Received 27 March 2023, accepted 23 April 2023, date of publication 1 May 2023, date of current version 24 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272001

RESEARCH ARTICLE

A Weighted *k*NN Fault Detection Based on Multistep Index and Dynamic Neighborhood Scale Under Complex Working Conditions

XIAOYI QIAN[1], TIANHE SUN[1], BAOSHI WANG[1], AND YUXIAN ZHANG[2]
[1]Liaoning Key Laboratory of Power Grid Energy Conservation and Control, Shenyang Institute of Engineering, Shenyang 110036, China
[2]School of Electrical Engineering, Shenyang University of Technology, Shenyang 110870, China

Corresponding author: Xiaoyi Qian (qianxiaoyi123@163.com)

This work was supported in part by the Natural Science Foundation of Liaoning Province (Doctoral Initiation Funding) under Grant 2022-BS-222, 2022BS-223, and in part by the Basic Research Project of the Educational Commission of Liaoning Province under Grant LJKQZ2021085.

ABSTRACT Fault detection based on k-nearest neighbor (FD-kNN) is one of the most widespread fault detection techniques for industrial processes under complex working conditions, owing to its characteristic of local modeling. However, its state separation ability tends to worsen when the operating data is heterogeneous distribution. To tackle this challenge, a weighted k-nearest neighbor fault detection method based on multistep index and dynamic neighbor scale is proposed. The multistep nearest neighbor index is defined to evaluate the state separation ability, and a weighted k-nearest neighbor fault detection framework is formed by the assigned weights obtained from kernel principal component analysis. On the basis above, a dynamic neighborhood scale correction method and a dynamic threshold setting strategy are proposed to deal with the heterogeneous distribution of operating data and track the abrupt change of the operation state. 10 common faults of wind turbines with complex operation conditions are used to verify the effectiveness of the proposed method.

INDEX TERMS Complex working condition, fault detection, k-nearest neighbor, multistep index, dynamic neighborhood scale, dynamic threshold.

I. INTRODUCTION

Fault detection has received widespread attention in the operation and maintenance of industrial processes due to avoid further faults deterioration [1]. In recent years, fault detection methods can be divided into data-based [2], model-based [3], and knowledge-based methods [4], among which data-based methods have the advantage of not relying on priori models and relevant domain expert knowledge. In addition, with the rapid development of information collection, transmission technology, and data mining methods, data-driven fault detection technology has been widely applied [5], [6].

Data-driven fault detection method contains signal processing [7], [8], statistical analysis [9], [10], and artificial intelligence methods [11], [12], [13]. In [7], a blind source separation method for sparse component analysis is proposed, which can accurately estimate the number of signal sources and further recover signal characteristics from mixed observation signals. In [8], the power spectral density of the relative axial acceleration of the bearing and the nacelle are developed for six fault conditions of different severity. In [9], a three-layer Bayesian network structure based on fault trees is proposed to simplify the Bayesian network structure for fault diagnosis of wind turbines. In [10], a PCA-based multi-criteria feature selection algorithm for the state feature selection of wind turbines is proposed, and then a fault detection model through artificial neural networks is established. In addition, artificial intelligence methods represented by deep learning have obtained significant achievements in the fault detection field in recent years. In [11], a new Multiple-Order Graphical Deep Extreme Learning Machine (MGDELM) algorithm for unsupervised fault

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

diagnosis of rolling bearing is proposed to juggle the manifold structure information with multiple-order similarity from the massive unlabeled industrial data. In [12], a self-adaptation graph attention network via meta-learning (SGANM) is proposed. In [13], a novel self-training semi-supervised deep learning (SSDL) approach is proposed to train a fault diagnosis model with few labeled and abundant unlabeled samples.

A. RELATED WORK

The data-driven fault detection methods have obtained successful achievements in various fields. However, the operating data under complex operating conditions and multi-modes usually have the characteristics of nonlinearity, multi-center, inconsistent covariance structure, etc. [14], such as semiconductor etching process [15], penicillin fermentation process [16], and wind turbine pitch system operation process [17], etc. The traditional solution for the above problems is to divide the operating conditions/modes and establish a distributed fault detection model to deal with the complex operating conditions/modes [18], [19], [20]. In addition, the working conditions division methods still have the disadvantages of blindness, randomness, and poor timeliness.

The fault detection method based on k-nearest neighbors (FD-kNN) was proposed by *He* in [15], which can better deal with the nonlinear problems as the k-nearest neighbor algorithm is a nonlinear classifier. The anomaly identification by kNN only relies on similar operating states among offline data, which can realize the function of local modeling for online samples without setting up a distributed model. Therefore, FD-kNN is more suitable for the fault detection of complex working conditions.

To improve the FD-kNN, *He* proposed a PC-kNN method in [21] which adopted Principal Component Analysis (PCA) for feature extraction and measured the nearest neighbor distance by principal component features. Therefore, PC-kNN reduces the feature dimension and improves detection efficiency. In [22], *Zhou* proposed a k-nearest neighbor fault detection method based on distance-preserving projection to achieve distance-preserving in space. In [23], a weighted k-nearest neighbor fault detection method based on a dynamic feature matrix is proposed, which considers the Spatio-temporal correlation of features and the distance weights of different features. Forthe multi-condition process, literature [24] proposed a fault detection based on the sparse residual distance statistical index to obtain the sparse residual space through a similar operating state distance. Then, the sparse residual k-nearest neighbor distance is calculated to construct the fault detection statistic. In [25], the neighbor mean is used to calculate the sample estimated score and obtain the residual through the estimated score. This method can eliminate the influence of data structure on process fault detection. In [26], a fault-symptom table is established through the variable contribution analysis in the calculation of the nearest neighbor distance, which realizes the identification of abnormal variables.

B. OUR CONTRIBUTIONS

The above-mentioned research on fault detection methods based on k-nearest neighbors has made certain improvements in feature extraction, distance measurement, fault root tracing, etc. However, the operating data in the actual engineering process has an obvious characteristic of heterogeneous distribution. The traditional nearest neighbor scale is mostly determined by the trial and error method. Since the distribution diversity of neighbors, thus the coverage degree of the normal state and the separation ability of the abnormal state based on the FD-kNN have specific differences. The performance of FD-kNN for anomaly identification is limited by the fixed neighbor scale, that is also one of the critical factors for the false alarm rate (FAR) and the missing alarm rate (MAR).

This paper develops a weighted k-nearest neighbor fault detection method based on the separation index and dynamic neighborhood scale (MI-DNS-WkNN) to solve the above problems. According to the fault detection principle of FD-kNN, a weighted k-nearest neighbor fault detection framework is established. A separation degree index based on multistep nearest neighbor distance is defined, and on this basis, a dynamic neighborhood scale correction method is proposed to deal with the heterogeneous distribution of operating data. A dynamic threshold-setting method is introduced to track the sudden change in the operation state.

The rest of this paper is organized as follows. Section II outlines the principle of FD-kNN. In Section III, we solve the fixed neighbor scale and threshold problem and provide a dynamic FD-kNN method based on the multistep index and dynamic neighborhood scale. In Section IV, by comparing with the simulation results of existing schemes, we prove the advantages of our scheme. We conclude the paper in Section V.

II. PRINCIPLE OF FD-kNN

The fault detection process of FD-kNN is independent and only determined by the $k$nearest neighbor distance of online samples among offline samples, which is suitable for the process fault detection problem of complex working conditions. FD-kNN separates normal operation data from fault data based on the following theories [15]:

1) The normal state data is similar to the historical normal state data;

2) There is a certain deviation between the fault state data and the historical normal state data.

The FD-kNN fault detection method includes two stages: offline modeling and online detection.

A. OFFLINE MODELING

Step 1: For all samples in the training set, find their k-nearest neighbor samples.

Step 2: The kNN distance is calculated for each sample, and the kNN distance D_i^2 for sample i is defined as the sum of the distances from sample i to its k nearest neighbors:

$$D_i^2 = \sum_{j=1}^{k} d_{ij}^2 \tag{1}$$

where d_{ij}^2 isthe Euclidean Distance from sample $\boldsymbol{x}_\mathbf{i}$ to its jth nearest neighbor.

Step 3: Determine the abnormal state discrimination threshold.

A common threshold setting method is to determine the threshold D_α^2 with given confidence α according to the coverage rate of D_i^2 in the training sample set.

B. ONLINE DETECTION

Step 1: For the online sample $\boldsymbol{x}_\mathbf{t}$, find its k nearest neighbor samples from the training data set.

Step 2: Calculate the kNN distance $D_\mathbf{t}^2$ of $\boldsymbol{x}_\mathbf{t}$.

Step 3: Compare the magnitude relationship between $D_\mathbf{t}^2$ and the threshold D_α^2. If $D_\mathbf{t}^2 \le D_\alpha^2$, it is a normal operation state, otherwise, it is a fault state.

The scale of nearest neighbor k is the most important parameter in FD-kNN. Basically, if the k value is larger, the influence of noise on fault detection is smaller, but the boundary between normal and faulty states will be less obvious. A common approach is to try several different k values on offline data and then choose the k value that corresponds to the best cross-validation result. However, in practice, not only for different data sets, but even for the offline state of different operating conditions, the influence of the k value on it is also different, so the optimal k value among samples is various as well. The offline and online samples independently determine their optimal k value, which can determine the local range more effectively, thereby improving the ability to separate the normal state and the abnormal state.

III. MI-DNS-WkNN

In this paper, the dynamic neighbor scale and the dynamic threshold are used to deal with the difference in data sparsity under different working conditions, to improve the separation ability of FD-kNN for fault states, thereby reducing the false alarm rate and the missing alarm rate. The proposed method defines the multistep neighbors to evaluate the separation ability of abnormal states. On this basis, an iterative partition method of neighbor clusters and a fusion strategy of dynamic thresholds are proposed. The details of the proposed method are described as follows.

A. FEATURE EXTRACTION AND WEIGHTING STRATEGY

Kernel Principal Component Analysis (KPCA) is applied to reduce computational complexity. KPCA performs feature extraction on the standardized offline samples to achieve feature dimensionality reduction. The principal component score is used as the feature weight. Therefore, the nearest neighbors calculation in the algorithm proposed in this paper is the weighted feature after dimensionality reduction. The weighted Euclidean Distance is calculated as follows [27]:

$$D_P^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} w_j d_{ij}^2 \tag{2}$$

where d_{ij}^2 is the Euclidean Distance from the sample $\boldsymbol{x}$ to its ith nearest neighbor of feature j, p is the feature dimension extracted by KPCA, and w_j is the jth eigenvalue of the kernel matrix in KPCA.

B. DYNAMIC NEIGHBORHOOD METHOD BASED ON MULTISTEP NEIGHBORS

This paper proposes a k-value division method for offline samples based on multistep neighbors. It is aimed to assign different k values to the samples in different sparse spaces, and to reduce the volatility of the neighbor distances of samples, thereby improving the separation ability for different states. The specific method of sample k-value division is as follows.

Firstly, the concept of multistep neighbors is defined, assuming that the number of offline samples is n and the number of neighbors is k. Define the sum from the first to the kth nearest neighbor of sample $\boldsymbol{x}_\mathbf{i}$ as the first-step neighbor, and the sum from $(k+1)$th to the $2k$th nearest neighbor as the second-step neighbor. Then the multistep nearest neighbor distances can be expressed as follows:

$$D_{i1}{}^2 = \sum_{j=1}^{k} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^2 \tag{3}$$

$$D_{i2}{}^2 = \sum_{j=k+1}^{2k} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^2 \tag{4}$$

On this basis, the separation index based on multistep neighbors is defined as follows:

$$SEP = \sum_{i=1}^{n} \frac{\sum\limits_{j=k+1}^{2k} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^2}{\sum\limits_{j=1}^{k} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^2} \tag{5}$$

The separation degree is advanced based on the similarity between the online state and nearest offline state. If the second-step neighbor with the highest similarity can be separated from the first-step neighbor, the abnormal state with stronger dissimilarity will be separated better.

According to the default k value, the weighted k nearest neighbor distances of all offline samples are calculated and arranged in descending order, and the mean and variance of the nearest neighbor distances are also calculated.

$$E_D = \frac{1}{n}\left[\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{k_i} d(i,j)^2\right] \tag{6}$$

$$\sigma_D = \frac{1}{n}\left[\sum\nolimits_{i=1}^{n} \left(D(i) - E_D\right)^2\right] \tag{7}$$

in which, the mean value E_D represents the average level of the neighbor distance of each sample, which is used to guide the updated direction of the k value in the iterative process, and the variance value σ_D represents the volatility of the offline sample neighbor distance, which is an important index used to evaluate the effectiveness of iteration.

The k values of offline samples are corrected based on E_D and SEP, the correction rules are shown in (8):

$$k_{new} = \begin{cases} \max\limits_{\forall k_{new}} SEP(i), \\ k_{new} \in \{k_{new} \,| k_{\min} \le k_{new} \le k_{\max}\} \\ \cap \{k_{new} \,\left| \left| D_{k_{new}}(i) - E_D \right| \right. < |D_k(i) - E_D|\} \\ \text{remain,} \;\; \text{other conditions.} \end{cases} \tag{8}$$

The above correction rules can be regarded as a two-step process. For sample $\boldsymbol{x}_\mathbf{i}$, the first step is to find all k values that meet the constraints $\left|D_{k_{new}}(i) - E_D\right| < |D_k(i) - E_D|$ within the preset range $[k_{\min}, k_{\max}]$. The purpose is to make the neighbor distance $D_{knew}(i)$ under the new k value tend to the mean E_D and reduce the volatility of the neighbor distance distribution. Then, the k value that satisfies the constraints of the first step and maximizes the $SEP(i)$ is selected as the new k value for $\boldsymbol{x}_\mathbf{i}$. If no k value meets the above conditions, the original k value remains unchanged.

The k values of all samples are updated and the mean values are recalculated to complete the iteration. The stopping criterion is the variance no longer decreases or reaches the maximum iterations. Since the convergence direction of the $k$value is affected by the initial value, an iterative calculation is performed for all k values within the preset $[k_{min}, k_{max}]$ range in the offline process. The optimal k value is the value corresponding to the largest separation degree.

For online samples, the selection of the k value is determined by the most similar samples in the offline samples, that is, it is equal to the k value corresponding to the nearest neighbor sample.

C. DYNAMIC THRESHOLD SETTING METHOD

The identification of each online sample is independent in the fault detection process of FD-kNN, that is, whether the online state is abnormal or not only depends on its neighborhood state. This theoretical basis makes it possible to set the dynamic threshold. This paper proposes a dynamic threshold-setting method linked to the scale of neighbors and the neighborhood state. The static component is determined by the overall offline sample distribution, and the dynamic component is determined by the local neighborhood state of the online sample. The specific setting method is as follows:

1) STATIC COMPONENT FOR THRESHOLD

For the preset confidence coefficient α, take it as the tolerance for noise data, that is, the static component for threshold T_1 can cover (1-α)% of the offline normal operating samples:

$$T_1 = D_{Rank}(\lfloor 0.05n \rfloor) \tag{9}$$

where D_{Rank} is the descending order of the neighbor distance of offline samples, n is the number of offline samples, $\lfloor\ \rfloor$ is the rounding function.

2) DYNAMIC COMPONENT FOR THRESHOLD

The dynamic component of the threshold is used to track the changes in complex operating conditions. This paper proposes a dynamic component-determining method for the threshold associated with the neighborhood scale. For the online sample $\boldsymbol{x}_i(i = 1, 2, \ldots, n)$, find its nearest neighbor samples $\boldsymbol{kx}_i^j(j = 1, 2, \ldots, k_i)$, calculate the mean value of the neighbor distances of $\boldsymbol{kx}_i$ in the offline process as the dynamic threshold component of $\boldsymbol{x}_i$, that is:

$$T_2(i) = (1/k_i) \sum_{j=1}^{k_i} D_{k_j}^2(j)] \tag{10}$$

where k_i *is* the neighborhood scale of the online sample $\boldsymbol{x}_i$, and k_j is the neighborhood scale of the jth nearest neighbor of $\boldsymbol{x}_i$.

3) THRESHOLD FUSION

The threshold static component of the overall distribution and the threshold dynamic component of the current state is fusion as the final threshold T_D of $\boldsymbol{x}_i$. Since the amplitudes of the static component and dynamic component are both based on the amplitude of the online state, and the influence on noise in the static component is considered, therefore, the static and dynamic components are superimposed as follows:

$$T_D(i) = [D_{Rank}(\lfloor 0.05n \rfloor) + (1/k_i) \sum_{j=1}^{k} D_{kj}^2(j)]/2 \tag{11}$$

In summary, the flowchart of the MI-DNS-WkNN fault detection method proposed in this paper is shown in Fig. 1.

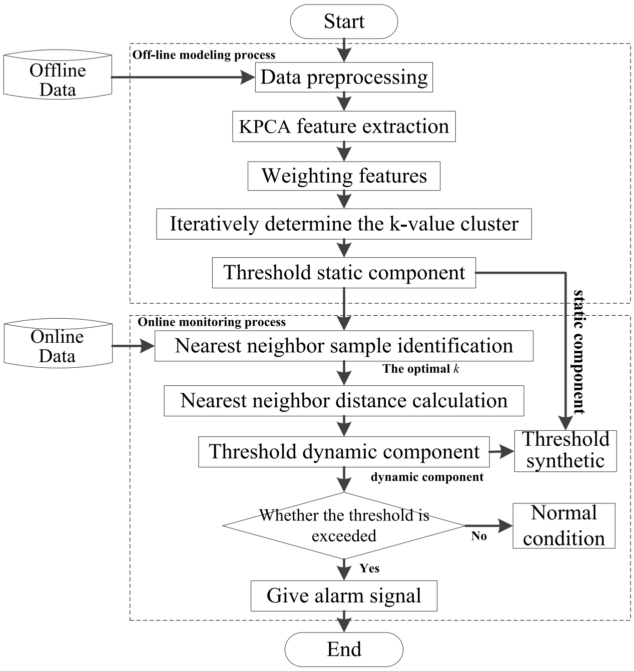

Start
Off-line modeling process
Offline Data
Data preprocessing
KPCA feature extraction
Weighting features
Iteratively determine the k-value cluster
Threshold static component
static component
Online monitoring process
Online Data
Nearest neighbor sample identification
The optimal k
Nearest neighbor distance calculation
Threshold dynamic component
Threshold synthetic
dynamic component
Whether the threshold is exceeded
No
Normal condition
Yes
Give alarm signal
End

FIGURE 1. Flow chart of MI-DNS-WkNN.

IV. SIMULATION EXPERIMENTS AND RESULTS ANALYSIS

A. EXPERIMENT DESCRIPTION

In this paper, 10 common fault data of megawatt-scale wind turbines are selected for simulation experiments, to verify the effectiveness of the proposed method for the fault detection process under complex operating conditions. The results are compared with FD-*k*NN [15], PC-*k*NN [22], FD-KPCA [28], and Kmeans-GMM [20].

Wind turbines are affected by the intermittency and randomness of wind. The operating states are frequently switched and affected by mechanical, electrical, and other factors, resulting in complex and changeable operation modes. According to the power characteristics of wind turbines, the operation process is divided into four stages: start-up and grid connection stage, maximum wind energy capture stage, constant power control stage, and over-wind speed cut-out stage.

In this paper, 10 common faults of NREL-5MW offshore wind turbines are selected. NREL 5 MW, a three-blade offshore WT with horizontal variable speed, is designed based on the Multibrid M5000 WT and Repower 5M WT [29], [30]. The parameters of NREL 5 MW offshore WT are shown in Table 1.

TABLE 1. Parameters of NERL 5MW offshore WT.

Parameters	Values
Rated power (P_n)	5 MW
Blade number	3
Tower height	87.6 m
Diameter of wind rotor	126 m
Cut-in wind speed, rated wind speed, cut-out wind speed	3, 11.4, and 25 m/s
Ratio of gearbox	98
Generator speed ($w_{g,n}$)	1173.7 rpm

Fifteen sensors are used to monitor the wind speed v_w, generator speed w_g, generator torque τ_g, rotor speed w_r, generating power p_e, pitch angle β_i of the blade i, root torque M_i of blade i, the azimuth angle of low-speed side ϕ, horizontal acceleration X_{acc} of tower top, vertical acceleration Y_{acc}, and deviation error Ξ. Fig. 2 shows the locations of the sensors.

The fault frequency of critical components in offshore WT is high due to its harsh working environment. The common fault types include sensor faults and actuator faults. In this paper, six sensor faults and four actuator faults are involved. Sensor faults contain small deviation (fault 6), proportion error (fault 1, fault 3, and fault 5), and numerical deviation (fault 2 and fault 4) between the measured value and the real value. Actuator faults include pitch system faults (fault 7 and fault 8), generator torque fault (fault 9), and yaw system fault (fault 10). The fault description and risk degree are shown in Table 2 [31].

TABLE 2. Fault description and risk degree.

No.	Fault location	Fault representation	Risk degree
1	Blade root sensor	Scaling	low
2	Accelerometer	Offset	medium
3	Generator speed sensor	Scaling	medium
4	Pitch angle sensor	Stuck	medium
5	Generator power sensor	Scaling	medium
6	The low-speed shaft position encoder	Bit error	low
7	Pitch actuator	Abrupt change in dynamics	medium
8	Pitch actuator	The slow change in dynamics	high
9	Torque offset	Offset	medium
10	Yaw drive	Stuck drive	high

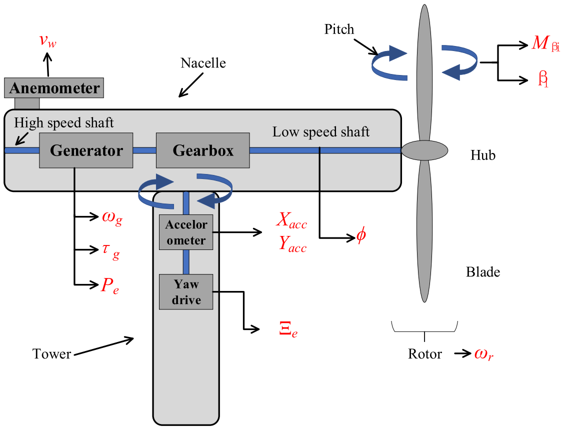


FIGURE 2. Locations of sensors.

In this experiment, 500 normal samples before the failure and 700 samples after the failure are collected for each fault as test data, and 10,000 groups of samples are randomly selected as training data from all the remaining normal operation data.

The effectiveness of the proposed method is verified by comparing the false alarm rate in the normal operating state and the missing alarm rate in the abnormal state. The false alarm rate (FAR) and the missing alarm rate (MAR) are calculated as follows [32]:

$$FAR = \frac{FP}{FP + TN} \tag{12}$$

$$MAR = \frac{FN}{TP + FN} \tag{13}$$

Among them, *FP* (false positives) and *TN* (true negatives) represent the number of identifying normal states as abnormal states and normal states, respectively. *FN* (false negatives) and *TP* (true positives) represent the number of identifying abnormal states as normal and abnormal states, respectively.

B. EXPERIMENTAL RESULTS AND ANALYSIS

1) OFFLINE STATE

The dimension of offline normal samples is 10000 (number of samples) × 15 (number of monitoring features). KPCA is adopted to reduce the computational complexity and assign weights to the features. The KPCA parameters are set as 85 % cumulative contribution; the Gaussian kernel function width is 20; the confidence coefficient is 99 %. The results of feature extraction and the weights of corresponding features are shown in Fig. 3.

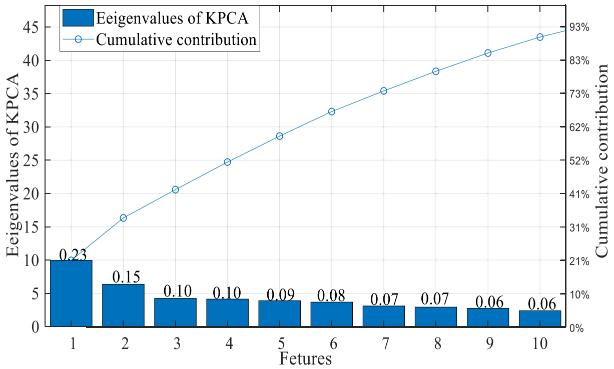


FIGURE 3. KPCA feature extraction results and corresponding weights.

Fig. 3 illustrates that the first 10 eigenvalues can reach the set cumulative contribution ratio. Therefore, these 10 eigenvalues are selected for subsequent offline calculation and normalized as the feature weights.

Each iteration is performed through the preset initial k value range $[k_{\min}, k_{\max}]$. The results show that when the initial value is 7 can achieve the maximum separation degree. The corresponding iteration results are shown in Fig. 4, in which Fig. 4 (a) is the distribution of the neighbor distance when the fixed k value is 7; Fig. 4 (b) is the distribution of the neighbor distance after the iterative correction of the dynamic neighborhood; Fig. 4 (c) is the changing trend of variance and separation degree with iterative correction; Fig. 4 (d) is the optimal neighborhood scale distribution corresponding to each sample after iteration.

Fig. 4 shows that the fluctuation degree of the neighbor distance distribution of the offline samples is significantly reduced by modifying the neighborhood scale, and a similar conclusion can be obtained from the changing trend of the variance during the iterative process. The separation degree of the samples also increases as the variance decreases, which verifies the effectiveness of the neighborhood scale correction method proposed in this paper.

There are still some cases with higher neighbor distances, which might be noise points or relatively sparse operating conditions. These cases will be processed through the subsequent dynamic threshold setting.

2) ONLINE PHASE

To verify the effectiveness of the method proposed in this paper, the experimental results are compared with a nonlinear process fault detection method FD-KPCA, the method Kmeans-GMM which considers the working conditions division, and the other two fault detection methods based on k- nearest neighbors FD-kNN and PC-kNN.

It should be noted that different from the bias verification between the predicted value and the actual value in the traditional data-based fault detection method, the bias in the kNN-based fault detection methods refers to the difference between the k-nearest neighbor distance of the online sample in the offline samples and the k-nearest neighbor distributions of offline samples.

The parameter settings of the comparison method are shown in Table 3. The hyper-parameters in the proposed algorithm include the initial value of the neighbor size; the range of k value $[k_{\min}, k_{\max}]$; the maximum iterations g, and the confidence level α.

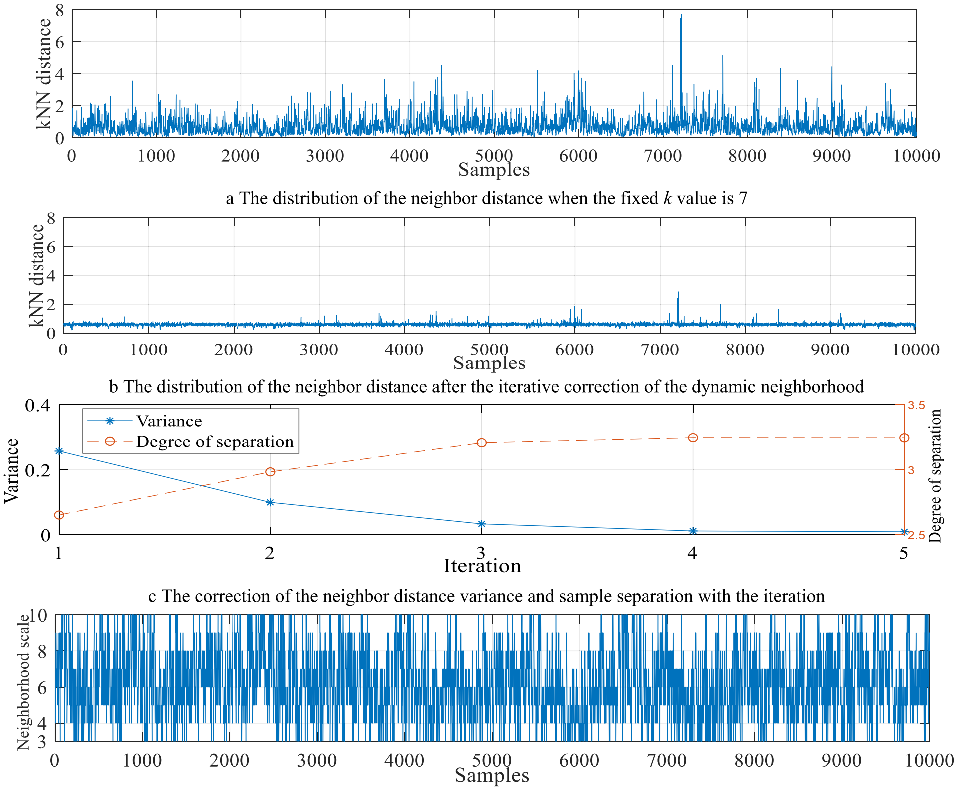
In the above parameters, the initial value of the neighbor size is the same as the fixed neighborhood size in other k-nearest neighbor-based methods, which is obtained from the trial and error method. If the range of k is too small, the approximate state is insufficient to describe the inherent characteristics of the operating state. If the range of k is too large, it may cover offline samples that are quite different from the current state, resulting in an increase in the discriminant index and the false alarm. Therefore, the range of k is usually taken as 3∼10 according to experience. The iteration number is used to control the number of corrections to the neighborhood size of the sample. The confidence level is used to control the noise tolerance, in this paper, the confidence level of all methods is unified to 5%. In addition, the clusters number of Kmeans-GMM is obtained by clustering validity function.

The false alarm rate of the first 500 samples and the missing alarm rate of the last 700 samples are counted, and the results are shown in Table 4 and Table 5, respectively.

The MAR is used to describe the separation ability of fault detection methods for different states (normal state and fault state). From the statistical results in Table 4, it is obvious that all methods can accurately detect fault 4, which is due to the obvious data bias caused by the stuck fault. For the other 9 faults, the proposed method achieved the lowest MAR compared with the other 4 fault detection methods. It is verified that the strategy of dynamic neighborhood scale based on the separation degree proposed in this paper can effectively separate the fault state from the normal state.

The FAR is used to describe the tolerance for sudden changes or noise in normal operating conditions. From the statistics of the FAR shown in Table 5, it can be seen that for most of the faults, the method proposed in this paper can obtain the lowest FAR compared with the other 4 methods. That is because the threshold setting of traditional methods usually set the fixed threshold according to the distribution characteristics of offline samples, while the dynamic threshold setting method proposed in this paper is suitable for FD-kNN, that is, it only depends on the local similar operating states. When the nearest neighbor distance of similar working states is large, the threshold of online samples is also increased accordingly, which reduces the false alarm caused by the sudden change of the operation state.

Take the detailed detection process of fault 5 (generator power sensor fault) and fault 8 (pitch system actuator fault) as an example to compare the performance of different fault detection methods more clearly, as shown in Fig. 5 and Fig. 6.



a The distribution of the neighbor distance when the fixed *k* value is 7

b The distribution of the neighbor distance after the iterative correction of the dynamic neighborhood

c The correction of the neighbor distance variance and sample separation with the iteration

d the optimal neighborhood scale distribution corresponding to each sample after iteration

FIGURE 4. Results of offline sample neighborhood scale after correction.

TABLE 3. Parameter settings of the comparison methods.

methods	parameters
FD-KPCA [28]	Cumulative contribution: 85%; statistical type: SPE/ T^2; confidence α: 5%; Kernel function: Gaussian.
Kmeans-GMM [20]	The number of clusters: 6; confidence α: 5%; Kernel function: Gaussian.
FD-kNN [15]	The number of nearest neighbors k: 7; confidence α: 5%.
FD-PC-kNN [22]	The number of nearest neighbors k: 7; the confidence α: 5%; the cumulative contribution rate is 85%.
MI-DNS-WkNN	The initial number of nearest neighbors k: 7; the minimum k value $k_{\min}$: 3; the maximum k value $k_{\max}$: 10; the maximum number of iterations g: 20; confidence α: 5%.

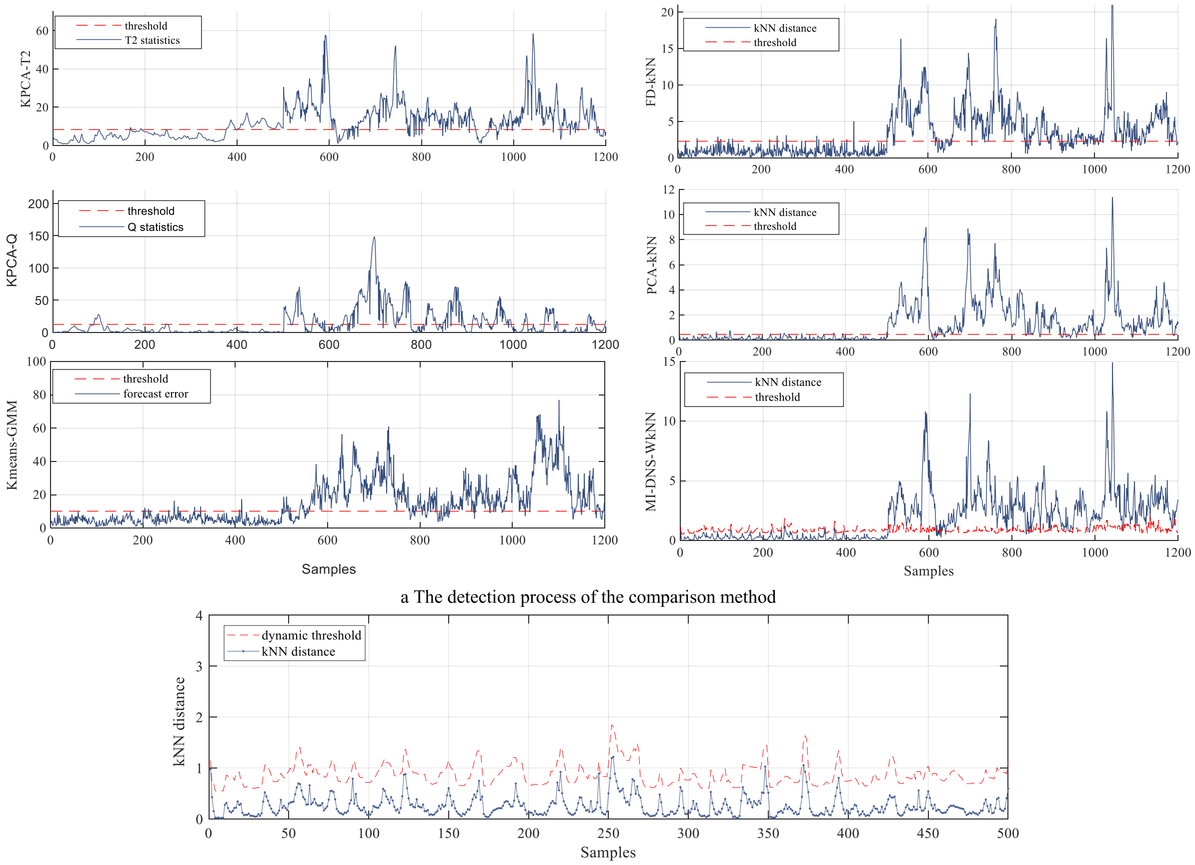
TABLE 4. Comparison of the missing alarm rate.

No.	FD-KPCA		FD-Kmeans -GMM	FD - *k*NN	FD-PC-*k*NN	MI-DNS-W*k*NN
	Q	T^2				
Fault 1	44.31	4.29	2.73	3.37	1.52	**1.46**
Fault 2	39.23	10.74	10.42	8.62	8.85	**3.37**
Fault 3	29.00	5.25	5.54	3.83	1.52	**0.51**
Fault 4	19.79	0	0	0	1.1	0
Fault 5	45.36	15.25	11.64	14.02	8.87	**2.24**
Fault 6	67.64	16.05	3.36	0	0	0
Fault 7	61.03	25.93	15.63	11.62	9.27	**3.17**
Fault 8	47.49	9.93	8.41	5.60	6.17	**2.10**
Fault 9	61.17	16.97	5.52	3.62	1.77	0
Fault 10	44.84	4.25	7.73	5.86	3.76	**0.21**
Average	45.98	10.87	7.10	5.65	4.28	**1.31**

From Fig.5 and Fig.6, it is obvious that the performance for fault detection of FD-KPCA is unsatisfactory, that is because the changes in complex working conditions of wind turbines are not considered. The fault detection effect of

TABLE 5. Comparison of the false alarm rate.

No.	FD-KPCA		FD-Kmeans -GMM	FD - *k*NN	FD-PC-*k*NN	MI-DNS-W*k*NN
	Q	T^2				
Fault 1	3.39	3.51	3.35	3.73	**1.58**	2.04
Fault 2	7.28	3.09	3.30	5.65	3.27	**1.26**
Fault 3	**2.56**	9.92	7.62	8.87	5.51	2.70
Fault 4	16.65	12.84	1.46	0	3.09	**0**
Fault 5	15.92	25.39	3.52	2.56	3.99	**2.27**
Fault 6	15.35	**0**	8.42	7.07	2.76	2.48
Fault 7	41.80	32.60	9.52	6.52	4.32	**0.79**
Fault 8	20.69	6.38	7.17	6.26	7.81	**1.58**
Fault 9	13.38	46.24	7.94	4.03	3.51	**1.04**
Fault 10	15.49	3.84	10.36	**3.25**	4.78	3.59
Average	15.25	14.38	6.26	4.79	4.06	**1.78**



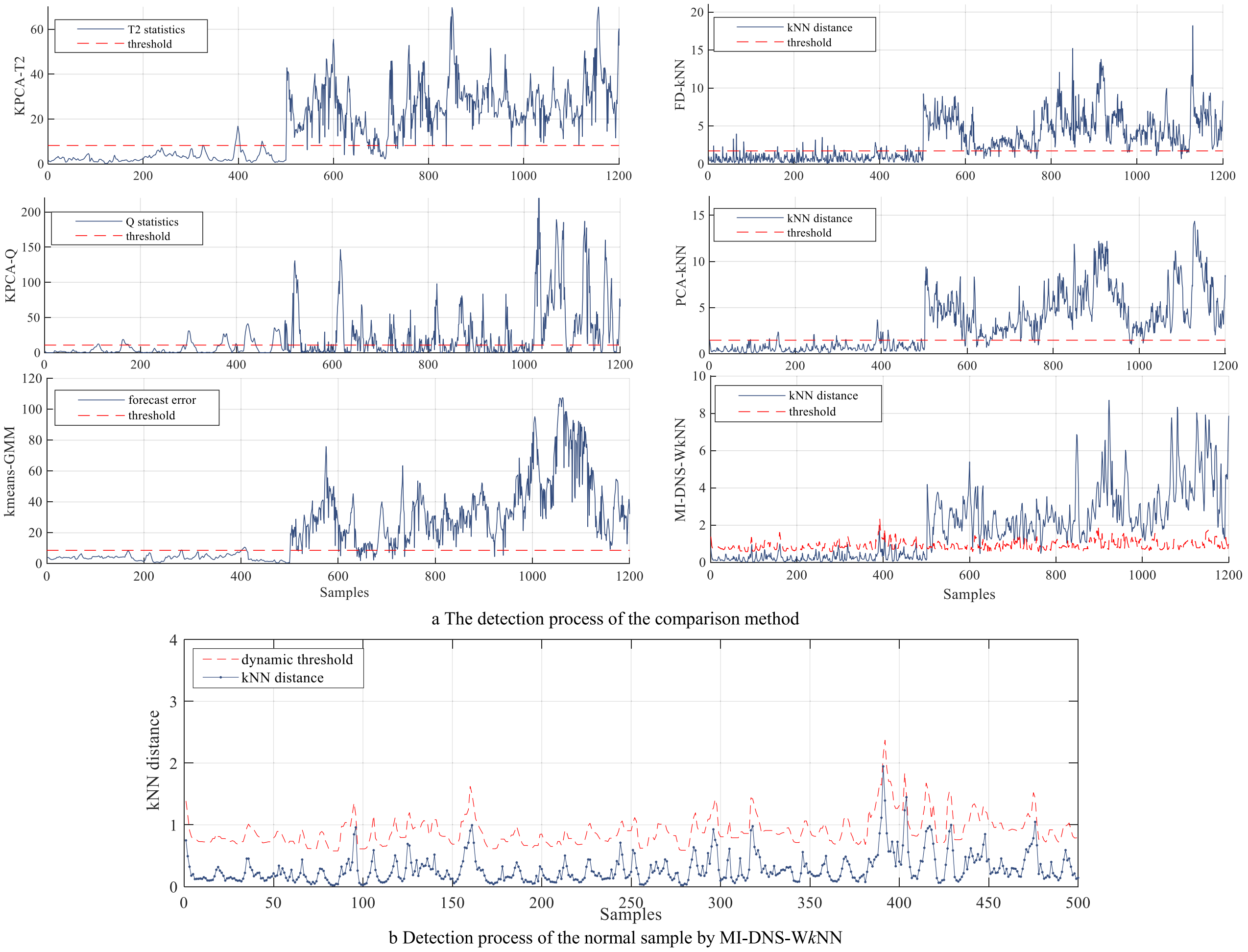
a The detection process of the comparison method

b Detection process of the normal sample by MI-DNS-W*k*NN

FIGURE 5. Fault 5 (generator power sensor fault) detection process.

the method based on the division of working conditions is better than FD-KPCA but worse than the methods based on *k*NN, this is due to the stronger local modeling capability of *k*NN-based methods and thus is more suitable for the fault detection process of wind turbines with complex and changing working conditions. Furthermore, compared with the other two similar *k*NN-based fault detection methods, the MI-DNS-W*k*NN proposed in this paper can separate the different states with a threshold of lower amplitude, which verified that the dynamic neighborhood scale modified by the separation index can effectively improve the separation ability of FD-*k*NN for normal state and abnormal state.

In addition, it can be seen from the enlarged part of the normal operating state that the dynamic threshold proposed in this paper can effectively follow the change of the wind turbine operating state, and effectively reduce the



a The detection process of the comparison method

b Detection process of the normal sample by MI-DNS-W*k*NN

FIGURE 6. Fault 8 (pitch system actuator fault) detection process.

false alarm caused by the abrupt change of the working condition.

V. CONCLUSION

In this paper, a weighted *k*-nearest neighbor fault detection method based on multistep index and dynamic neighborhood scale is proposed. The main innovations include the definition of the multistep neighbor index, the dynamic neighborhood strategy, and the setting of the dynamic threshold. The common faults of wind turbines with complex operating conditions are used to verify the effectiveness of the proposed method, and the experimental results show that:

(1) The proposed correction method of dynamic neighborhood scale can effectively reduce the fluctuation degree of the neighbor distance;

(2) The improved FD-*k*NN with dynamic neighborhood scale mechanism can effectively separate the abnormal state from the normal state, thereby reducing the MAR;

(3) The proposed dynamic threshold strategy can effectively track the change of the operating state, and reduce the FAR caused by the abrupt change of operating state.

The further research parts include fusion mechanism models and interpretive analysis of detection processes.

REFERENCES

[1] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, J. Keane, and G. Nenadic, "Machine learning methods for wind turbine condition monitoring: A review," *Renew. Energy*, vol. 133, pp. 620–635, Apr. 2019.

[2] Y. Zhang, X. Su, K. Meng, and Z. Y. Dong, "Robust fault detection approach for wind farms considering missing data tolerance and recovery," *IET Renew. Power Gener.*, vol. 14, no. 19, pp. 4150–4158, Dec. 2020.

[3] M. Schmid, E. Gebauer, C. Hanzl, and C. Endisch, "Active model-based fault diagnosis in reconfigurable battery systems," *IEEE Trans. Power Electron.*, vol. 36, no. 3, pp. 2584–2597, Jul. 2020.

[4] J. Feng, Y. Yao, S. Lu, and Y. Liu, "Domain knowledge-based deep-broad learning framework for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3454–3464, Apr. 2021.

[5] K. Zhong, M. Han, and B. Han, "Data-driven based fault prognosis for industrial systems: A concise overview," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 2, pp. 330–345, Mar. 2020.

[6] Y. G. Hu, "Overview of fault diagnosis and life prediction of wind turbine yaw system," *Chin. Soc. Electr. Eng.*, vol. 42, no. 13, pp. 4871–4884, Nov. 2022.

[7] C.-Z. Hu, Q. Yang, M.-Y. Huang, and W.-J. Yan, "Sparse component analysis-based under-determined blind source separation for bearing fault feature extraction in wind turbine gearbox," *IET Renew. Power Gener.*, vol. 11, no. 3, pp. 330–337, Feb. 2017.

[8] M. Ghane, A. R. Nejad, M. Blanke, Z. Gao, and T. Moan, "Diagnostic monitoring of drivetrain in a 5 MW spar-type floating wind turbine using Hilbert spectral analysis," *Energy Proc.*, vol. 137, pp. 123–204, Oct. 2017.

[9] H. N. Pan, M. Qin, J. Zhang, C. Chang, and P. Lei, "Bayesian networks in electric reliability assessment of doubly-fed wind turbine generator," *Appl. Mech. Mater.*, vols. 494–495, pp. 1791–1794, Feb. 2014.

[10] Y. Wang, X. Ma, and P. Qian, "Wind turbine fault detection and identification through PCA-based optimal variable selection," *IEEE Trans. Sustain. Energy*, vol. 9, no. 4, pp. 1627–1635, Oct. 2018.

[11] X. Zhao, M. Jia, J. Bin, T. Wang, and Z. Liu, "Multiple-order graphical deep extreme learning machine for unsupervised fault diagnosis of rolling bearing," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[12] J. Long, R. Zhang, Z. Yang, Y. Huang, Y. Liu, and C. Li, "Self-adaptation graph attention network via meta-learning for machinery fault diagnosis with few labeled data," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[13] J. Long, Y. Chen, Z. Yang, Y. Huang, and C. Li, "A novel self-training semi-supervised deep learning approach for machinery fault diagnosis," *Int. J. Prod. Res.*, vol. 70, pp. 1–14, Feb. 2022.

[14] X. Peng, Y. Tang, W. Du, and F. Qian, "Multimode process monitoring and fault detection: A sparse modeling and dictionary learning method," *IEEE Trans. Ind. Electron.*, vol. 64, no. 6, pp. 4866–4875, Jun. 2017.

[15] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 20, no. 4, pp. 345–354, Nov. 2007.

[16] Q. Jiang and X. Yan, "Multimode process monitoring using variational Bayesian inference and canonical correlation analysis," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1814–1824, Oct. 2019.

[17] Y. Zhang, K. Wang, X. Qian, and M. Gendeel, "Robust fault-detection based on residual K–L divergence for wind turbines," *IET Renew. Power Gener.*, vol. 13, no. 13, pp. 2400–2408, Oct. 2019.

[18] X. X. Zheng, M. N. Li, and J. Wang, "Classification of offshore wind turbine operating conditions based on PSO optimization kernel principal element analysis," *Power Syst. Protection Control*, vol. 44, no. 16, pp. 28–35, May 2016.

[19] Y. Q. Liu, "Classification method of wind turbine operating conditions based on support vector machine," *Chin. J. Sol. Energy*, vol. 31, no. 9, pp. 1191–1197, Nov. 2010.

[20] Y. L. Dong, Y. Q. Li, and H. B. Cao, "Real-time evaluation method of wind turbine health status based on operating condition identification," *Chin. J. Electr. Eng.*, vol. 33, no. 11, pp. 88–95, May 2013.

[21] Q. P. He and J. Wang, "Large-scale semiconductor process fault detection using a fast pattern recognition-based method," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 2, pp. 194–200, May 2010.

[22] Z. Zhou, C. Wen, and C. Yang, "Fault detection using random projections and k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 1, pp. 70–79, Feb. 2015.

[23] X. Y. Qian and Y. X. Zhang, "Fault detection method for k-nearest neighbor wind turbines based on dynamic feature matrix," *J. Instrum.*, vol. 40, no. 6, pp. 202–212, Sep. 2019.

[24] X. Guo, S. Liu, and Y. Li, "Research on fault detection method of multi-condition process based on sparse residual distance," *J. Autom.*, vol. 45, no. 3, pp. 617–625, Oct. 2019.

[25] Y. Li and Z. Yao, "Research on fault detection based on improved neighbor normalized principal polynomial," *Comput. Simul.*, vol. 39, no. 2, pp. 501–506+517. Nov. 2022.

[26] P. Wang, Y. Hu, and Y. Li, "Industrial process fault diagnosis based on k-nearest neighbor variable contribution and reconstruction theory," *Control Theory Appl.*, vol. 37, no. 3, pp. 639–650, Feb. 2020.

[27] F. Yuan, X. Xia, J. Shi, H. Li, and G. Li, "Non-linear dimensionality reduction and Gaussian process based classification method for smoke detection," *IEEE Access*, vol. 5, pp. 6833–6841, 2017.

[28] H. Sun, Y. Guo, and W. Zhao, "Fault detection for aircraft turbofan engine using a modified moving window KPCA," *IEEE Access*, vol. 8, pp. 166541–166552, 2020.

[29] R. Barcena, T. Acosta, A. Etxebarria, and I. Kortabarria, "Wind turbine structural load reduction by linear single model predictive control," *IEEE Access*, vol. 8, pp. 98395–98409, 2020.

[30] P. F. Odgaard and K. E. Johnson, "Wind turbine fault detection and fault tolerant control-an enhanced benchmark challenge," presented at the Amer. Control Conf., Washington, DC, USA, 2013.

[31] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and XGboost," *IEEE Access*, vol. 6, pp. 21020–21031, 2018.

[32] Z. Shi and W. O'Brien, "Development and implementation of automated fault detection and diagnostics for building systems: A review," *Automat. Construct.*, vol. 104, pp. 215–229, Aug. 2019.

XIAOYI QIAN received the M.Sc. and Ph.D. degrees in control theory and control engineering from the Shenyang University of Technology, Shenyang, China, in 2016 and 2020, respectively. He is currently with the College of Electric Power, Shenyang Institute of Engineering. His current research interests include data mining, intelligent optimization, and data-driven fault diagnosis.

TIANHE SUN received the M.Sc. and Ph.D. degrees in electrical engineering from the Shenyang University of Technology, Shenyang, China, in 2016 and 2020, respectively. From 2018 to 2019, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Aalborg University, Aalborg, Denmark. He is currently an Associate Professor with the College of Electric Power, Shenyang Institute of Engineering. His current research interests include fault diagnosis, intelligent optimization, and optimizing dispatching of IES.

BAOSHI WANG received the B.S. degree from the Shenyang Institute of Engineering, Shenyang, China, in 2008, and the M.Sc. degree in electrical engineering from the Shenyang University of Technology, Shenyang, in 2012. He is currently an Associate Professor with the College of Electric Power, Shenyang Institute of Engineering. His current research interests include intelligent optimization, intelligent optimization, and electric power system analysis.

YUXIAN ZHANG received the M.Sc. and Ph.D. degrees in control theory and control engineering from Northeastern University, China. He was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, from 2007 to 2009. He is currently a Professor with the Shenyang University of Technology. His current research interests include data mining, intelligent control, and intelligent optimization.