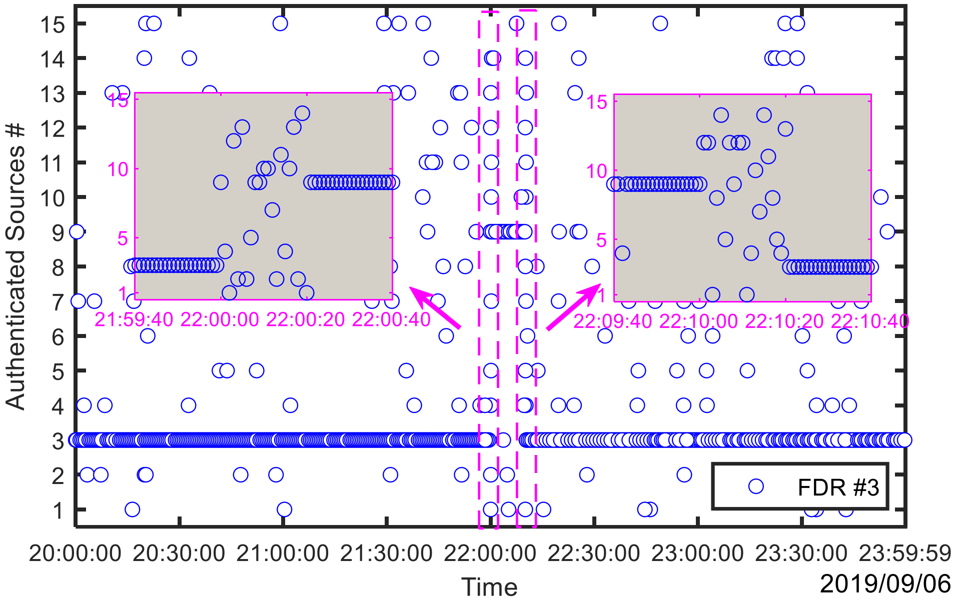


Fig. 7. Data source authentication for FDR #3 in the U.S. Western Interconnection Grid in practice.

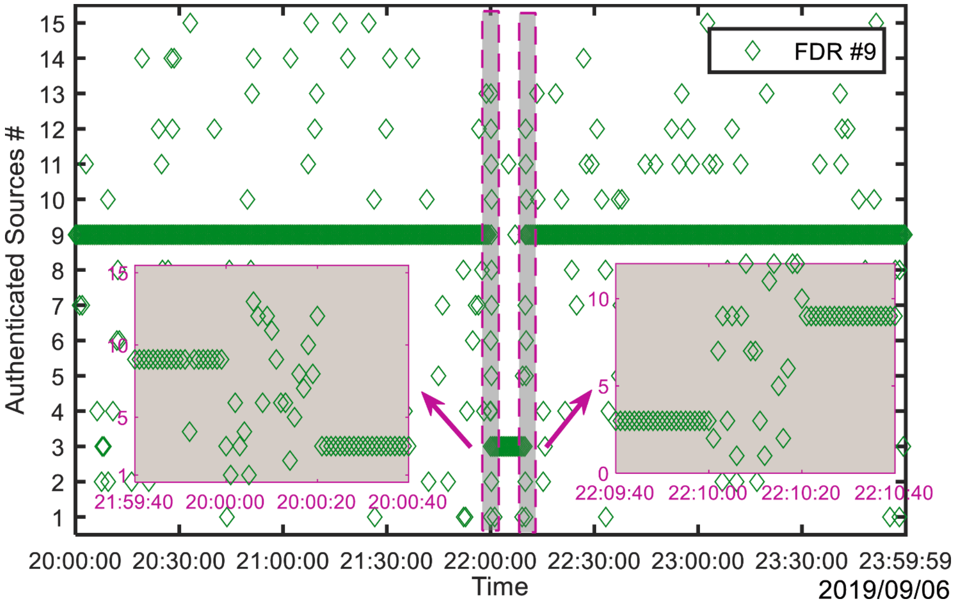


Fig. 8. Data source authentication for FDR #9 in the U.S. Western Interconnection Grid in practice.

Table 3
Off-line Training time and On-line Authentication Time of the Proposed LFC-QKSVM Algorithm for Scenario 1.

Algorithm	Window Length	Off-line Training Time	On-line Authentication Time
MM-gcForest [5]	10 min	59.32 s	0.0524 s
MM-RFC [18]	10 min	18.64 s	0.0073 s
LFC-QKSVM	1 s	322.63 s	0.0651 s

in Fig. 9. Besides, the off-line training time and online authentication time for this scenario are given in Table 4.

It can be seen from Fig. 9 that data from FDRs #8, #30 and #50 are correctly authenticated as in sources #8, #30 and #50 when no data exception arises (i.e., 23:00:00 ~ 23:00:59 and 23:03:00 ~ 23:05:00). In the period of data exception (i.e., 23:01:00 ~ 23:02:59), it can be seen from the monitoring graph that the data from mixed FDRs #8, #30 and #50 are respectively recognized as sources #30, #50 and #8, which indicates a data exception happened and the data sources of FDRs #8, #30 and #50 are mixed from each other. Therefore, it can be concluded that the proposed algorithm is also effective when applied for large-scale power systems with a large number of FDRs.

As for the computation time, it can be seen from Table 4 that the proposed LFC-QKSVM spends 1565.38 s for training while 0.3433 s for authentication. Although the time of the on-line stage increase with the increase of the number of FDRs, the on-line authentication time is still within 1 s and is acceptable for on-line application. It is noted that the window lengths of MM-gcForest and MM-RFC algorithms are 10 min while the window length of the proposed LFC-QKSVM is 1 s. Therefore, the data samples for MM-gcForest and MM-RFC algorithms are fewer than the data samples for the LFC-QKSVM algorithm, although the data points in each sample for MM-gcForest and MM-RFC algorithms are

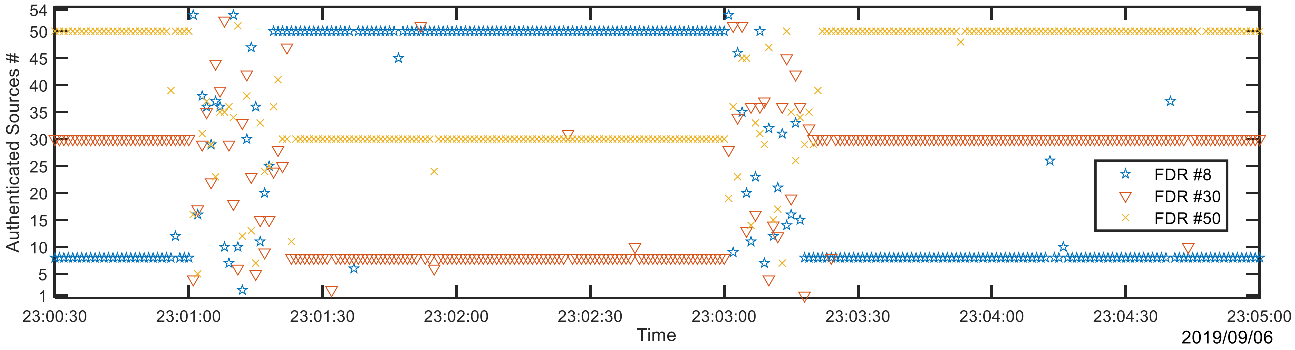


Fig. 9. Monitoring graph for FDRs #8, #30 and #50 in the U.S. Eastern Interconnection Grid in practice.

Table 4
Off-line Training time and On-line Authentication Time of the Proposed LFC-QKSVM Algorithm for Scenario 2.

Algorithm	Window Length	Off-line Training Time	On-line Authentication Time
MM-gcForest [5]	10 min	186.90 s	0.1654 s
MM-RFC [18]	10 min	65.45 s	0.0269 s
LFC-QKSVM	1 s	1565.38 s	0.3433 s

more than the data points in each sample for the LFC-QKSVM algorithm. In fact, the data sample and data point are different concepts. The relationship between them is: one data sample is described by several features as shown in equation (10), while the features are extracted based on a large number of original data points that belong to this data sample. It can be seen that the number of data samples of the LFC-QKSVM algorithm utilized for training is much more than the ones of MM-gcForest and MM-RFC algorithms, although the points in each sample are fewer. Therefore, the computation time of the LFC-QKSVM algorithm shown in Tables 3 and 4 is much longer than MM-gcForest and MM-RFC algorithms.

Indeed, the training time of the proposed algorithm is much higher than the other two algorithms in Tables 3 and 4, while it is worth for the improvement of data source authentication accuracy as shown in Tables 1 and 2. Besides, it should be clarified that the computation time of the proposed LFC-QKSVM algorithm shown in Tables 3 and 4 is not obtained by running in parallel for each FDR. On the one hand, the trained model can be used for a quite long time as long as the power system model does not have large change and the old FDRs are not replaced by new ones. Therefore, the LFC-QKSVM model do not need to be updated too frequently, and once for several days is enough. Hence, the training time is feasible in practical applications. On the other hand, there are several methods that can be utilized to accelerate the computation process. If more FDRs are required to be authenticated in the future, dedicated servers instead of a single desktop can be further deployed. Furthermore, parallel techniques can be utilized as well to improve the training efficiency for QKSVM.

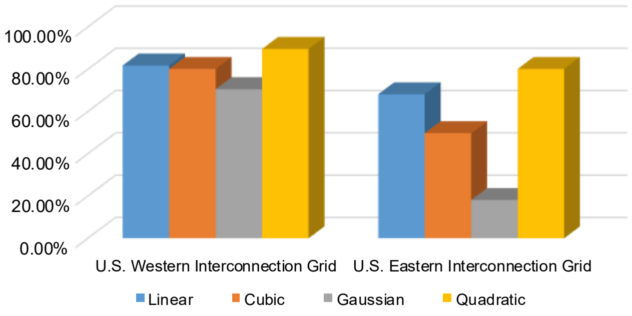


Fig. 10. Comparisons among different kernel functions for SVM.

6. Discussions

6.1. Discussions on the kernel functions of SVM

In fact, several kernel functions can be applied for SVM. To select the best kernel function for the proposed data source authentication algorithm, linear, cubic, Gaussian and quadratic kernel functions are compared. Their results in the U.S. Western and Eastern Interconnection Grids are given in Fig. 10.

It can be seen that the accuracy obtained by Gaussian kernel function is the lowest; the ones obtained by polynomial (i.e., linear, quadratic, and cubic) kernel functions are all higher and the quadratic kernel function achieves the best performance. The reasons are that i) the extracted spatial signatures may not be in Gaussian distribution; ii) under-fitting and over-fitting problems may exist if linear or cubic functions are used. It can be seen that the quadratic kernel function can always achieve high accuracy, indicating that the quadratic kernel function is the most suitable kernel function for SVM in data source authentication. To be honest, it is hard to give an affirmative conclusion that whether there is an over-fitting problem when using the quadratic kernel function. However, it can be seen that the results obtained by the quadratic kernel function are better when comparing the results obtained by linear and cubic kernel functions. Therefore, it can be at least concluded that the influence of the under/over-fitting problem when using the quadratic kernel function is much lighter than the influence when using the linear or cubic kernel function. In other words, it is suggested to use QKSVM for data source authentication in this work.

6.2. Discussions on the importance of cooperation between LFC-based spatial signature extraction and QKSVM

To demonstrate the importance of cooperation between LFC-based spatial signature extraction and QKSVM, comprehensive comparisons among different combinations of feature extraction methods and machine learning algorithms are performed, and their results in the U.S. Western and Eastern Interconnection Grids are given in Figs. 11 and 12, respectively. The x-axis denotes different machine learning algorithms (i.e., gcForest, RFC, BP and QKSVM) and the y-axis denotes different feature extraction methods (FFT, DWT, MM and LFC). It is noted that all the tests are performed on the same data set with the same window length (i.e., 1 s) and reporting rate (i.e., 10 Hz). The results of using original frequency measurements (i.e., feature extraction is not performed before utilizing classifiers) are also given in Figs. 11 and 12 for comparisons. It can be seen that using feature extraction methods can always achieve better performance than using the original frequency measurements, which also demonstrates the necessity of feature extraction.

It can be seen from Figs. 11 and 12 that: i) The LFC-based spatial

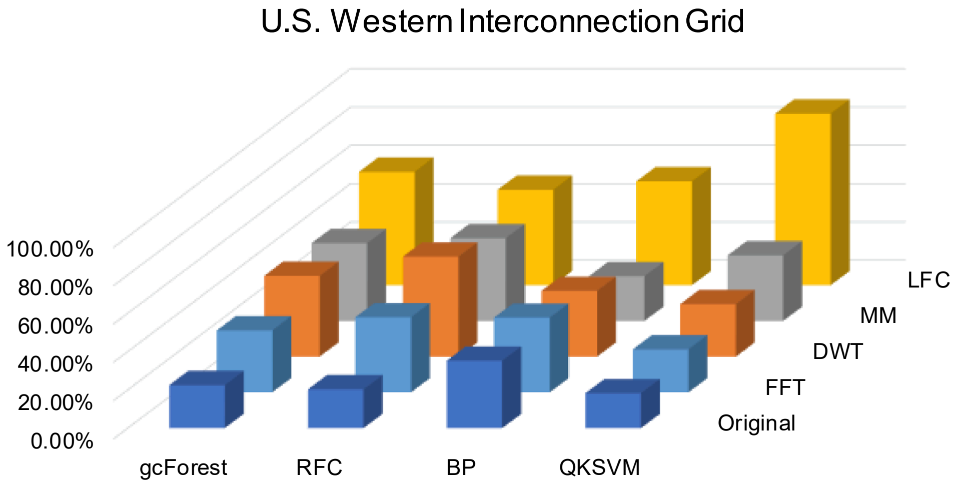


Fig. 11. Accuracies in the U.S. Western Interconnection Grid by using different feature extraction methods and machine learning algorithms.

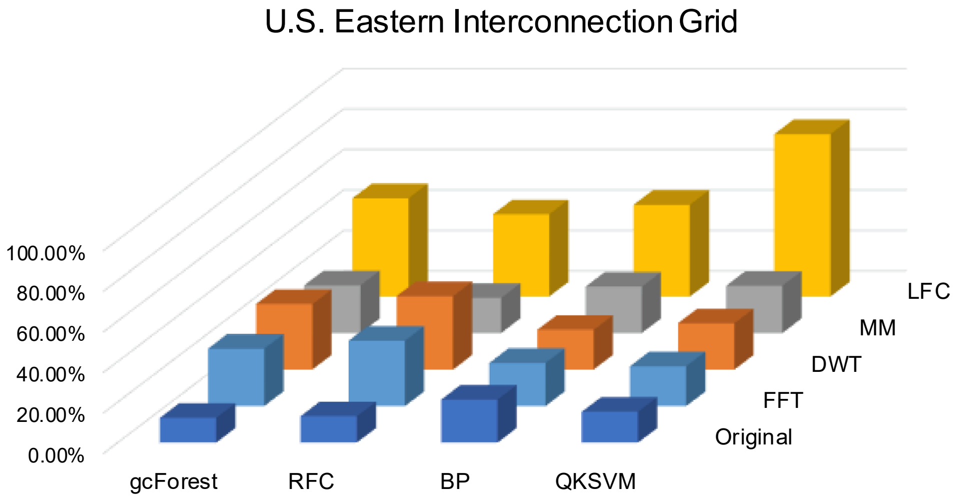


Fig. 12. Accuracies in the U.S. Eastern Interconnection Grid by using different feature extraction methods and machine learning algorithms.

signature extraction method can always help achieve the highest accuracy no matter what machine-learning algorithms are followed. ii) The QKSVM-based data source authentication algorithm can achieve the highest accuracy among different machine learning algorithms when combined with LFC. iii) The performance of the QKSVM-based data source authentication algorithm is inferior to other ones if its input features are extracted by the other three feature extraction methods, which means that it is quite important to utilize LFC and QKSVM together to achieve the best performance for data source authentication in this situation.

The reasons behind the improvement of accuracy are that: i) the quality of feature extraction can greatly influence the effectiveness of machine-learning algorithms, and using original data means no feature extraction is performed before data authentication. ii) Although the feature extraction methods such as FFT, DWT and MM have good performance for signal processing and other fields, they do not consider the special characteristics of power systems. Therefore, they may not suitable for data source authentication for power systems and cannot achieve as good results as in other fields. iii) The proposed LFC-based feature extraction method considers the inherent load–frequency characteristics of power systems, so it is more suitable to extract the spatial signatures from the measured data. In other words, the proposed LFC-based feature extraction method considers the special characteristics of power systems so as to achieve better performances than others.

6.3. Discussions on the impact of data reporting rate

In the previous study (i.e., Ref. [18]), it is stated that the MM-RFC algorithm can achieve high accuracy even when two SMDs are located several miles away, which seems to conflict with the phenomena in this work. However, it should be clarified that the data reporting rate of PMUs (i.e., 120 Hz) in Ref. [18] is much higher than the one of FDRs (i.e., 10 Hz) in this work. If a lower reporting rate is utilized, the performances of the MM-gcForest and MM-RFC algorithms will decrease dramatically (see Fig. 4 in [18]) for the SMDs located closely. To demonstrate this point, tests with the 10 Hz reporting rate in the same system as [18] are utilized and the results are: the LFC-QKSVM algorithm can achieve 94.12% accuracy while the MM-RFC algorithm can

Table 5
Evaluation Indexes of Data Source Authentication in The U.S. Western and Eastern Interconnection Grids by Using Different Parameters *L* for the LFC-Based Extraction Method.

Index	$L = 10$		$L = 3$	
	Western	Eastern	Western	Eastern
Accuracy	89.60%	81.53%	88.10%	80.12%
Recall	76.92%	87.24%	75.32%	86.96%
F-1	79.37%	76.85%	77.67%	75.11%

achieve about 48% accuracy. Therefore, it is the difference in reporting rate that causes the uncoordinated phenomena between Ref. [18] and this work. In fact, this work aims to authenticate the data source of a large number of FDRs in bulk power systems using low-reporting rate measurement data with a relatively short time window. If high-reporting data are available, the algorithms [3,5,18] in the previous work can be employed to authenticate the data source of SMDs that are located close to each other.

6.4. Discussions on the impact of parameter L for the LFC-Based extraction method

Generally, the larger L could help to extract more signatures and achieve high accuracy. However, it also results in more equations to be solved and more computation time. Therefore, to obtain the trade-off between authentication accuracy and computation time, it is recommended to select a relatively large L for power systems with a small number of FDRs to be authenticated and select a relatively small L for power systems with a large number of FDRs to be authenticated. To demonstrate that the deterioration of evaluation indexes is mainly caused by the increase of FDR number rather than the parameter L, the test results on the U.S. Western Interconnection Grid with $L = 3$ and the test results on the U.S. Eastern Interconnection Grid with $L = 10$ are also given in Table 5 for comparisons. It can be seen that the results obtained by $L = 10$ are slightly better than the results obtained by $L = 3$ for both the two interconnection grids. However, even the parameter $L = 10$ is used for the U.S. Eastern Interconnection Grid, its results are still worse than the one of the U.S. Western Interconnection Grid with $L = 3$. Therefore, it can be concluded that performance decrease is mainly caused by the increase of FDT number rather than the reduction of L from 10 to 3 in Section 5.

6.5. Discussions on the impact of data exception in practice

From the power system application perspective, lots of potential impacts would be caused if the data source information is mixed. For example, the error of event location estimation for power systems would be greatly increased. At 2019–10-19 21:59:03, there was a generation trip event occurs in the U.S. Eastern Interconnection Grid and the FDRs *UsOhAkron998*, *UsNyLewiston1436*, *UsNyLeroy985*, *CaOnToronto703* and *UsOhChilliecothe670* are the first five FDRs that received the electromechanical waves of the frequency drop. Based on the different delays of wave arrival time, the event location can be estimated as the GPS coordinate (43.6103, −83.4856), which is 92.21 miles away from the actual location (42.3048, −83.1527). However, if the data source information of FDRs is mixed, then the estimation errors will increase dramatically. For example, if the source information of *UsOhAkron998* and *UsOhChilliecothe670* is mixed, then the estimated location would be (45.6578, −84.5223), which is 242.75 miles away from the actual location. If the source information among more FDRs is mixed, the errors could increase even up to 800 miles. Therefore, data exception could cause a great impact in practice and data source authentication is meaningful for practical applications in power systems.

6.6. Discussions on the relationship between data exception detection and data spoofing prevention

It should be mentioned that in addition to detecting data exceptions, preventing data spoofing is another essential point for data source authentication. For data spoofing prevention, there are several mature techniques and protocols. For example, the secure sockets layer (SSL), also known as transport layer security (TLS) later, is used for data transmission on websites. SSL consists of a record layer and a transport layer. The record layer protocol determines the encapsulation format of the data in the transport layer. The transport layer security protocol uses the asymmetric encryption calculus to authenticate the communication party, after which the symmetric key is exchanged as the session key [30]. Besides, other asymmetric cryptographic/authentication algorithms such as RSA and Elgamal can also be utilized for preventing data spoofing. However, on the one hand, these asymmetric cryptographic/authentication algorithms need additional information during the data transmission, which cannot be employed for the current FDRs directly. On the other hand, they mainly aim to prevent data spoofing rather than detect ongoing data exceptions. This work mainly focuses on detecting data exceptions and authenticate the data sources only based on the measurement data, while how to prevent data spoofing is not the main point of this work. Therefore, the asymmetric cryptographic/authentication algorithms are considered in detail in this work.

7. Conclusions

In this work, a novel feature extraction method based on LFC and a data source authentication algorithm based on QKSVM are proposed. First, the spatial signatures of FDRs located in different regions are extracted. Then, the QKSVM algorithm is employed to authenticate the data source of each FDR. Case studies in the U.S. Western and Eastern Interconnection Grids demonstrate the effectiveness of the proposed algorithm in authenticating measurement data in actual systems. Some conclusions can be drawn as follows.

i) The cooperation of the proposed LFC-based spatial signature extraction method and the QKSVM-based data source authentication algorithm can achieve higher authentication accuracy compared with existing algorithms. The proposed data source authentication algorithm can achieve higher accuracy with a much shorter time delay.

ii) It is essential to determine a suitable kernel function for SVM, which has large impact on the final authentication result. Concretely, linear and cubic/Gaussian kernel functions might respectively cause under-fitting and over-fitting problems in this case, so the most suitable kernel function is determined as the quadratic one in this work.

iii) The highest authentication accuracy for data sources can be achieved if the LFC-based method and the QKSVM-based algorithm are utilized simultaneously. Combining one of them with another feature extraction method or machine learning algorithm will reduce accuracy.

iv) The geographic distances among different FDRs have an obvious but not decisive influence on the authentication accuracy of the proposed algorithm. Generally, the authentication accuracy increases with the geographic distance between different data sources. However, this correlation does not always hold due to other factors (i.e., electrical distance and inherent signatures of the devices).

Although several analyses have been done in this work, there are still some issues and work that are required to be further studied as follows in the future.

i) More attention is required to be paid to the inherent signatures of SMDs and their physical meanings. Once this issue can be handled, the performance of data authentication will be much better and more robust.

ii) The module can be operated within the PDCs. More specifically, for the FNET/GridEye system, the LFC-QKSVM module can be operated in the action layer of the OpenPDC framework [29], which is designed to handle multiple FDRs/PMUs. Currently, the OpenPDC in FNET/GridEye has been processing data from more than 300 FDRs in real-time. Our future work will study the computational complexity/overhead of this module in real-time operation on this actual monitoring system.

iii) The results of this work can support building an authenticated dataset of system frequency. With the help of this dataset, electrical network frequency filtered from audios or videos can be compared with the dataset so as to achieve corresponding forensic analysis.

iv) Data exception detection and data spoofing prevention are the two essential points for data source authentication. This work mainly focuses on data exception detection and data spoofing prevention is worth further studying in the future.

CRediT authorship contribution statement

Shengyuan Liu: Conceptualization, Methodology, Writing – original draft. **Shutang You:** Conceptualization, Discussion, Methodology. **He Yin:** Discussion. **Zhenzhi Lin:** Supervision. **Yilu Liu:** Supervision. **Yi Cui:** Discussion. **Wenxuan Yao:** Discussion. **Lakshmi Sundaresh:** Discussion.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Funding for this research was provided by the NSF Cyber-Physical Systems (CPS) Program under award number 1931975. This work also made use of Engineering Research Center shared facilities supported by the Engineering Research Center Program of the National Science Foundation and the Department of Energy under NSF Award Number EEC-1041877 and the CURENT Industry Partnership Program.

References

[1] Phadke AG, Bi T. Phasor measurement units, WAMS, and their applications in protection and control of power systems. J Mod Power Syst Clean Energy 2018;6 (4):619–29.

[2] Liu S, Lin Z, Zhao Y, Liu Y, Ding Y, Zhang B, et al. Robust system separation strategy considering online wide-area coherency identification and uncertainties of renewable energy sources. IEEE Trans Power Syst Sep. 2020;35(5):3574–87.

[3] Liu S, You S, Yin H, Lin Z, Liu Y, Yao W, et al. Model-free data authentication for cyber security in power systems. IEEE Trans Smart Grid Sep. 2020;11(5):4565–8.

[4] Qiu W, Tang Q, Wang Y, Zhan L, Liu Y, Yao W. Multi-view convolutional neural network for data spoofing cyber-attack detection in distribution synchrophasors. IEEE Trans Smart Grid 2020;11(4):3457–68. https://doi.org/10.1109/TSG.2020.2971148.

[5] Cui Y, Bai F, Liu Y, Fuhr PL, Morales-Rodriguez ME. Spatio-temporal characterization of synchrophasor data against spoofing attacks in smart grids. IEEE Trans Smart Grid Sep. 2019;10(5):5807–18.

[6] Fan Y, Zhang Z, Trinkle M, Dimitrovski AD, Song JB, Li H. A cross-layer defense mechanism against GPS spoofing attacks on PMUs in smart grids. IEEE Trans Smart Grid 2015;6(6):2659–68.

[7] IEEE standard for synchrophasor data transfer for power systems, *IEEE Std C37.118.2*, 2011.

[8] Zhang Z, Gong S, Dimitrovski AD, Li H. Time synchronization attack in smart grid: impact and analysis. IEEE Trans Smart Grid 2013;4(1):87–98.

[9] Risbud P, Gatsis N, Taha A. Vulnerability analysis of smart grids to GPS spoofing. IEEE Trans Smart Grid 2019;10(4):3535–48.

[10] Chaojun G, Jirutitijaroen P, Motani M. Detecting false data injection attacks in AC state estimation. IEEE Trans Smart Grid 2015;6(5):2476–83.

[11] Manandhar K, Cao X, Hu F, Liu Y. Detection of faults and attacks including false data injection attack in smart grid using Kalman filter. IEEE Trans Control Network Syst 2014;1(4):370–9.

[12] Moslemi R, Mesbahi A, Velni JM. A fast, decentralized covariance selection-based approach to detect cyber attacks in smart grids. IEEE Trans Smart Grid 2018;9(5): 4930–41.

[13] Sedghi H, Jonckheere E. Statistical structure learning to ensure data integrity in smart grid. IEEE Trans Smart Grid 2015;6(4):1924–33.

[14] Ashok A, Govindarasu M, Ajjarapu V. Online detection of stealthy false data injection attacks in power system state estimation. IEEE Trans Smart Grid 2018;9 (3):1636–46.

[15] Fan X, Du L, Duan D. Synchrophasor data correction under GPS spoofing attack: a state estimation-based approach. IEEE Trans Smart Grid 2018;9(5):4538–46.

[16] Esmalifalak M, Liu L, Nguyen N, Zheng R, Han Z. Detecting stealthy false data injection using machine learning in smart grid. IEEE Syst J Sep. 2017;11(3): 1644–52.

[17] Ozay M, Esnaola I, Yarman Vural FT, Kulkarni SR, Poor HV. Machine learning methods for attack detection in the smart grid. IEEE Transactions on Neural Networks and Learning Systems, Aug 2016;27(8):1773–86.

[18] Cui Y, Bai F, Liu Y, Liu Y. A measurement source authentication methodology for power system cyber security enhancement. IEEE Trans Smart Grid Jul. 2018;9(4): 3914–6.

[19] Liu S, You S, Zeng C, Zhao Y, Yao W, Liu Y, et al. Impact of simultaneous activities on frequency fluctuations: comprehensive analyses based on the real measurement data from FNET/GridEye. CSEE J Power Energy Syst Mar. 2021;7(2):421–31.

[20] Liu S, You S, Lin Z, Zeng C, Li H, Wang W, et al. Data-driven event identification in the U.S. power systems based on 2D-OLPP and RUSBoosted trees. IEEE Trans Power Syst 2022;37(1):94–105.

[21] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies Feb. 2011;2(1):37–63.

[22] V. Vittal, J. D. McCalley, P. M. Anderson, A. A. Fouad, “Power system stability and control,” 3rd ed. New York, NY, USA, Wiley-IEEE Press, 2019.

[23] P. Kundur, “Power system stability and control,” New York, NY, USA, McGraw-Hill, 1994.

[24] Huang H, Li F. Sensitivity analysis of load-damping characteristic in power system frequency regulation. IEEE Trans Power Syst May 2013;28(2):1324–35.

[25] ‘Parameter estimation using grid search with cross-validation’ [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html.

[26] K.P.·Murphy, “Machine learning: a probabilistic perspective,” Cambridge, MA, USA, The MIT Press, 2012.

[27] C., Koby, and Y. Singer. “On the algorithmic implementation of multiclass kernel-based vector machines,” *Journal of Machine Learning Research*, vol. 2, pp. 265-292, Dec. 2001.

[28] Yao W, Zhao J, Till MJ, You S, Liu Y, Cui Y, et al. Source location identification of distribution-level electric network frequency signals at multiple geographic scales. IEEE Access May 2017;5:11166–75.

[29] ‘OpenPDC’ [Online] https://www.gridprotectionalliance.org/docs/products/openpdc/brochure20.pdf.

[30] ‘Transport Layer Security’ [Online] https://en.wikipedia.org/wiki/Transport_Layer_Security.