ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20210102034143
Preliminary communication

A Kernel Entropy Method and its Application in Monitoring and Assessment of Wind Turbine Degradation Performance

Yong-Sheng QI*, Chao REN, Xue-Jin GAO, Li-Qiang LIU, Chao-Yi DONG

Abstract: To overcome the problems of wind turbine (WT) degradation assessment, a new kernel entropy method based on supervisory control and data acquisition (SCADA) was proposed. This approach can be used to effectively monitor and assess WT performance degradation. First, a new condition monitoring method based on a kernel entropy component analysis (*KECA*) was developed for nonlinear data. Then, the squared prediction error (*SPE*) was used to monitor the WT health state. Due to the diversity and nonlinearity of SCADA data, fault features are easily overwhelmed by other vibration signals. To address this, a new kernel entropy partial least squares (*KEPLS*) algorithm was introduced. The proposed kernel entropy method improves the performance prediction by considering higher order information. Furthermore, changes in the prediction residual can be used to define certain limits to realize early warning of WT faults. Finally, the method was applied to actual SCADA data of a wind farm. The results show that the method can accurately evaluate the health state of WTs, thus verifying the effectiveness and feasibility of the proposed method.

Keywords: degradation performance monitoring; health assessment; Kernel Entropy Component Analysis (*KECA*); Kernel Entropy Partial Least Squares (*KEPLS*); SCADA data

1 INTRODUCTION

With the intensification of the energy crisis, wind energy has become an important source of renewable energy and has gradually played a significant role in the global energy mix. In general, wind turbines (WTs) operate in harsh environments under dramatically fluctuating operating conditions and are often subjected to strong mechanical stress. Their operating state and load bearing conditions are random and unstable. Therefore, methods to improve availability of WTs, reduce operation and maintenance costs, and enhance the economic benefits of wind farms are of great significance. To date extensive research has been carried out on condition monitoring and fault diagnosis of key components of the WT [1-6]. Condition monitoring systems (CMSs) are capable of processing high frequency signals. However, due to data acquisition problems, incomplete data, and the inability to analyze all fault information, many scholars have recently adopted prognostics and health management (PHM) systems [7-13]. Failure PHM of WTs, including the development of assessment systems for monitoring and managing the health status of WTs across their entire life cycle, have become an important research focus. Various algorithms and intelligent models are used as part of PHM, which can monitor, predict, and manage the health of WT systems in order to achieve condition-based or predictive maintenance. Performance degradation prognostics and health assessment of WTs form the basis of PHM research.

In recent years, data-driven multivariate statistical monitoring methods have been widely used for condition monitoring of industrial processes [14]. The core idea is to transform the input space into a feature space and a residual space through dimensionality reduction. A set of low-dimensional variables containing important features that summarize the information carried by high-dimensional data can be constructed using multivariate statistical monitoring methods, such as principal component analysis (*PCA*), partial least squares (PLS) approach, independent component analysis (ICA), and other related algorithms [15]. The data are then projected onto a higher-dimensional squared prediction error (*SPE*) and Hotelling's T^2 statistical relationship is calculated to determine whether the component has exceeded a predefined control limit and whether the situation is abnormal.

Principal component analysis is the most widely used algorithm in multivariate statistical monitoring and can effectively reduce the dimensionality of data while retaining the maximum variance of the original data. However, the *PCA* algorithm targets linear systems, whereas WT monitoring data, which mainly come from supervisory control and data acquisition (SCADA) systems, are multivariate and nonlinear, and exhibit strong coupling among variables. The application of *PCA* to SCADA data is not ideal. Accordingly, Scholkopf et al. [16] introduced kernel *PCA* (*KPCA*), which combines the kernel function with the *PCA* algorithm. The kernel function can be used to extend *PCA* to a high-dimensional feature space by eliminating the nonlinearity of process variables to achieve more effective process monitoring. In 2010, Robert Jenssen [17] proposed kernel entropy component analysis (*KECA*), which combines the kernel function with the concept of information entropy. Through kernel mapping, data are mapped to a high-dimensional space, which solves the nonlinear data problem, and the dimensionality of the data in the high-dimensional feature space is reduced based on the size of the kernel entropy. Therefore, more feature information can be retained by building deep kernels. In contrast to *PCA* and *KPCA*, which use the eigenvalue size as an index, *KECA* reduces the dimension by disclosing the structure of the dataset through information entropy to more effectively reveal the data.

Wind turbine condition monitoring based on *KECA* can only determine whether the state is abnormal and whether a fault has occurred within a certain period of time, but cannot predict fault trends in advance, identify potential faults, or evaluate the health status of the WT. Yi Liu, Qing-Yang Wu and Junghui Chen investigated an active selection of information data for sequential quality enhancement of soft sensor models with latent variables [18].Yi Liu, Chao Yang, Zengliang Gao and Yuan Yao proposed ensemble deep kernel learning with application to quality prediction in industrial polymerization processes

[19]. Kaixin Liu, Zhengyang Ma, Yi Liu, Jianguo Yang and Yuan Yao reported enhanced defect detection in carbon fiber reinforced polymer composites via generative kernel principal component thermography [20]. Hongying Deng, Keyun Yang, Yi Liu and Shengchang Zhang proposed actively exploring informative data for smart modeling of industrial multiphase flow processes [21]. To this end, considering the non-linear and non-stationary characteristics of SCADA data, a new fault prediction method based on kernel entropy PLS (*KEPLS*) is proposed. The algorithm uses Renyi entropy to realize feature vector arrangement and dimensionality reduction, which can better characterize the angle information between different nonlinear features. In addition, *KEPLS* can effectively extract high-order statistics and overcome the problem of non-stationary data to a certain extent. It can also solve the problem faced by traditional *KPLS*, which can only represent second-order statistics, and often ignores the problem of fault information implicit in high-order statistics.

Thus, to better monitor the condition of WT components and predict future faults, this paper proposes a *KEPLS* predictive model that not only extracts multi-scale information more effectively but also analyzes the residual error to achieve more accurate fault prediction. In addition, a fuzzy comprehensive evaluation method and radar chart method were used to perform a visual analysis of a WT gearbox health assessment in one dimension and multiple dimensions to better understand the experimental results from an objective and visual perspective.

The rest of this paper is organized as follows. Section 2 describes a generalized model framework for condition monitoring and fault prediction of WTs. Section 3 outlines the experimental data cleaning process. Section 4 summarizes the condition monitoring method based on the *KECA* and *SPE*. Section 5 describes the failure prediction framework using the *KEPLS* model. In section 6, the results of applying the method to specific case study are presented. Finally, the main conclusions of this work are summarized in section 7.

2 GENERALIZED MODEL FRAMEWORK

The condition monitoring and fault prediction of WTs were studied. A flowchart of the algorithm is presented in Fig. 1. The algorithm is divided into three stages, as follows.

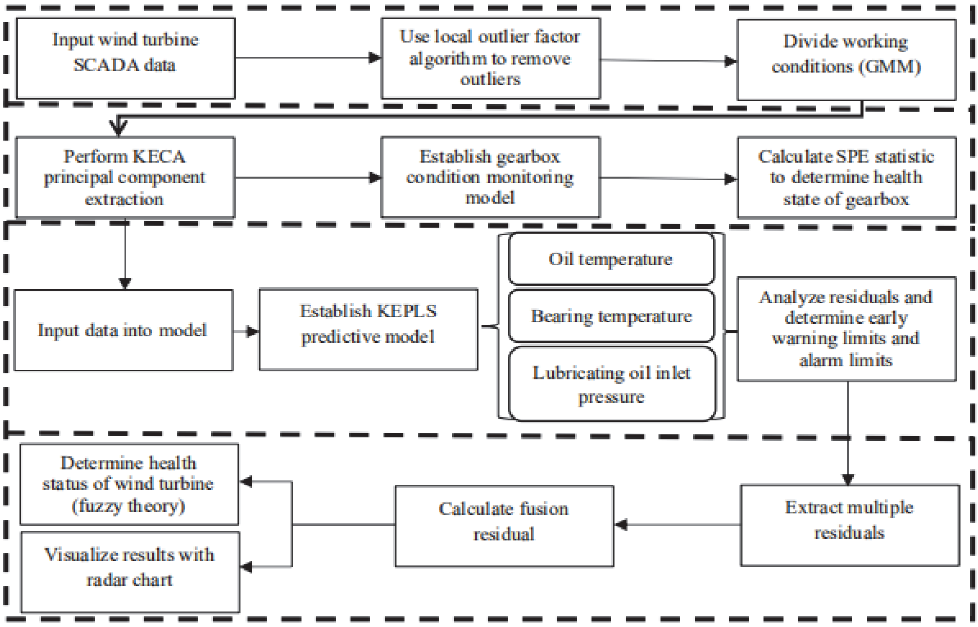

Input wind turbine SCADA data
Use local outlier factor algorithm to remove outliers
Divide working conditions (GMM)
Perform KECA principal component extraction
Establish gearbox condition monitoring model
Calculate SPE statistic to determine health state of gearbox
Input data into model
Establish KEPLS predictive model
Oil temperature
Bearing temperature
Lubricating oil inlet pressure
Analyze residuals and determine early warning limits and alarm limits
Determine health status of wind turbine (fuzzy theory)
Visualize results with radar chart
Calculate fusion residual
Extract multiple residuals

Figure 1 Flowchart of proposed algorithm

(1) Condition monitoring based on *KECA*:

There is a complex nonlinear relationship between parameters of the SCADA system. When a fault occurs, several parameters will change. However, changes in each parameter may be caused by multiple faults. To characterize the SCADA data, the *KECA* algorithm is combined with *SPE* statistics and applied to WT condition monitoring.

(2) Fault prediction based on *KEPLS*:

Principal component information extracted with the *KECA* algorithm was more detailed and has the advantage of containing deep hidden relationships. After extracting the *KECA* features, a *KEPLS* prediction model is established and applied to WT gearbox fault prediction. The model was compared with the *KPLS* predictive model to verify that the *KEPLS* algorithm can avoid the influence of nonlinear data on the analysis results and effectively improve the prediction accuracy of model.

(3) Health assessment based on fusion residuals:

According to the changes in residual characteristics generated by the *KEPLS* predictive model, potential gearbox faults were detected, and the health status of the gearbox was assessed. The single monitoring and analysis of a residual characteristic cannot comprehensively analyze the health status of the gearbox. Predictions of the gearbox oil temperature residual, bearing temperature residual, and lubricating oil inlet pressure residual were combined to enhance the reliability of the assessment.

3 DATA PREPROCESSING

Data preprocessing refers to operations such as cleaning, filtering, removing, and converting redundant data. Furthermore, data quality can be improved through appropriate parameter selection [13], structured model representation of unstructured data, and fusion of multiple multi-scale data, and further development of methods for acquiring fault characteristic data [22].

3.1 Outlier Elimination

Outliers in the data include both outliers and duplicates, which are usually caused by measurement errors, missing records, or abnormalities. Duplicate points can be easily found and filtered out by observing the wind speed and power time series. Since there are no labels in the data set, an unsupervised detection method was adopted to define outliers from the perspectives of distance, probability, and density of the feature space. The local outlier factor (LOF) method based on the density local outlier idea was used to identify outliers.

The LOF algorithm can be described as follows. The LOF algorithm judges whether the point is abnormal by comparing the density of each point p and the neighbouring point k; the lower the density of the point p, the more likely that it is an abnormal point. The density of points is calculated by the distance between points. The farther the distance between the points, the lower the density. Conversely, the closer the distance, the higher the density. The local outlier factor algorithm produces the anomaly score of a sample by calculating the "local reachable density". The greater the local density of a sample point, the more likely that this point is an anomaly.

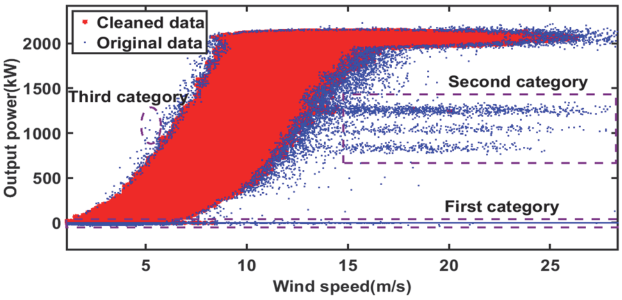


Figure 2 Comparison of data before and after preprocessing

To further analyze valid data between the cut-in and cut-out wind speed, as shown in Fig. 2, three types of abnormal data under normal operating conditions of the WT and their typical characteristics are defined:

(1) Accumulation point (first category, second category): A horizontal line of accumulation points at the bottom or middle part of the curve is often caused by wind abandonment, power rationing, communication failures, etc. First category accumulation points correspond to the initial points identified by expert screening. In this case, the output power is very small or continuously less than or equal to zero for a period of time. Second category accumulation points are abnormal points that occur when the output power is lower than the normal output and does not change (or rarely changes) with wind speed for a continuous period of time. These data cannot directly indicate abnormal functioning or malfunctioning of the WT and will affect the accuracy of the WT health status prediction. The key characteristic parameters such as wind speed, output power and pitch angle of the accumulation point have typical time series characteristics, and the first type of accumulation point and the second type of accumulation point are identified in turn based on engineering experience. First, the original data set is screened based on expert experience and physical laws. Second, the abnormal points caused by wind curtailment and power rationing are screened out based on the pitch angle information. Therefore, the accumulation points of the first and second types are eliminated.

(2) Outliers (third category): Outlier data with scattered characteristics are often caused by random factors such as sensor abnormalities, noise, and fluctuations in operating conditions. The fluctuation of outliers is random but objectively reflects the actual operating conditions of the WT to a certain extent. However, accuracy of the probabilistic density-based health state prediction model of the WT is affected by high outlier ratios and large dispersion.

Results of the wind power data preprocessing are shown in Fig. 2. Accumulation points and outliers can be observed in the raw data. For example, when the output power is 1000 kW, a small number of data points fall within a wind speed of 15 - 25 m/s, which are considered abnormal data. After the LOF method is applied to filter out abnormal data, the abnormal data detection method is effective.

3.2 GMM Classification

The Gaussian mixture modeling (GMM) method was used to adaptively divide the working conditions and capture the quasi-linear wind power data.

The Gaussian mixture model clusters data based on its statistical distribution. If "natural sub-categories exist" in the sample data, then observations in a certain sub-category are considered to come from a certain statistical distribution, and the whole observation comes from multiple statistically distributed random samples with a finite mixture distribution. Data with a non-Gaussian distribution is usually decomposed into a linear combination of several Gaussian distribution functions:

$$H\left(x\right)=\sum_{i=1}^{n}p_{i}N\left(\mu_{i},\,\delta_{i}\right) \tag{1}$$

where p_i is the prior probability of the i-th component, which satisfies $p_i \geq 0, \sum_{i=1}^{n} p_i = 1, N\left(\mu_i, \delta_i\right)$ is the gaussian distribution function of the i-th component; and parameters μ_i and δ_i are the mean and variance of the density function, respectively. Using the expectation-maximization (EM) clustering algorithm, when sample data X is known and class z is unknown, the logarithmic likelihood function is maximized through iteration to determine the estimated value of component parameters $\hat{\mu}$ and $\hat{\delta}$, as follows:

$$LL(\mu,\delta|X,z)=\log\prod_{i=1}^{n}N(x_i,z_i|\mu,\delta)= \\ =\sum_{z}p\sum_{i}\log N(x_i,z_i|\mu,\delta) \tag{2}$$

A finite Gaussian mixture model for wind power data was established. According to mean value μ_i (centroid-like), variance δ_i and the proportion (mixing probability) p_i of wind speed variables of each component, the segmented quasi-linear wind speed and output power data were obtained.

From the raw data, the state of each WT part cannot be clearly observed. Therefore, raw data must be labeled using the clustering method.

SCADA data were clustered using the GMM method, and the clustering number was 3. The wind power data are the initial input data, and the data are expanded to all categories according to the wind data classification results. First, the Gaussian mixture model was fitted. The mixing ratio of the first type was 0.395318 and the average value was 5.375820, that of the second type was 0.389258, with an average value of 9.196038, and that of third type was 0.215425 and the mean value was 14.237322. Then, the Gaussian mixture model was clustered, and the labels of each vector were obtained and stored in a file for future invocation.

The classification results of the Gaussian mixture model were clearly divided into three categories, as shown in Fig. 3. It can be seen from Fig. 3 that the wind speed range of the second category is approximately 5 ~ 13.2 m/s ($\mu \pm 2\sigma$). Considering that the performance degradation of the wind turbine will lead to an increase in the rated wind speed, and the subspace of the working conditions will be further divided when extracting features, the wind power data in the range of 5 ~ 15 m/s are captured to form a new data set. Next, the model passed to the *KECA* model as training data for the next step of the *KECA*.

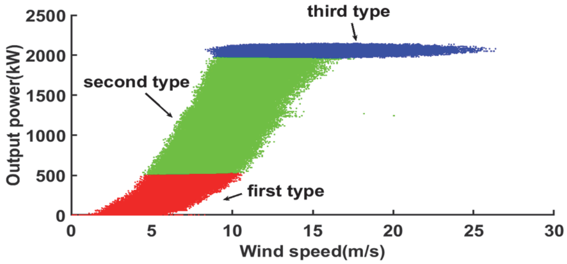


Figure 3 Visualization of classification result

4 WIND TURBINE CONDITION MONITORING BASED ON *KECA*

4.1 Kernel Entropy Component Analysis

According to the principle of *KECA*, after the principal component model is established, monitoring statistics between the principal component model and the data to be detected determine whether any abnormalities are present. Here, the *SPE* of residual spatial statistics was selected. The *SPE* statistic reflects the degree of deviation between the model and test value at any given moment. The formula for calculating the *SPE* statistic is:

$$SPE_i = E_i E_i^{\mathrm{T}} = t_i\left(I - P_R P_R^{\mathrm{T}}\right)t_i^{\mathrm{T}} \tag{3}$$

where t_i is the i-th core element of input vector x in the feature space and P_R is the feature vector extracted by *KECA*.

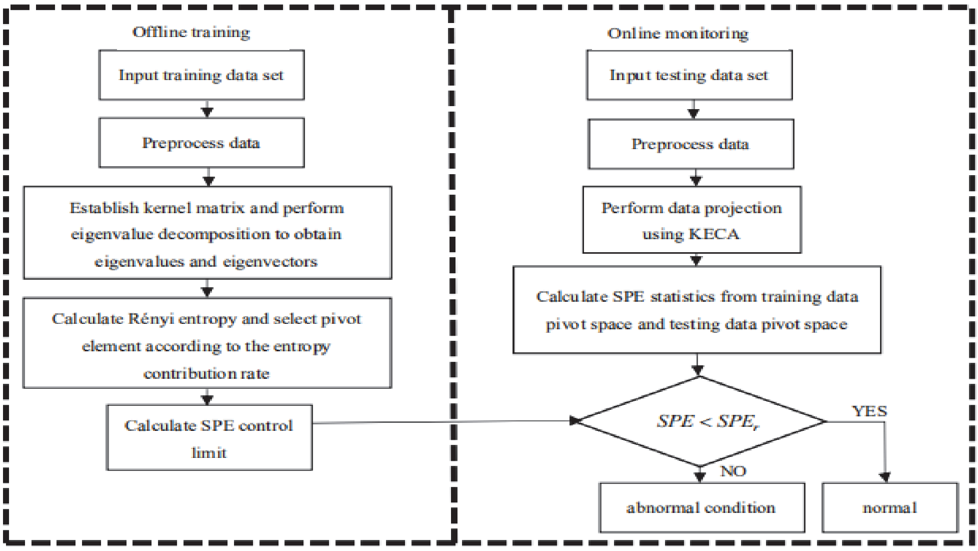

Offline training
Input training data set
Preprocess data
Establish kernel matrix and perform eigenvalue decomposition to obtain eigenvalues and eigenvectors
Calculate Rényi entropy and select pivot element according to the entropy contribution rate
Calculate SPE control limit
Online monitoring
Input testing data set
Preprocess data
Perform data projection using KECA
Calculate SPE statistics from training data pivot space and testing data pivot space
$SPE < SPE_r$
YES
NO
abnormal condition
normal

Figure 4 Kernel entropy component analysis (*KECA*) algorithm for wind turbine gearbox condition monitoring

When the confidence level is γ, the control limit of the *SPE* statistic can be calculated using the following formula:

$$SPE_{\gamma} = \theta_1\left[\frac{c_{\gamma}\sqrt{2\theta_2 h^2}}{\theta_1} + 1 + \frac{\theta_2 h(h-1)}{\theta_1^2}\right]^{1/h} \tag{4}$$

where $\theta_i = \sum_{j=p+1}^{n}\lambda_j^i;\ i = 1,\ 2,\ 3;\ h = 1 - \frac{2\theta_1\theta_3}{\theta_2^2};\ C\gamma$ is the critical value of the standard normal distribution test level γ. If $SPE < SPE\gamma$, the *SPE* statistic of the input vector is normal.

In this paper, the *KECA* algorithm was proposed to realize WT condition monitoring. The specific implementation can be divided into two parts: offline training and online monitoring. A flowchart of the *KECA* algorithm for WT condition monitoring is presented in Fig. 4. (1) Offline training:

Step 1: Select normal working condition data, then preprocess and standardize the data.

Step 2: Build a kernel matrix for the sample data. The kernel matrix is then decomposed to obtain the corresponding eigenvalues and eigenvectors, and the quadratic entropy of Rényi is calculated. The corresponding pivot element is selected according to the contribution rate of Rényi's quadratic entropy. The principal component with a cumulative contribution rate higher than 95% was selected in this work.

Step 3: Establish the *KECA* monitoring model for data under normal working conditions.

(2) Online monitoring:

Step 1: Preprocess and standardize the data online in the same way as the training data.

Step 2: Calculate the corresponding kernel matrix and pivot matrix.

Step 3: Calculate the *SPE* statistics and compare them with the control limit of 99% confidence to determine whether a fault will occur.

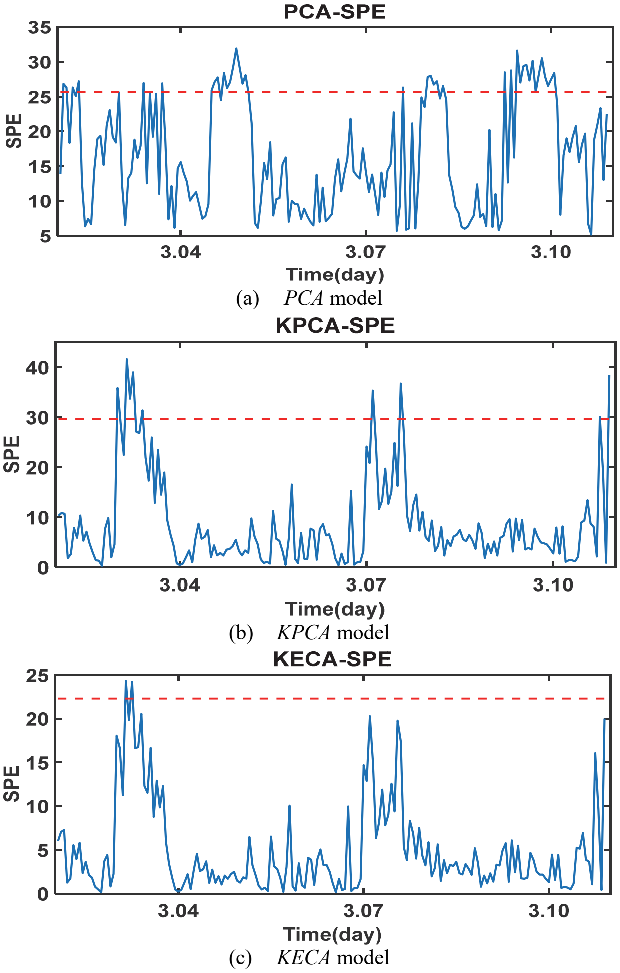
4.2 Analysis and Comparison of Condition Monitoring Results Based on *KECA*

Data from the SCADA system of a wind farm was adopted. Wind speed, output power, gearbox oil temperature, gearbox bearing temperature, and gearbox lubricating oil inlet pressure were selected as the experimental parameters. Parameter data with a sampling time of 1 h from February 21 to March 1 were used as the experimental training data set. The corresponding data from March 2 to March 11 were used to verify the model and data from March 12 to April 16 as the testing data set.

To verify the effectiveness of the *KECA* algorithm for WT condition monitoring, principal elements of the experimental training samples were first extracted to establish a monitoring model, then the model verification samples were used as the test data set in the monitoring phase and the *SPE* statistics were calculated. It can be seen from the actual situation that the gearbox is under normal operating conditions during this time period and the operating state is normal. *KECA-SPE* has fewer false alarm points. The dotted line in the figure represents the 99% control limit, and the solid line represents the *SPE* change curve under normal operation. If it is higher than the threshold limit, it is regarded as a fault, otherwise, it is regarded as normal. Therefore the *SPE* statistic calculated by the model should be below the threshold limit.

Fig. 5 shows the change in *SPE* of the WT gearbox under normal operating conditions obtained using *PCA*, *KPCA*, and *KECA*. The dotted line represents the 99% control limit. The main reason for dramatic changes in the *SPE* over time is that environmental conditions such as wind speed, wind direction, and ambient temperature can cause frequent changes in the working conditions. Several false alarm points can be observed in Fig. 5. Compared with the *KPCA* and *KECA*, *PCA* exhibited the worst prediction performance mainly due to nonlinearity of the WT gearbox data. The effect of the principal component after the nonlinear change is better. The *KECA* result is better than that of the *KPCA* because the *KECA* uses Rényi's second entropy to select the principal elements, which can reveal the structure of the data and extract deep information.

The test sample was used as the input of the monitoring model. The actual data show that the WT was shut down on April 16 due to the high temperature of the main shaft caused by a faulty gearbox. The WT condition was monitored using the *PCA*, *KPCA*, and *KECA*. The results are shown in Fig. 6. False alarm points can be observed on each of the three graphs. However, the *SPE* value of the false alarm point on the *PCA* graph was larger than that of the fault point. Due to abnormal data points generated between March 30 and April 10, the method cannot distinguish between a false alarm point or fault point. This phenomenon was rarely observed with *KPCA* and *KECA*, showing the limitations of *PCA* for nonlinear data processing. Although both *KPCA* and *KECA* can effectively monitor the fault state of the WT, the *KECA* model exhibited better monitoring effects.



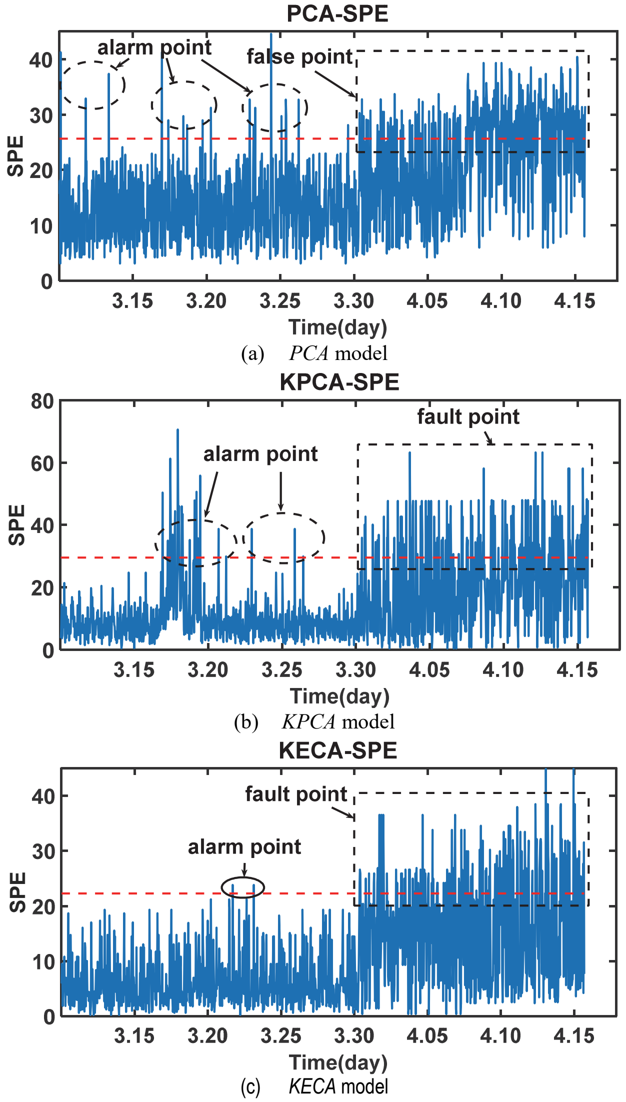
(a) *PCA* model

(b) *KPCA* model

(c) *KECA* model

Figure 5 Comparison of squared prediction error (*SPE*) results obtained using the *PCA*, *KPCA*, and *KECA* model for monitoring the condition of the wind turbine. The graphs show changes in the *SPE* over time under normal operating conditions

The *KECA* model also discovered the fault state faster than the other two methods. Although the *KECA* produces a small number of false alarm points, these were likely caused by the increase in the volume of data and insufficient data preprocessing or due to environmental factors. False alarm points cannot be avoided with the frequently changing WT operating conditions. However, in general, the *KECA* resulted in fewer false alarm points compared with the other methods.



(a) *PCA* model

(b) *KPCA* model

(c) *KECA* model

Figure 6 Comparison of squared prediction error (*SPE*) results obtained using the *PCA*, *KPCA*, and *KECA* models for monitoring the condition of the wind turbine. The graphs show changes in the *SPE* over time under normal operating conditions

5 PREDICTION MODEL BASED ON *KEPLS*
5.1 Principles of *KEPLS* Algorithm

Similar to the core idea of *KPCA*, the *KPLS* algorithm extracts feature vectors by maximizing the variance. However, in general, if the eigenvectors are only extracted according to the magnitude of the eigenvalues, the high-order information entropy of the raw sample variable will not be expressed well. Entropy is a concept representing the amount of information contained. Studies have shown that by introducing entropy information, the *KECA* algorithm can achieve better nonlinear processing results than the *KPCA* algorithm [23-25]. Therefore, drawing on the principle of *KECA*, the *KPLS* algorithm can also be used to extract the eigenvalues and eigenvectors according to the information entropy. Thus, the *KEPLS* algorithm is proposed for WT fault prediction.

The *KEPLS* algorithm projects the data from the low-dimensional input space to the high-dimensional feature space through kernel mapping, converts nonlinear data into linear data, and then selects features in the high-dimensional feature space according to the entropy, in order to achieve data dimensionality reduction.

Information entropy is a measure of uncertainty in a system. Here, Rényi entropy was used in the *KPLS* kernel entropy component analysis. Rényi entropy is defined as:

$$H(p)=-\log\int p^2(x)\mathrm{d}x \tag{5}$$

where $p(x)$ is the pdf of the data D.

Since Eq. (5) is a monotone function, it can be expressed as:

$$V(p)=\int p^2(x)\mathrm{d}x \tag{6}$$

The Parzen window density estimator can be used to estimate Eq. (6) and the following formula is obtained:

$$\hat{p}(x)=\frac{1}{N}\sum_{x_i\in D}W_\sigma(x,x_i) \tag{7}$$

Substituting Eq. (7) into Eq. (6), the following equation is obtained:

$$\begin{aligned}\hat{V}(p)&=\int p^2(x)\mathrm{d}x=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int W_\sigma(x,x_i)k_\sigma(x,x_j)\mathrm{d}x=\\&=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int W_{\sqrt{2}\sigma}(x,x_i)\end{aligned} \tag{8}$$

The volume integral of the Gaussian function is still a Gaussian function and the sign simplification of Eq. (8) can be obtained as $W_{\sqrt{2}\sigma}(x_i,x_j)$. Using the Gaussian function as the kernel function, it can be expressed as follows:

$$\hat{V}(p)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k(x_i,x_j)=\frac{1}{N^2}I^{\mathrm{T}}KI \tag{9}$$

where I is a vector of all ones and K is an $N\times N$ kernel matrix.

Eigen-decomposition, $K=\phi^{\mathrm{T}}\phi=EDE^{\mathrm{T}}$, is performed on kernel matrix K where $D=diag(\lambda_1,...,\lambda_N)$; E is the corresponding eigenvector matrix $E=(e_1,...,e_N)$. Hence, the following equation can be obtained [17, 26]:

$$\hat{V}(p)=\frac{1}{N^2}\left(\sqrt{\lambda_i}e_i^{\mathrm{T}}1\right)^2 \tag{10}$$

The generalized eigenvector in the PLS algorithm is $X^{\mathrm{T}}YY^{\mathrm{T}}X$ and weight vector ω is the eigenvector corresponding to the largest eigenvalue of the generalized eigen matrix.

$$X^{\mathrm{T}}YY^{\mathrm{T}}X\omega=\lambda\omega \tag{11}$$

Scoring vector t of X is calculated as

$$t = X\omega \tag{12}$$

However, in the kernel feature space, ω and t cannot be directly obtained. First, the nonlinear iterative partial least squares (NIPALS) algorithm needs to be kernelized. From Eq. (11) and Eq. (12), the following formula can be obtained:

$$XX^{\mathrm{T}}YY^{\mathrm{T}}X\omega = \lambda X\omega \tag{13}$$

That is to say:

$$XX^{\mathrm{T}}YY^{\mathrm{T}}t = \lambda t \tag{14}$$

$$KYY^{\mathrm{T}}t = \lambda t \tag{15}$$

In the *KPLS* algorithm, the principal component score vector t is the eigenvector corresponding to the maximum eigenvalue of $XX^{\mathrm{T}}YY^{\mathrm{T}}$. In the *KEPLS* algorithm, the corresponding Rényi entropy value can be calculated according to Eq. (10) and the eigenvalue with the largest contribution to the Rényi entropy estimation and its corresponding eigenvector can be selected, which is score matrix t in *KEPLS*. Then, score u of the predictor variable is calculated according to the value of t.

In summary, the specific steps of the *KEPLS* algorithm can be listed as follows:

(1) Calculate score vector t_i of process variable X in the high-dimensional space using the above method and unitize t_i.

(2) Load the matrix of predictor variables: $q_i = Y^{\mathrm{T}}t_i$.

(3) Calculate the score matrix of the predictor variable as $u_i = Yq_i$ and unitize u_i.

(4) Repeat Steps (2) - (4) until u_i values converge.

(5) Calculate the residual information that reflects $\phi(x)$ and Y, as follows:

$$K = \left(I - t_i t_i^{\mathrm{T}}\right)K_i\left(I - t_i t_i^{\mathrm{T}}\right) = K_i - t_i t_i^{\mathrm{T}}K_i - K_i t_i t_i^{\mathrm{T}} + t_i t_i^{\mathrm{T}}K_i t_i t_i^{\mathrm{T}} \tag{16}$$

$$Y_{i+1} = Y_i - t_i t_i^{\mathrm{T}}Y_i \tag{17}$$

5.2 Results and Analysis of Trend Prediction Based on *KEPLS*

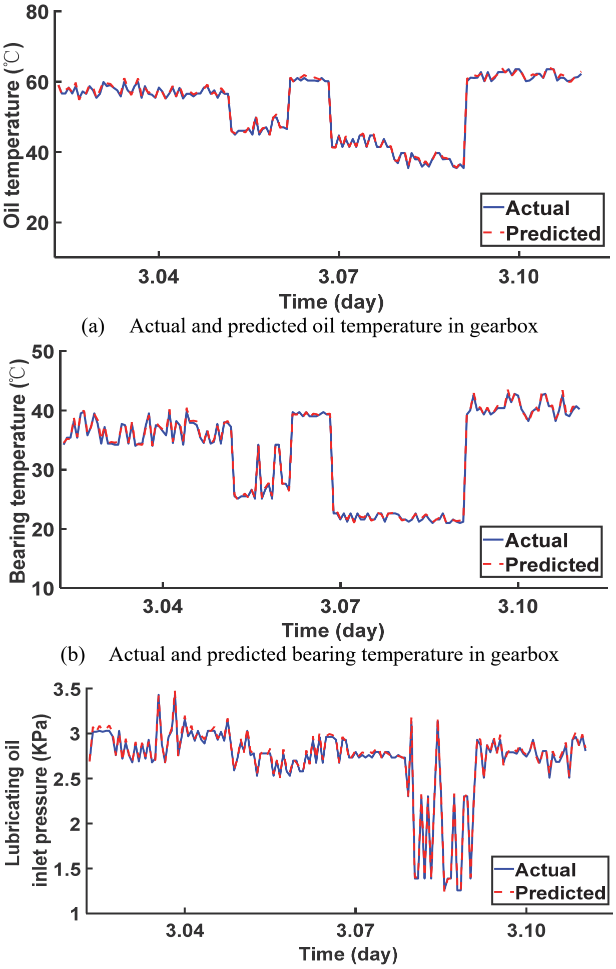
To verify the effectiveness of the *KEPLS* predictive model, information from the experimental training samples was extracted and used to build the predictive model. Information from the model verification sample was used as test data. The prediction outputs are the target parameter values, which were compared with the actual target parameters to determine the accuracy and reliability of the proposed model. The mean absolute error (*MAE*) and mean relative error (*MRE*) were used to evaluate the predictive model, and are calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y(i) - \hat{y}(i)\right| \tag{18}$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y(i) - \hat{y}(i)}{y(i)}\right| \tag{19}$$

where $y(i)$ is the predicted value of the target parameter, $\hat{y}(i)$ is the actual value of the target parameter, and n is the number of samples.

Fig. 7 shows the results of the *KEPLS* model. The predictions were compared with those of the traditional *KPLS* model to demonstrate advantages of the method, as shown in Fig. 8.



(a) Actual and predicted oil temperature in gearbox

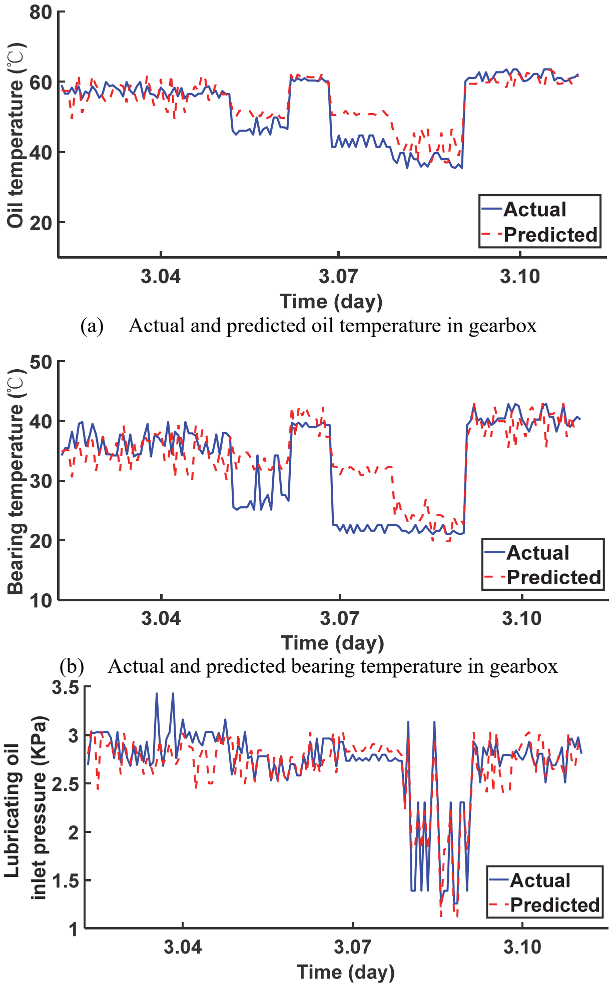
(b) Actual and predicted bearing temperature in gearbox

(c) Actual and predicted lubricating oil inlet pressure in gearbox

Figure 7 Comparison of actual target parameters and *KEPLS* model predictions

Table 1 Mean absolute error (*MAE*) and mean relative error (*MRE*) of *KPLS* and *KEPLS* model predictions

Target parameters	Model	*MAE*	*MRE*
Oil temperature	*KPLS*	3.3620	0.0725
	KEPLS	0.4128	0.0078
Bearing temperature	*KPLS*	3.6275	0.1310
	KEPLS	0.2727	0.0082
Lubricating oil inlet pressure	*KPLS*	0.1462	0.0594
	KEPLS	0.0220	0.0081



(a) Actual and predicted oil temperature in gearbox

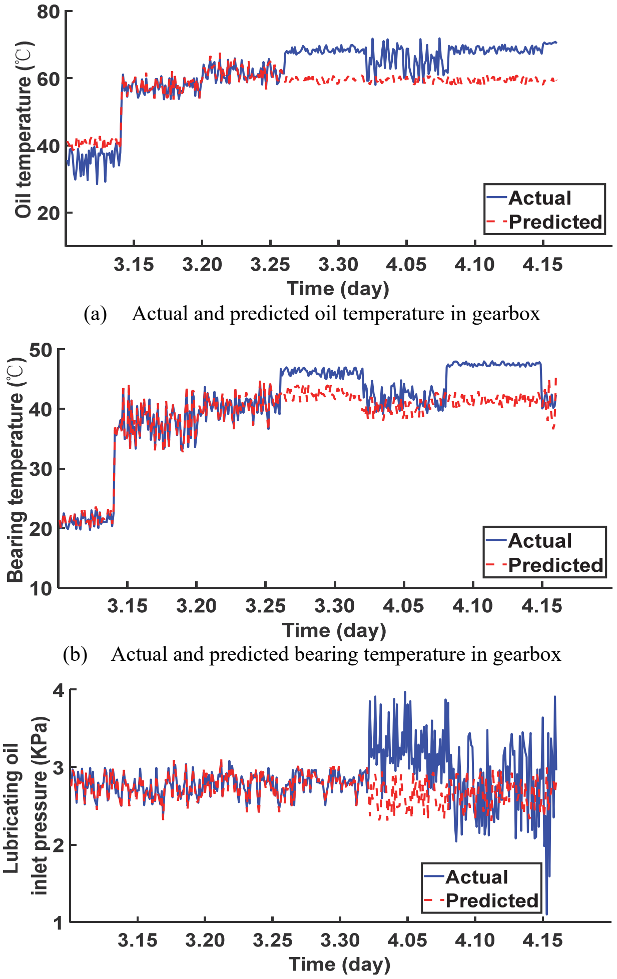
(b) Actual and predicted bearing temperature in gearbox

(c) Actual and predicted lubricating oil inlet pressure in gearbox

Figure 8 Comparison of actual target parameters and *KPLS* model predictions

Fig. 7 and Fig. 8 show the differences between the predicted values of the *KEPLS* model and the *KPLS* model, respectively, versus the actual parameters. Tab. 1 compares the error values obtained for the *KPLS* model and *KEPLS* model. Based on the error results presented in Tab. 1, the *MAE* and *MRE* of the *KEPLS* model are relatively small, suggesting that the output of the model can be directly compared with the actual temperature to evaluate whether the WT is in an abnormal operating state. If the difference between the predicted value and the actual value of the continuous data increases, that is, if the residual increases continuously for a period of time, it indicates a fault. Comparing the error results presented in Tabl. 1, it can be found that the proposed model has a higher fitting accuracy than the *KPLS* model.

Fig. 9 shows the predicted gearbox oil temperature, gearbox bearing temperature, and gearbox lubricating oil inlet pressure. Excluding the influence of model accuracy, the gearbox oil temperature and bearing temperature of this WT were abnormally high around March 26. However, the gearbox lubricating oil inlet pressure was only slightly higher than normal around April 1, representing a lag. The residual value was calculated to further analyze the prediction residual in order to determine the cause and timeframe of the temperature abnormality. The result is shown in Fig. 10.



(a) Actual and predicted oil temperature in gearbox

(b) Actual and predicted bearing temperature in gearbox

(c) Actual and predicted lubricating oil inlet pressure in gearbox

Figure 9 Comparison of actual target parameters and *KEPLS* model predictions over time based on data from March 12 to April 16

5.3 Residual Prediction Analysis Based on *KEPLS*

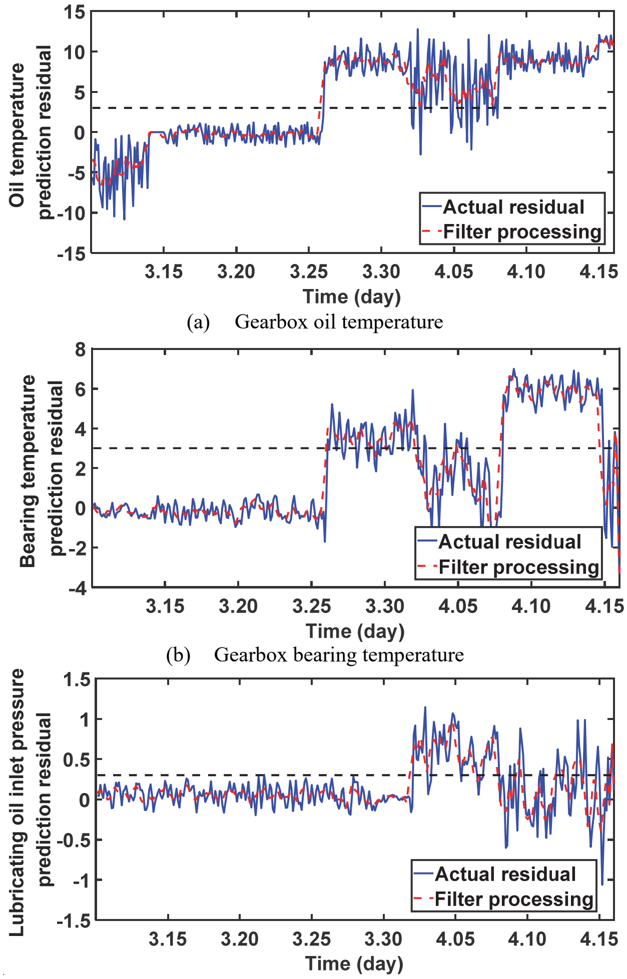
When the gearbox is operating normally, the mean value of the predicted residuals should be close to zero with a very small standard deviation due to the high prediction accuracy of the *KEPLS* model. When the gearbox produces fault symptoms, the residual characteristics will exhibit abnormal fluctuations and the hidden gearbox fault can be detected based on changes in the statistical characteristics of the residual.

Here, the kernel density estimation was used to calculate the residual error threshold. When the residual error of the predictive model exceeds a certain set threshold, it indicates that there may be a fault and an early fault alarm will be issued. However, the residual of a single parameter is not enough to determine the overall health of the gear. The fusion analysis of multiple residuals of all predicted parameters is more stable and reliable. To this end, this paper proposes the fusion residual concept, defined as:

$$\beta=\begin{cases}0, & \forall\lambda_i<L_i\\ \dfrac{\sum\limits_{i=1}^{n}\alpha_i\lambda_i}{n}, & \exists\lambda_i>L_i\\ \max\left(\lambda_i\right), & \forall\lambda_i>L_i\end{cases} \tag{20}$$

where λ_i is the i-th residual error (i = 1, 2, …, n); L_i is the threshold value of the i-th residual; and α_i is the weight factor affecting the health state of the gearbox. The weight factor will be different for different WTs. The value of the weight factor can be deduced from the early operating state of the WT.

Fig. 10 presents the target parameter residuals predicted using the *KEPLS* model. The upper threshold value of the residuals is calculated using the kernel density estimation (KDE) as the high temperature (high pressure) alarm line. When the residual value exceeds the threshold value, the gearbox is in an abnormal state and the high temperature (high pressure) alarm is issued. After March 26, the temperature of the gearbox spindle exceeded the high temperature alarm line and the temperature of the spindle remained abnormal, as shown in Fig. 10. Furthermore, the oil temperature continued to exceed the high-temperature alarm threshold after March 26. The residual pressure diagram of the gearbox lubricating oil inlet shows that residual fluctuation occurred after April 1. Compared with gearbox oil temperature and bearing temperature, the abnormal situation of gearbox lubricating oil inlet pressure lags behind.



(a) Gearbox oil temperature

(b) Gearbox bearing temperature

(c) Gearbox lubricating oil inlet pressure

Figure 10 Prediction residual

Based on the trends of the two residuals, namely, temperature of the gearbox spindle and oil temperature, it can be inferred that the abnormal temperatures of the gearbox components of the WT may be due to a high temperature fault caused by bearing friction. The gearbox fault is caused by heating the oil in the gearbox via heat transfer. In addition, high temperature failure will also cause the oil inlet pressure to change, but more slowly. Therefore, monitoring the target parameters using the predictive model can achieve early fault prediction when abnormal conditions occur in the gearbox. However, the target parameters do not change concurrently and the maximum time difference is about 5 days. Therefore, monitoring only one parameter will likely lead to errors in the prediction results. In view of the above possible problems, the method of multi-parameter fusion residuals was adopted. Through fusion analysis of the residuals of multiple target parameters of the gearbox, mutual restrictive relationships were considered in order to avoid one-sidedness and errors.

According to Eq. (20), the fusion residuals of the three target residuals generated the *KEPLS* model were calculated. The weight factor was selected using trial and error, and the results are shown in Fig. 11. Comparing Fig. 9 and Fig. 10, it can be seen that after fusion, the residual error is more sensitive to abnormal conditions and fewer abnormal assessments are made under normal conditions.

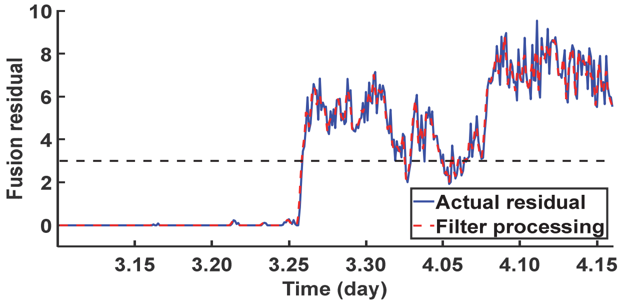


Figure 11 Fusion residual

6 CASE STUDY

According to the health status classification principle, the prediction results for the abnormal state parameters of the WT were divided into five grades: $L = \{l_1, l_2, l_3, l_4, l_5\} = \{\text{health, good, attention, deterioration, disease}\}$.

In this study, the combined triangular and half trapezoidal distributed membership function was adopted to define the health status class membership function according to the actual degradation of WT performance during the early stage:

$$l_1 = \begin{cases} 1, & \beta < a_1 \\ \dfrac{a_2 - \beta}{a_2 - a_1}, & a_1 \le \beta < a_2 \\ 0, & \beta \ge a_2 \end{cases} \tag{21}$$

$$l_2 = \begin{cases} 1, & \beta < a_1 \\ \dfrac{\beta - a_1}{a_3 - a_1}, & a_1 \le \beta < a_3 \\ \dfrac{a_4 - \beta}{a_4 - a_3}, & a_3 \le \beta < a_4 \\ 0, & \beta \ge a_4 \end{cases} \tag{22}$$

$$l_3 = \begin{cases} 1, & 0 \le \beta < a_3 \\ \dfrac{\beta - a_3}{a_4 - a_3}, & a_3 \le \beta < a_4 \\ \dfrac{a_5 - \beta}{a_5 - a_4}, & a_4 \le \beta < a_5 \\ 0, & \beta \ge a_5 \end{cases} \tag{23}$$

$$l_4 = \begin{cases} 1, & 0 \le \beta < a_4 \\ \dfrac{\beta - a_4}{a_5 - a_4}, & a_4 \le \beta < a_5 \\ \dfrac{a_6 - \beta}{a_6 - a_5}, & a_5 \le \beta < a_6 \\ 0, & \beta \ge a_6 \end{cases} \tag{24}$$

$$l_5 = \begin{cases} 0, & 0 \le \beta < a_5 \\ \dfrac{\beta - a_5}{a_6 - a_5}, & a_5 \le \beta < a_6 \\ 1, & \beta \ge a_6 \end{cases} \tag{25}$$

According to the actual operating data of the WT in the early stage, the following values were selected: $a_1 = 1.5$, $a_2 = 4.86$, $a_3 = 8.54$, $a_4 = 11.56$, $a_5 = 14.58$ and $a_6 = 16$. Then, a fuzzy comprehensive evaluation was carried out to analyze the health status of the WT as accurately as possible.

The radar chart, also referred to as the network chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables on a single axis starting from the same point. The relative position and angle of the shaft are usually uniform. The chart is equivalent to a parallel coordinate diagram with radially arranged axes. The radar chart is mainly used to evaluate operating conditions in terms of profitability, productivity, liquidity, and safety. Here, the radar chart is applied to the health assessment of a WT to obtain a more intuitive representation of the results.

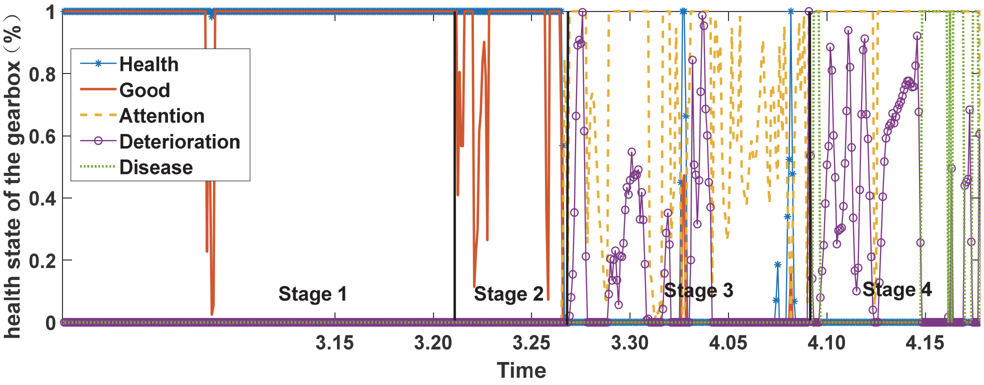
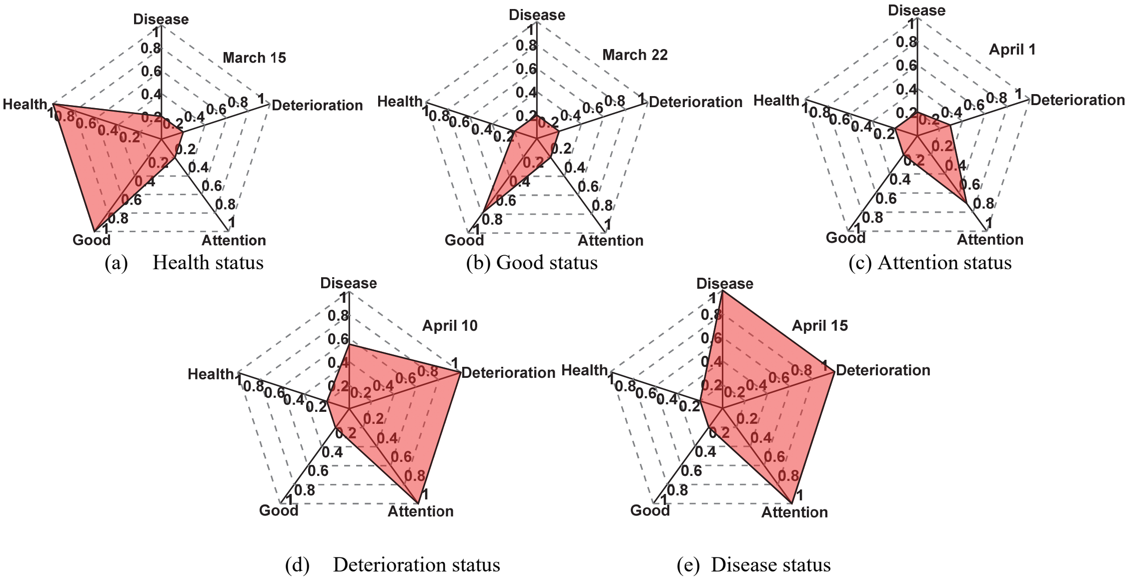


Figure 12 Wind turbine health assessment results



(a) Health status

(b) Good status

(c) Attention status

(d) Deterioration status

(e) Disease status

Figure 13 Radar charts of wind turbine health indicators on different dates

7 CONCLUSION

This paper proposed an algorithm for monitoring and assessment of wind turbine degradation performance based on the kernel entropy method. The method can be used to monitor the state of a wind turbine gearbox, predict multivariate process data, and comprehensively evaluate the health state of the gearbox according to multi-model information fusion and residual error prediction. The *KECA* algorithm was applied to assess the wind turbine state by introducing the Rényi quadratic entropy and selecting the principal element based on the entropy value to process the nonlinear SCADA data. This approach reveals structural information of the data, and also ensures that information is not lost, to the greatest extent possible. Then, the proposed *KEPLS* method was used to improve the prediction performance by considering higher-order information. The method not only extracts multi-scale information better but can also be used to perform a fusion analysis on multiple residuals obtained from the predictive model, thereby enhancing the stability and reliability of the algorithm. Finally, an actual wind farm case study was used to demonstrate that the prediction model and assessment method are accurate, simple, and intuitive, and can be used to evaluate the health status of WT components. The proposed method has good application prospects for WT health monitoring.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China (Grant number 61763037, 61803029), the Natural Science Foundation of Inner Mongolia(2019LH06007, 2020MS05029), and the Inner Mongolia Autonomous Region Science and Technology Plan Project (2019, 2020GG0283).

8 REFERENCES

[1] E1-Thalji, I. & Jantunen, E. (2015). A summary of fault modelling and predictive health monitoring of rolling element bearings. *Mechanical Systems & Signal Processing, 60-61*, 252-272. https://doi.org/10.1016/j.ymssp.2015.02.008

[2] Zhao, Y. Y., Li, D. S., Dong, A., Kang, D. H., Lv, Q. & Shang, L. (2017).Fault prediction and diagnosis of wind turbine generators using scada data. *Energies*, *10*(8), 1210. https://doi.org/10.3390/en10081210

[3] Tautz-Weinert, J. & Watson, S. (2017). Using SCADA data for wind turbine condition monitoring-areview. *IET Renewable Power Generation*, *11*(4), 382-394. https://doi.org/10.1049.jet-rpg.2016.0248

[4] Cao, M. N., Qiu, Y. N., Feng, Y. H., Wang, H., & Li, D. (2016). Study of wind turbine fault diagnosis of based on unscented Kalman filter and SCADA data. *Energies*, *9*(10). https://doi.org/10.3390/en9100847

[5] Li, H., Yang, C., Li, X. W., Ji, H. T., Qin, X., Chen, Y. J., Yang, D., & Tang, X. H. (2014). Condition characteristic parameters mining and outlier identification for electric pitch system of wind turbine. *Proceedings of the CSEE*, *34*(12), 1922-1930.

[6] Li, H., Hu, Y. G., Li, Y., Yang, D., Liang, Y. Y., Ouyang, H. L., & Lan, Y. S. (2016). Overview of condition monitoring and fault diagnosis for grid-connected high-power wind turbine unit. *Electric Power Automation Equipment*, *36*(1), 6-16.

[7] Leahy, K., Hu, R. L., Konstantakopoulos, I. C., Spanos, C. J., & Agogino, A. M. (2016). Diagnosing wind turbine faults using machine learning techniques applied to operational data. *2016 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE*, 20-22. https://doi.org/10.1109/ICPHM.2016.7542860

[8] Lapira, E., Brisset, D., Ardakani, H. D., Siegel, D., & Lee, J. (2012). Wind turbine performance assessment using multi-regime modeling approach. *Renewable Energy*, *45*, 86-95. https://doi.org/10.1016/j.renene.2012.02.018

[9] Wang, S. Y., Huang, Y. X., Li, L., & Liu, C. L. (2016). Wind turbines abnormality detection through analysis of wind farm power curves. *Measurement*, *93*, 178-188.

[10] Guo, P., Jiang, M. L., & Li, H. T. (2016). Performance analysis and monitoring based on scada data and gaussian process regression for wind turbine power generation. *Electric Power Automation Equipment*, *36*(8), 10-15.

[11] Sun, P., Li, J., Kou, X. K., Lv, Z. B., Yao, D. G., Wang, J., Wang, L. L., & Teng, W. J. (2017). Wind turbine status parameter anomaly detection based on prediction models and fuzzy theory. *Electric Power Automation Equipment*, *37*(8), 90-98.

[12] Zhou, Q., Xu, Q. P., Li, J., Wang, M. B., & Xiang, C. M. (2017). Operating state assessment based on set-pair analysis and evidential reasoning decision-making for wind turbine generator unit. *Electric Power Automation Equipment*, *37*(7), 38-45.

[13] Du, N., Yi, J., Mazidi, P. M., Cheng, L., & Guo, J. B. (2017). A parameter selection method for wind turbine health management through scada data. *Energies*, *10*(2), 253. https://doi.org/10.3390/en10020253

[14] Garcia-Alvarez, D., Fuente, M. J., & Sainz, G. I. (2012). Fault detection and isolation in transient states using principal component analysis. *Journal of Process Control, 22*(3), 551-563.https://doi.org/10.1016/j.jprocont.2012.01.007

[15] Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, *36*(2), 220-234. https://doi.org/10.1016/j.arcontrol.2012.09.004

[16] Scholkopf, B., Smola, A., & Muller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, *10*(5), 1299-1319. https://doi.org/10.1162/089976698300017467

[17] Jenssen, R. (2010). Kernel entropy component analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *32*(5), 847-860. https://doi.org/10.1109/TPAMI.2009.100

[18] Liu, Y., Wu, Q. Y., & Chen, J. H. (2017). Active selection of informative data for sequential quality enhancement of soft sensor models with latent variables. *Industrial & Engineering Chemistry Research*, *56*(16), 4804-4817. https://doi.org/10.1021/acs.iecr.6b04620

[19] Liu, Y., Yang, C., Gao, Z. L., & Yao, Y. (2018). Ensemble deep kernel learning with application to quality prediction in industrial polymerizationprocesses. *Chemometrics & Intelligent Laboratory Systems*, *174*, 15-21. https://doi.org/10.1016/j.chemolab.2018.01.008

[20] Liu, K. X., Ma, Z. Y., Liu, Y, Yang, J. G., & Yao, H. (2021). Enhanced defect detection in carbon fiber reinforced polymer composites via generative kernel principal component thermography. *Polymers*, *13*(5), 825. https://doi.org/10.3390/polym13050825

[21] Deng, H. Y., Yang, K. Y., Liu, Y., Zhang, S. C., & Yao, Y. (2020). Actively exploring informative data for smart modeling of industrial multiphase flow processes.Actively exploring informative data for smart modeling of industrial multiphase flow processes. *IEEE Transactions on Industrial Informatics*, *PP*(99), 1-1. https://doi.org/10.1109/TII.2020.3046013

[22] Chen, Y., Zhu, F. B. & Jay, L. (2013).Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method. *Computers in Industry*, *64*(3), 214-225. https://doi.org/10.1016/j.compind.2012.10.005

[23] Qi, Y. S., Zhang, H. L., Gao, X. J., & Wang, P. (2016). Novel Fault Monitoring Strategy for Chemical Process Based on *KECA*. *CISEC Journal*, *67*(3), 1063-1069

[24] Yang, Y. H., Li, X. L., Liu, X. Z., & Chen, X. B. (2015). Wavelet kernel entropy component analysis with application to industrial process monitoring. *Neurocomputing*, *147,* 395-402. https://doi.org/10.1016/j.neucom.2014.06.045

[25] Jiang, Q. C., Yan, X. F., Lv, Z. M., & Guo, M. J. (2013). Fault detection in nonlinear chemical processes based on kernel entropy component analysis and angular structure. *Korean Journal of Chemical Engineering*, *30*(6), 1181-1186. https://doi.org/10.1007/s11814-013-0034-7

[26] Gomez-Chova, L., Jenssen, R., & Camps-Vall, G. (2012). Kernel entropy component analysis for remote sensing image clustering. *IEEE Geoscience and Remote Sensing Letters*, *9*(2), 312-316. https://doi.org/10.1109/LGRS.2011.2167212

Contact information:

Yong-Sheng QI, Professor, PhD
(Corresponding author)
Inner Mongolia University of Technology,
School of Electric Power,
49 Aimin Street, Xincheng District, Hohhot, China
E-mail: qys@imut.edu.cn

Chao REN, postgraduate
Inner Mongolia University of Technology,
School of Electric Power,
49 Aimin Street, Xincheng District, Hohhot, China
E-mail: 3180734266@qq.com

Xue-Jin GAO, Professor, PhD
Beijing University of Technology,
School of Information Department,
100 Ping Leyuan, Chaoyang District, Beijing, China
E-mail: gaoxuejin@bjut.edu.cn

Li-Qiang LIU, Professor, PhD
Inner Mongolia University of Technology,
School of Electric Power,
49 Aimin Street, Xincheng District, Hohhot, China
E-mail: llqiang@imut.edu.cn

Chao-Yi DONG, Professor, PhD
Inner Mongolia University of Technology,
School of Electric Power,
49 Aimin Street, Xincheng District, Hohhot, China
E-mail: dongchaoyi@imut.edu.cn