IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 January 2023, accepted 2 February 2023, date of publication 9 February 2023, date of current version 14 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243625

RESEARCH ARTICLE

Wind Turbine Condition Monitoring Based on Improved Active Learning Strategy and KNN Algorithm

CHENGJIA BAO[1], TIANYI ZHANG[2], ZHIXI HU[2], WANGJING FENG[2], AND RU LIU[2]
[1]State Grid Gansu Electric Power Company, Lanzhou, Gansu 730030, China
[2]State Grid Lanzhou Electric Power Supply Company, Lanzhou, Gansu 730070, China

Corresponding author: Tianyi Zhang (z1073538532@163.com)

This work was supported by the second batch of comprehensive plan of State Grid Gansu Power Company, in 2022, under Grant 522701220005.

ABSTRACT As the damage of the gearbox of wind turbines (WTs) will cause economic losses, it is necessary to conduct online condition monitoring (CM) on the gearbox. Most WTs are equipped with SCADA system, and CM method based on Supervisory Control and Data Acquisition (SCADA) data is one of the most economical methods. K Nearest Neighbor (KNN) algorithm has good robustness, and WTs are typical nonlinear objects. Based on this, KNN regression model is established for CM, and Distance Correlation (DC) coefficient is used to select modeling variables to improve the shortcomings of traditional feature selection algorithm. A large amount of redundant data will be generated during the operation of WTs, and the efficiency of KNN algorithm is affected by the size of training set. Therefore, an active learning (AL) algorithm combining multiple strategies is proposed to select high-quality training data. The validity of the proposed method is verified by the data of an actual WT. The experimental results show that the method presented in this paper performs well in the comparative experiments, and the online CM results are about 20 days earlier than the SCADA system.

INDEX TERMS Wind turbine gearbox, condition monitoring, K-nearest neighbor algorithm, active learning, distance correlation coefficient.

I. INTRODUCTION

Due to the increasingly serious global environmental problems, the development of renewable energy has become a hot issue of global concern [1], [2]. Wind energy is a common renewable energy, which has the advantages of abundant reserves and low utilization cost. Therefore, many countries have begun to vigorously develop wind power generation technology [3], [4], [5]. With the popularization of wind power generation technology in the world, the installed capacity of wind turbines(WTs) has also increased, thus consuming high operation and maintenance costs [6]. WT is a complex equipment with hundreds of subsystems and components. According to the relative research [7], the downtime due to gearbox failure accounts for about 20% of all downtime, so the gearbox is considered to be one of the most troublesome components. Therefore, it is necessary to study the condition monitoring (CM) technology of WT gearbox [8].

The condition monitoring (CM) method can be divided into signal-trending analysis, model-based, and data-driven methods. Signal-trending analysis usually uses the vibration signal [10], acoustic emission signals [11] and electrical signals [12], etc. However, the acquisition of these signals requires the installation of professional sensors, which will cause additional costs. The model-based method is to establish accurate models of subsystems and components through rich expert knowledge [13]. However, due to the complexity of WT operation, expert knowledge is difficult to obtain; And the model usually focuses on a specific component, which cannot be migrated to other components.

The associate editor coordinating the review of this manuscript and approving it for publication was Chong Leong Gan.

Since almost all wind turbines have installed Supervisory Control and Data Acquisition (SCADA) system, data- driven methods have become the focus of research [14]. Wang et al. [15] proposed a data-driven CM method based on incremental learning and multivariate state estimation technique, which updates the training set in real time while keeping the running time unchanged. Zhang et al. [16] proposed an machine learning method based on random forest and the eXtreme Gradient Boosting to establish the data-driven WT fault detection framework. Shi et al. [17] proposed a CM method based on XGBoost algorithm, and also proposed a data preprocessing method based on Density-Based Spatial Clustering of Applications. The above methods not only propose CM methods, but also take the selection of training samples into consideration.

In fact, due to the complex operation conditions of WTs and the massive historical data recorded by SCADA system, how to select high-quality training samples for data-driven methods has always been a hot issue [18]. Some researchers [19], [20] pointed out that active learning (AL) is better than learning from samples, and selecting high-quality samples can effectively improve the generalization ability of learners. Huang et al. [21] proposed a principle based on the min-max view of AL to provide a systematic way for measuring and combining the informativeness and representativeness. Ozdemir et al. [22] proposed a new method combining the representativeness and uncertainty to estimate the ideal samples from a given data set.

K-nearest neighbor regression (KNNR) algorithm is used in many field due to its robustness and many scholars have improved it from different aspects. Song et al. [23] proposed a instance selection method for KNNR algorithm, which deleted outlier instances and the little-contribution instances to decrease the size of training set. Hu et al. [24] established a KNNR model using the adaptation of particle swarm optimazation, which minimizes the cross validation error in the capacity estimation. Zhang et al. [25] proposed data-driven CM method based on ensemble K-nearest neighbor (KNN), which can achieve the desired estimation accuracy and improve the operation efficiency. However, the optimization of the above methods to the model is a one-time optimization, and there is no dynamic optimization.

The operation process of the WT gearbox will produce a large number of repeated low-quality data. To address this phenomenon, a WTCM method based on AL strategy and KNNR algorithm is proposed. Firstly, select modeling features based on distance correlation (DC) coefficient. Then, design an AL sample selection algorithm based on uncertainty and representativeness to select high-quality samples to establish KNNR model, Finally, the Statistical Process Control (SPC) technology is used to process the residual of the model output value and observed value to achieve CM. The SCADA data collected from a WT is used to validate the feasibility of industrial application of the proposed approach. Results shows that the proposed method can realize gearbox CM and provide health rate indicators.

The rest of this paper is organized as follow. Section II gives a detailed description of KNNR algorithm, DC coefficient and AL strategy. Section III presents the framework of the proposed WTCM method. SectionIV shows the results of experiments to validate the proposed method. The experimental results are summarized and the conclusions are given in Section V.

II. METHODOLOGY

This section introduce the KNNR algorithm, AL strategy and DC coefficient.

A. KNNR ALGORITHM

KNNR algorithm is a common inert algorithm. Its basic principle is to estimate a testing sample by finding training samples similar to the testing samples through distance [25]. The specific steps are as follows:

For a testing sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n, y)$ and the training samples $\boldsymbol{X} = \{\boldsymbol{x_1}; \boldsymbol{x_2}; \ldots; \boldsymbol{x_m}\}, \boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{in}, y_i), i \in [1, m]$.

(1) Calculate the distance between $\boldsymbol{x}$ and all training samples:

$$d_j(\boldsymbol{x}, \boldsymbol{x_i}) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \tag{1}$$

where d_j is the distance between $\boldsymbol{x}$ and $\boldsymbol{x_i}$.

(2) Find K training samples $\boldsymbol{X^K} = \{\boldsymbol{x_1^K}; \boldsymbol{x_2^K}; \ldots; \boldsymbol{x_K^K}\}, \boldsymbol{x_p^K} = (x_{p1}^K, x_{p2}^K, \ldots, x_{pn}^K, y_p^K)$ closest to $\boldsymbol{x}$.

(3) Calculate the output of the model by averaging:

$$\hat{y} = \frac{1}{K} \sum_{p=1}^{K} y_p^K \tag{2}$$

where $\hat{y}$ is the output of the model.

B. DC COEFFICIENT

From the previous theoretical analysis, it can be seen that the selection of prediction variables has a great impact on the performance of the regression model. Therefore, selecting appropriate prediction variables is a prerequisite for establishing an ideal model. At present, the commonly used feature selection methods include Pearson correlation coefficient, Spearman correlation coefficient and mutual information. However, these methods can only measure the relationship between the variables with linear correlation, and has poor effect on the characteristics of nonlinear correlation.

For a large number of parameters with complex correlation in WTs, DC coefficient [26] overcomes the shortcomings of Pearson correlation coefficient, which can not only reflect the linear relationship between variables, but also represent the nonlinear relationship between variables.

For 2 random variables $\boldsymbol{x} \in R^n$ and $\boldsymbol{y} \in R^n$, $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i), i = 1, 2, \ldots, n\}$, the DC coefficient is defined as:

$$R^2(\boldsymbol{x}, \boldsymbol{y}) = \frac{\upsilon^2(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\upsilon^2(\boldsymbol{x}, \boldsymbol{x})\upsilon^2(\boldsymbol{y}, \boldsymbol{y})}} \tag{3}$$

where

$$\upsilon^2(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}\boldsymbol{B}_{i,j} \tag{4}$$

$$A_{i,j} = \left\|x_i - x_j\right\|_2 - \frac{1}{n}\sum_{k=1}^{n}\left\|x_k - x_j\right\|_2 - \frac{1}{n}\sum_{l=1}^{n}\|x_i - x_l\|_2 + \frac{1}{n^2}\sum_{k,l=1}^{n}\|x_k - x_l\|_2 \tag{5}$$

$$B_{i,j} = \left\|y_i - y_j\right\|_2 - \frac{1}{n}\sum_{k=1}^{n}\left\|y_k - y_j\right\|_2 - \frac{1}{n}\sum_{l=1}^{n}\|y_i - y_l\|_2 + \frac{1}{n^2}\sum_{k,l=1}^{n}\|y_k - y_l\|_2 \tag{6}$$

Similarly, $\upsilon^2(\boldsymbol{x},\boldsymbol{x})$ and $\upsilon^2(\boldsymbol{y},\boldsymbol{y})$ can be calculated as:

$$\upsilon^2(\boldsymbol{x},\boldsymbol{x}) = \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^2 \tag{7}$$

$$\upsilon^2(\boldsymbol{y},\boldsymbol{y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{B}_{i,j}^2 \tag{8}$$

The range of DC coefficient is [0,1]. If the DC coefficient between two variables is closer to 1, it indicates that their correlation is stronger. Therefore, the variable with strong correlation with the prediction variable is selected as the modeling variable.

C. AL STRATEGY

During the actual operation of WTs, a large number of samples labeled as "normal" will be generated [9]. If these samples are directly used as training samples, the following problems will exist: (a). Low quality samples may be mixed in the samples, making the established model difficult to achieve the desired prediction accuracy. (b). There is a lot of redundancy in the sample, which wastes storage space. To solve the above problems, AL strategy is used to develop appropriate sample selection methods, and actively select the samples that can best improve the performance of the current model, so as to maximize the performance of the model, and effectively alleviate the dependence of the model on the number of training samples. The research object of this paper is pool-based AL strategy.

As shown in Fig. 1, AL strategy is a process of iteratively selecting samples. First, use the existing training set training the model, score the candidates and determine the selection order according to the model and sample selection query Q, and then select the top ranked samples for manual judgment. If the conditions are met, add the training set and start a new round of sample selection until the stop condition is reached.

The three common strategies for pool-based AL are bases on uncertainty, diversity and representativeness. As shown in Fig. 2, green dots represent training samples, and other dots represent candidates in the sample pool, nd the black solid line divides the samples into two clusters. Uncertainty means that the candidate is located at the decision boundary or has the minimum conditional entropy, such as the blue dot in Fig. 2. Selecting such samples can expand the decision space. Representativeness means that the candidate is located in the center of the cluster or high density area, as shown in the orange point in Fig. 2. Select this type of sample will obtain a more typical sample. Diversity means that candidates are far away from training samples, such as the yellow dots in Fig. 2. Since such samples often introduce outliers, this property is not considered in this paper.

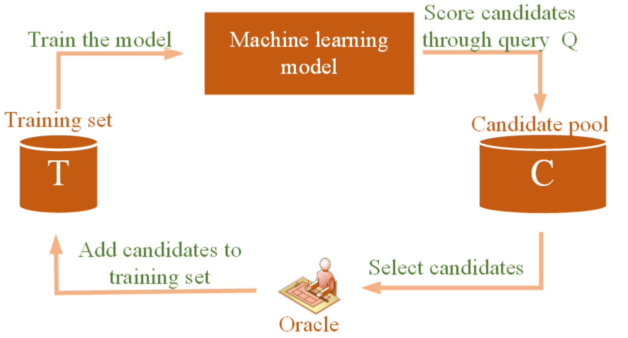

FIGURE 1. The framework of AL strategy.

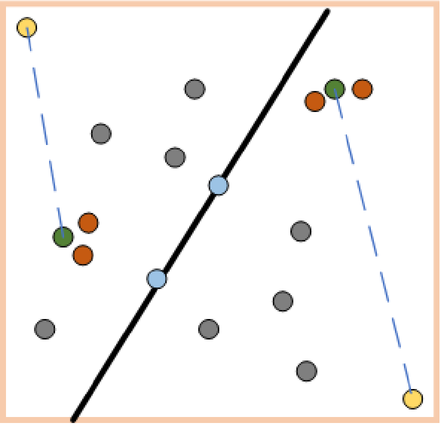

FIGURE 2. Schematic diagram of AL strategy.

III. PROPOSED WTCM MTHOD

This section will introduce the framework of the proposed WTCM method, and gives a detailed description of each phase.

A. AL SAMPLE SELECTION METHOD BASED ON UNCERTAINTY AND REPRESENTATIVENESS

In order to select high-quality samples to construct training set, an AL sample selection method considering uncertainty and representativeness is proposed. The flow chart of sample selection method is as Fig. 3.

Firstly, the uncertainty of a candidate is scored based on the model prediction error.

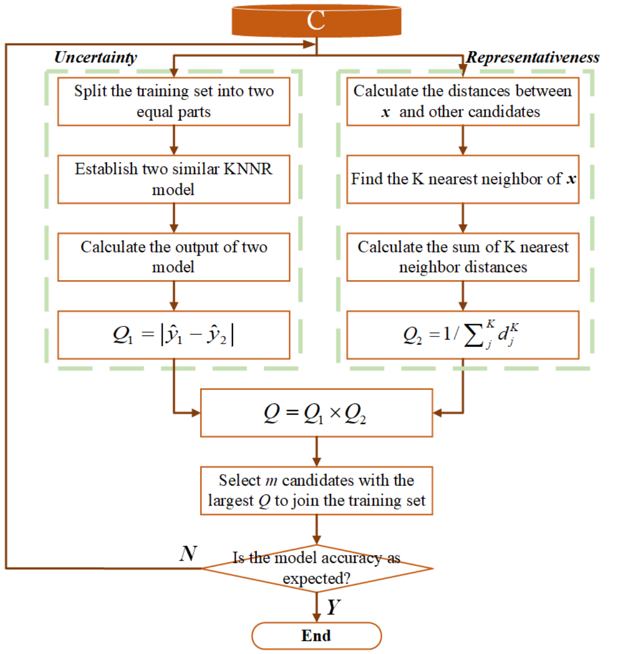
C
Uncertainty
Representativeness
Split the training set into two equal parts
Establish two similar KNNR model
Calculate the output of two model
$Q_1 = |\hat{y}_1 - \hat{y}_2|$
Calculate the distances between x and other candidates
Find the K nearest neighbor of x
Calculate the sum of K nearest neighbor distances
$Q_2 = 1/\sum_j^K d_j^K$
$Q = Q_1 \times Q_2$
Select m candidates with the largest Q to join the training set
Is the model accuracy as expected?
N
Y
End

FIGURE 3. The flow chart of AL-based sample selection method.

(1) Randomly divide the training set into two parts D_1 and D_2 with the same size, and establish KNNR model based on these two training sets.

(2) Input candidate $\boldsymbol{x}$ into the two models and obtain the output $\hat{y}_1$ and $\hat{y}_2$ of the models.

(3) The score of uncertainty can be defined as follows:

$$Q_1 = |\hat{y}_1 - \hat{y}_2| \tag{9}$$

The greater the Q_1, the greater the uncertainty of the candidate.

Then, the representativeness of a candidate is scored based on the K-nearest neighbor distance.

(1) Calculate the Euclidean distances between candidate $\boldsymbol{x}$ and other candidates in the pool.

(2) Find K nearest distances d_i^K, $i \in [1, K]$ of $\boldsymbol{x}$.

(3) Calculate the sum of the K nearest distances:

$$d_s = \sum_{i=1}^{K} d_i^K \tag{10}$$

And the score of representativeness can be defined as follows:

$$Q_2 = 1/d_s \tag{11}$$

That is, the smaller the K-nearest neighbor distance of the candidate, the higher the representativeness.

Finally, the comprehensive score Q of candidate $\boldsymbol{x}$ is calculated. In order to avoid the impact of different orders of magnitude, multiplication is used:

$$Q = Q_1 \times Q_2 \tag{12}$$

Therefore, candidates with high comprehensive scores should be preferred.

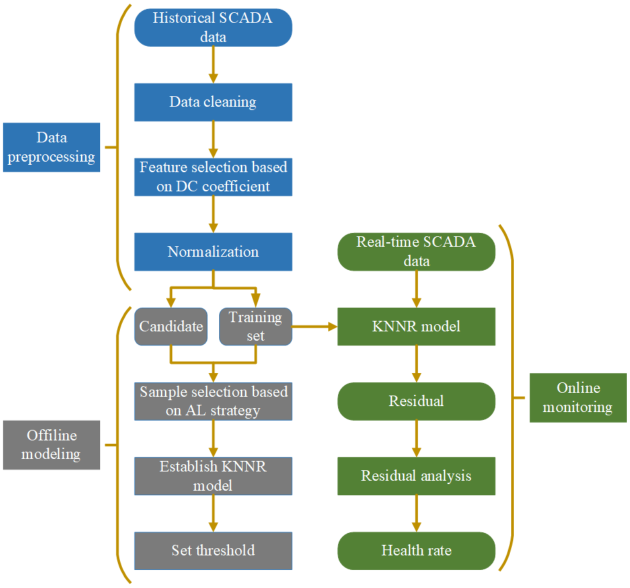

FIGURE 4. The flow chart of the proposed WTCM method.

B. THE PROPOSED WTCM METHOD BASED ON NORMAL BEHAVIOR MODELING

Normal behavior modeling (NBM) [27] is an important branch of data-driven WTCM. Its principle is to use historical normal data to establish a model about predictive variables, and to judge whether the equipment operation state is abnormal by analyzing the observed value and residual error of predictive variables. The proposed WTCM method is divided into data preprocessing, offline modeling and online monitoring:

(1) Data preprocessing includes removing missing or abnormal data for data cleaning, selecting features for model establishment based on DC coefficient, and performing normalization processing

(2) In the offline modeling stage, firstly, high-quality training samples should be selected from the candidates based on AL strategy. Then the KNNR model should be established based on NBM. Finally, the EWMA method should be used to analyze the output of the model and design the threshold.

(3) The online monitoring stage is to input the real-time SCADA data into the established KNNR model to obtain the residual of the output value and observation value of the prediction characteristics, and use the sliding window method to analyze the residual. Finally obtain the health rate of the equipment in combination with the set threshold.

IV. CASE ANALYSIS

A. DATA DESCRIPTION

The SCADA data used to verify the progressiveness of the proposed WTCM method is collected from a real WT with a rated power of 1.5MW in Hebei Province, China. The cut-in wind speed of this double fed WT is 3m/s, the rated wind speed is 12m/s, the cut-out wind speed is 25m/s, and the sampling interval of SCADA data is 1min. According

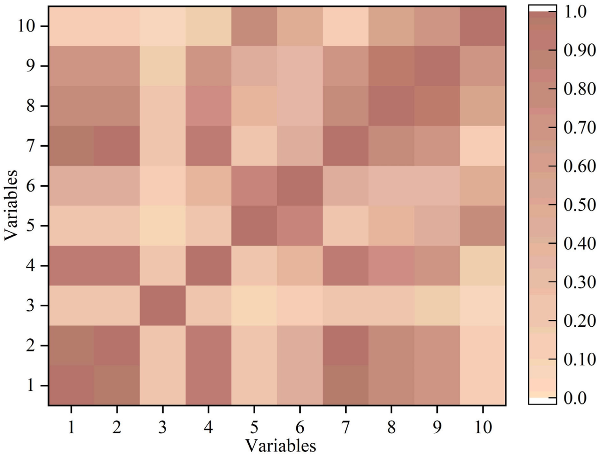

FIGURE 5. The DC coefficient of variables.

to the fault record of SCADA system, the WT had a fault named "gearbox oil temp higher than the upper limit" from 2017/11/17 8:30 to 2017/11/18 14:30.

The sampling interval of SCADA system is 1min, and the theoretical maximum sample size will more than 430,000. However, the data driven modeling method in this paper does not need to use a lot of training data, so the data set is de sampled, and the sampling interval is increased to 5 minutes to reduce the amount of data.

B. DATA PREPROCESSING

This experiment uses the SCADA data of a wind farm in Hebei Province, China from 2017/02/01 0:00-2017/11/17 8:30. Firstly, the unavailable data in historical SCADA data are deleted, including the following data: missing data, data with active power less than or equal to zero, data with wind speed less than the cut-in speed, and data with wind speed greater than the cut-off wind speed. The samples with abnormal operating parameters are removed based on the Laida criterion, and 52,000 samples are finally left.

There are about 60 operating parameters recorded in the SCADA system of the WT. Since this paper studies gearbox faults, the parameters related to the pitch system, yaw system and some grid side that are not closely related to the gearbox are not considered for the time being. Therefore, select the following variables: generator active power, generator speed, ambient temp, wind speed, gearbox oil inlet pressure, gearbox oil inlet filter pressure, main bearing speed, gearbox high drive bearing temp, gearbox high non-drive bearing temp and gearbox oil temp.

Since the gearbox oil temp is often used to reflect the health of the gearbox, the gearbox oil temp is selected as the prediction variable. The DC correlation coefficient between variables is shown in Fig. 5.

According to the DC coefficient between the variables, the correlation between gearbox high non-drive bearing temp and gearbox high drive bearing temp is as high as 0.9573, and the DC coefficient between generator active power, generator speed and main bearing speed also exceeds 0.9. Therefore, gearbox high non-drive bearing tempt and generator active power are retained in these five variables. The ambient temp and wind speed are added to the input variables as variables that can reflect the environmental factors. Finally, according to the DC coefficient, the selected input variables are generator active power, ambient temp, wind speed, gearbox oil inlet pressure, gearbox oil inlet filter pressure and gearbox high non-drive bearing temp. The range of the variables is shown in Tab. 1.

TABLE 1. The range of selected variables.

Variables	Range
Generator active power/kW	[0.05,1725.8]
Ambient temp/ °C	[-20.7,27.1]
Wind speed/$(m \cdot s^{-1})$	[3,15.67]
Gearbox oil inlet pressure/bar	[2.60,3.44]
Gearbox oil inlet filter pressure/bar	[3.94,6.59]
Gearbox high non-drive bearing temp/°C	[51.51,74.06]
Gearbox oil temp/ °C	[50.49,70.09]

Normalize the samples to avoid dimensional influence, the formula is as follows:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{13}$$

where x is the raw data, $x_{\min}$ is the minimum of the corresponding parameter, $x_{\max}$ is the maximum of the corresponding parameter.

Select 2,5000 data from 6/1 0:00 to 11/17 8:30 as the experimental data. The first 1,5000 data are used as the original training set, the 1,5001st-1,6000th data as the verification set, and the 1,6001st-2,5000th data as the testing set.

C. PERFORMANCE ANALYSIS OF AL-BASED SAMPLE SELECTION

This part will discuss the relevant parameters of sample selection based on AL, and compare the sample selection method proposed in this paper with other sample selection methods. The quality of the training set is measured by its Mean Absolute Error(MAE), Root Mean Square Error(RMSE) and R-square on the verification set:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{14}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{16}$$

where m is the number of testing samples, $\bar{y}$ is the mean value of observed value.

1,000 training samples are randomly selected from 1,5000 training samples as the benchmark training set, and the

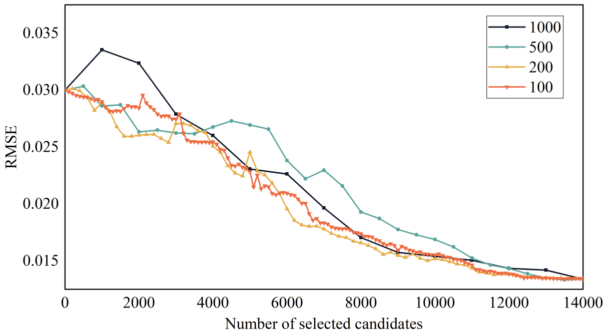

FIGURE 6. The RMSE with different number of selected candidates.

TABLE 2. The ablation experiments of proposed method.

Method	I	II	III	IV	V	VI
RMSE	0.061	0.045	0.030	0.028	0.025	**0.019**
MAE	0.049	0.038	0.024	0.023	0.022	**0.014**
R^2	0.78	0.80	0.84	0.85	0.86	**0.90**

remaining 1,4000 training samples are selected as candidates. First, the step size of the sample selection method should be determined, that is, the number of candidates selected each iteration.

Fig. 6 shows the change of RMSE at different step sizes, which also reflects the change of training set quality. When the step size is 200, most of the RMSE curves are at the lowest of the four curves, and the fluctuation amplitude is small. When the number of selected candidates reaches about 6,000, the RMSE with a step size of 200 is always the lowest. A few peaks on the curve may be due to the introduction of some outliers, but with the increase of the number of selected candidates, RMSE still decreases. When the step size is 100, rmse has several peaks, which indicates that a small step size may introduce outliers to reduce the generalization accuracy of the model. When the step size is large, RMSE does not have a peak, but fluctuates greatly. Therefore, the step size is determined to be 200.

Combined with the elbow rule, when the step size is 200, and the number of selected candidates reaches 6,000, the RMSE declines slowly and the curve trend gradually flattens. Therefore, the number of selected candidates is determined to be 6,000 (the number of training samples is 7,000), and the RMSE is 0.019 at this time.

We will conduct ablation experiments on the method proposed in this paper. The subjects are I. Benchmark training set without feature selection; II. Pearson correlation coefficient + benchmark training set; III. DC coefficient + benchmark training set; IV. DC coefficient + AL sample selection based on uncertainty; V. DC coefficient + AL sample selection based on representativeness; VI. DC coefficient + proposed AL sample selection method.

According to the ablation experiment, method III performs better than method I and method II in all aspects of the verification set, so it can be seen that the DC coefficient has a good effect on the nonlinear object such as WT.

TABLE 3. The performance of KNNR model with different training sets.

Training set	Benchmark	AL selected	Original
RMSE	0.030	0.019	0.014
MAE	0.024	0.014	0.010
R^2	0.84	0.90	0.92

TABLE 4. The performance of KNNR model with different training sets.

Method	AL selected	PL	DAL
RMSE	**0.019**	0.024	0.026
MAE	**0.014**	0.019	0.022
R^2	**0.90**	0.86	0.85

Compared with Method III, IV, V and VI, the performance of Method VI is better than that of the other three methods in all aspects, which proves that the AL sample selection method proposed in this paper, which considers uncertainty and representativeness, is better than the AL method considering single strategy, and can select training samples with better quality.

The following table lists the performance and operation time of KNNR model on the verification set under different training sets.

The RMSE of the verification set on the training set selected based on AL decreased by 36.6%, MAE decreased by 41.1%, and R^2 increased by 7.1% compared with the benchmark training set. Compared with the original training set, RMSE increased by 35.7%, MAE increased by 40%, R decreased by 2.2%. Because the complexity of KNN model is about $O(n)$, n is the number of training samples, the operation time of the AL-selected training set is less than half of the original training set

Next, we investigate the performance of the sample selection method in this paper, passive learning (PL) method and distance based active learning (DAL) [28] methods when the number of selected candidates is consistent. PL refers to randomly selecting a certain number of candidates to join the training set each time, while DAL refers to selecting the candidate with the farthest distance to join the training set.

Compared with AL, the rmse of PL increased by 26.3%, MAE increased by 35.7%, and R^2 decreased by 4.4%. Meanwhile, the RMSE of DAL increased by 36.8%, MAE increased by 57.1%, and r increased by 5.5%. The worst performance of DAL may be due to the introduction of a large number of outliers by selecting the sample farthest from the training set. In contrast, the active learning method in this paper has the best effect.

D. CONDITION MONITORING OF GEARBOX FAULT

This section will calculate the output of the verification set on the established KNNR model, and set the threshold by analyzing the observed value and residual value of the predicted value of the output variable through SPC technology. Then

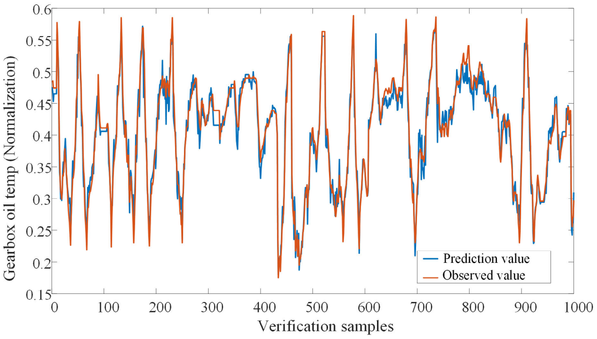

FIGURE 7. **Observed value and prediction value of the verification set.**

the testing samples is input into KNNR model to realize CM and fault early warning.

Fig. 7 shows the CM curve of the verification set. The blue line is the prediction value of verification set, and the red line is the observed value.

Fig. 7 shows the output curve of the KNNR model, but the fault degree of a single sample cannot be quantified based on only two curves. Therefore, it is necessary to calculate the residuals of observed and predicted values and set thresholds, so that the residual curve can be converted into a binary early warning.

SPC [29] technology refers to Process Control by means of mathematical statistics. It analyzes and evaluates the production process, timely finds signs of systematic factors according to feedback information, and takes measures to eliminate their effects, so as to maintain the process in a controlled state only affected by random factors, so as to achieve the purpose of quality control.

For a random variable $X\tilde{N}(\mu, \sigma^2)$, its probability of falling in $[\mu - 3\sigma, \mu + 3\sigma]$ is:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.99 \tag{17}$$

If X exceeds the range of Eq. 17 for a long time, it can be considered that the process is affected by abnormal factors and has faults.

In practical application, the mean $\bar{X}$ and standard deviation S of the residual sequence are used to replace the μ and σ of the normal distribution. The calculation formula is as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{18}$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{X})^2} \tag{19}$$

where n is the number of samples.

The upper control limit (UCL) and low control limit (LCL) is calculated as follows:

$$UCL = \bar{X} + 3S \tag{20}$$

$$LCL = \bar{X} - 3S \tag{21}$$

If the residual exceeds the control limit for a long time, it is considered that the gearbox has had a significant failure at this time. In this pare, $UCL = 0.0485$ and $LCL = -0.0453$.

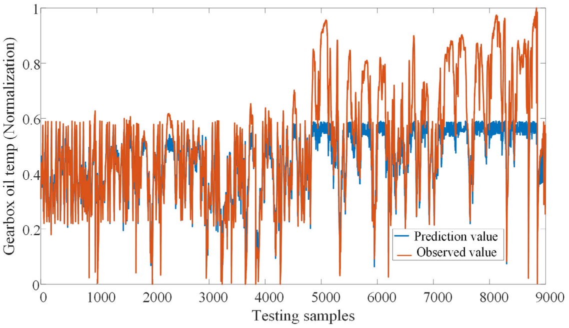

FIGURE 8. **Observed value and prediction value of the testing set.**

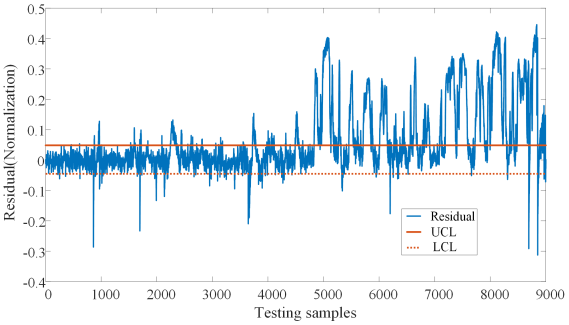

FIGURE 9. **Residual of the testing set.**

The CM results of the testing set on the KNNR model are shown in Figure 8. The blue line is the predicted value of the model, and the red line is the observed value.

As shown in Fig. 8, the observed values of about the first 1,000 samples are normal, and the oil temp is maintained within a certain range at this time, which is in a stable operating state. The oil temp change frequency of the 1,001st to 5,000th samples is slower than before, and the oil temp has reached a lower level many times, indicating that the equipment may be in an early deterioration stage at this time. After about 5,000 samples, the oil temp rises significantly and fluctuates violently, indicating that the equipment has had an obvious fault at this time.

The blue line in Fig. 9 represents the residual of the observed and predicted values, and the red line represents the control limit. About the first 1,000 samples basically have no overrun, and about the 1,000th to 5,000th samples have exceeded the limit for many times, but the range is not high. After the 5,000th sample, the residual error exceeds the control limit significantly and for a long time. In reality, the residual may exceed the limit due to the sensor jumping at a certain time. If the residual triggers an alarm every time it exceeds the limit, it may give a false alarm, so the sliding window method is used to deal with the residual error.

If the window length is $M = 1000$, the average value E_i of the residual sequence in ith window is:

$$E_i = \frac{1}{M}\sum_{j=i}^{i+M-1} e_j \tag{22}$$

where e_j is the jth residual of the ith window.

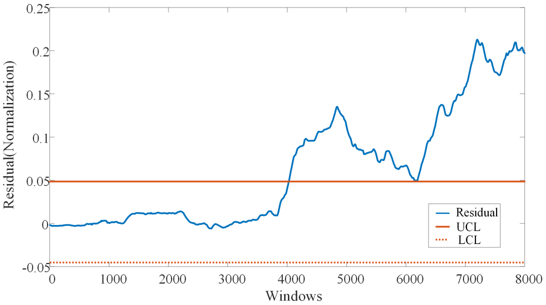

FIGURE 10. **CM result of the testing set.**

Fig. 10 shows the residual curve processed by the sliding window method. The previous curve did not exceed the limit. At the 4,024 window, the curve exceeded the threshold and was still rising. After that, the curve fluctuated greatly, but the overall trend was upward, indicating that the equipment had a serious fault at this time. According to relevant records, the sampling time of the 5,023th sample is 2017/10/28 21:45. Therefore, the method proposed in this paper can detect faults and trigger alarms about 20 days earlier than SCADA system.

V. CONCLUSION

In this paper, a WTCM method based on KNNR algorithm is proposed, and its effectiveness is proved by actual SCADA data. The DC coefficient is used to select modeling variables. In order to select high-quality samples to construct training set, an AL-based sample selection method considering multiple strategies is proposed. Based on the experimental results, we draw the following conclusions:

(1) For WT, a nonlinear object, DC coefficient has advantages over the commonly used Pearson correlation coefficient in feature selection, and the generalization accuracy of the model has been significantly improved.

(2) Compared with only considering a single AL strategy, the AL-based sample selection method considering two strategies can select higher quality training samples, and greatly reduce the size of the training set within the allowable range of reduced model generalization accuracy, which reduces the storage space and improves the operation efficiency.

(3) The KNNR model proposed in this paper can realize CM of the WT gearbox. Compared with the SCADA system, the method in this paper can trigger the fault alarm about 20 days earlier, which is helpful to find the early deterioration phenomenon. And the method in this paper is also applicable to other objects.

The disadvantage of this method is that the training set cannot be updated over time to adapt to the performance changes of WT. The next research direction will be to solve the common concept drift problem of WTCM.

REFERENCES

[1] A. I. Osman, L. Chen, M. Yang, G. Msigwa, M. Farghali, S. Fawzy, D. W. Rooney, and P.-S. Yap, "Cost, environmental impact, and resilience of renewable energy under a changing climate: A review," *Environ. Chem. Lett.*, Oct. 2022, doi: 10.1007/s10311-022-01532-8.

[2] Z. Luo, C. Liu, and S. Liu, "A novel fault prediction method of wind turbine gearbox based on pair-copula construction and BP neural network," *IEEE Access*, vol. 8, pp. 91924–91939, 2020.

[3] S. Ren, X. Feng, and M. Yang, "Solution of issues in energy theory caused by pathway tracking: Taking China's power generation system as an example," *Energy*, vol. 262, Jan. 2023, Art. no. 125596.

[4] S.-P. Breton and G. Moe, "Status, plans and technologies for offshore wind turbines in Europe and North America," *Renew. Energy*, vol. 34, no. 3, pp. 646–654, Mar. 2009.

[5] A. Novikau, "Current challenges and prospects of wind energy in Belarus," *Renew. Energy*, vol. 182, pp. 1049–1059, Jan. 2022.

[6] B. Snyder and M. J. Kaiser, "Ecological and economic cost-benefit analysis of offshore wind energy," *Renew. Energy*, vol. 34, no. 6, pp. 1567–1578, Jun. 2009.

[7] Q. Wei and D. Lu, "A survey on wind turbine condition monitoring and fault diagnosis—Part I: Components and subsystems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6546–6557, Oct. 2015.

[8] S. W. Sheng, "Wind turbine condition monitoring," *Wind energy*, vol. 17, no. 5, pp. 671–672, 2014.

[9] A. Wang, Z. Qian, Y. Pei, and B. Jing, "A de-ambiguous condition monitoring scheme for wind turbines using least squares generative adversarial networks," *Renew. Energy*, vol. 185, pp. 267–279, Feb. 2022.

[10] L. Cao, Z. Qian, H. Zareipour, Z. Huang, and F. Zhang, "Fault diagnosis of wind turbine gearbox based on deep bi-directional long short-term memory under time-varying non-stationary operating conditions," *IEEE Access*, vol. 7, pp. 155219–155228, 2019.

[11] Z. Liu, X. Wang, and L. Zhang, "Fault diagnosis of industrial wind turbine blade bearing using acoustic emission analysis," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6630–6639, Sep. 2020.

[12] W. Qiao and D. Lu, "A survey on wind turbine condition monitoring and fault diagnosis—Part II: Signals and signal processing methods," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6546–6557, Oct. 2015.

[13] S. Dey, P. Pisu, and B. Ayalew, "A comparative study of three fault diagnosis schemes for wind turbines," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 5, pp. 1853–1868, Sep. 2015.

[14] J. Maldonado-Correa, S. Martín-Martínez, E. Artigao, and E. Gómez-Lázaro, "Using SCADA data for wind turbine condition monitoring: A systematic literature review," *Energies*, vol. 13, no. 12, p. 3132, Jun. 2020.

[15] Z. Wang, C. Liu, and F. Yan, "Condition monitoring of wind turbine based on incremental learning and multivariate state estimation technique," *Renew. Energy*, vol. 184, pp. 343–360, Jan. 2022.

[16] 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.

[17] Y. Shi, Y. Liu, and X. Gao, "Study of wind turbine fault diagnosis and early warning based on SCADA data," *IEEE Access*, vol. 9, pp. 124600–124615, 2021.

[18] Z. Wang and C. Liu, "Wind turbine condition monitoring based on a novel multivariate state estimation technique," *Measurement*, vol. 168, Jan. 2021, Art. no. 108388.

[19] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.

[20] F.-M. Schleif, B. Hammer, and T. Villmann, "Margin-based active learning for LVQ networks," *Neurocomputing*, vol. 70, nos. 7–9, pp. 1215–1224, Mar. 2007.

[21] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.

[22] F. Ozdemir, Z. Peng, P. Fuernstahl, C. Tanner, and O. Goksel, "Active learning for segmentation based on Bayesian sample queries," *Knowl.-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106531.

[23] Y. Song, J. Liang, J. Lu, and X. Zhao, "An efficient instance selection algorithm for k nearest neighbor regression," *Neurocomputing*, vol. 251, pp. 26–34, Aug. 2017.

[24] C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka, "Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery," *Appl. Energy*, vol. 129, pp. 49–55, Sep. 2014.

[25] H. Zhang, H. Niu, Z. Ma, and S. Zhang, "Wind turbine condition monitoring based on bagging ensemble strategy and KNN algorithm," *IEEE Access*, vol. 10, pp. 93412–93420, 2022.

[26] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *Ann. Statist.*, vol. 35, no. 6, pp. 2769–2794, 2007.

[27] J. Tautz-Weinert and S. J. Watson, "Using SCADA data for wind turbine condition monitoring—A review," *IET Renew. Power Gener.*, vol. 11, no. 4, pp. 382–394, 2017.

[28] F. Douak, F. Melgani, and N. Benoudjit, "Kernel ridge regression with active learning for wind speed prediction," *Appl. Energy*, vol. 103, pp. 328–340, Mar. 2013.

[29] D. S. Holmes and A. E. Mergen, "Using SPC in conjunction with APC," *Qual. Eng.*, vol. 23, no. 4, pp. 360–364, Oct. 2011.

CHENGJIA BAO is currently working as a Senior Engineer at State Grid Gansu Electric Power Company, China.

His current research interests include automation of power systems, intelligent transmission, and transformation lines.

TIANYI ZHANG is currently working as an Engineer at State Grid Lanzhou Power Supply Company, China.

His current research interest includes intelligent inspection of transmission lines.

ZHIXI HU is currently working as a Deputy Senior Engineer at State Grid Lanzhou Electric Power Supply Company, China.

His current research interest includes power system automation.

WANGJING FENG is currently working as a Senior Engineer at State Grid Lanzhou Power Supply Company, China.

His current research interests include power marketing and marketing big data analysis.

RU LIU is currently working as a Senior Engineer at State Grid Lanzhou Electric Power Supply Company, China.

Her current research interests include power big data and artificial Intelligence.