

Wind Turbine Condition Monitoring Based on Bagging Ensemble Strategy and KNN Algorithm

HONGMIN ZHANG¹, HAIMING NIU¹, ZENGHUI MA², AND SHUYAO ZHANG³

¹Beijing Engineering Research Center of Power Station Automation, Chn Energy ZhiShen Control Technology Co., Ltd., Beijing 102200, China

²College of Ocean Information Engineering, Hainan Tropical Ocean University, Sanya 572022, China

³School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China

Corresponding author: Shuyao Zhang (120202127027@ncepu.edu.cn)

This work was supported in part by the Chn Energy Science and Technology Innovation Project “Research and Demonstration of Wind Turbine Control System”, Funding No.:YZ2021001.

ABSTRACT The gearbox is an important component of a wind turbine (WT). Once the gearbox is damaged, problems such as long-term maintenance and high maintenance costs will occur. Therefore, it is necessary to carry out on-line condition monitoring (CM) of WTs. Because a large amount of data is accumulated by the supervisory control and data acquisition (SCADA) system, CMs based on data-driven methods have been widely investigated. In this paper, a CM method that is based on the KNN regression method and bagging ensemble strategy is proposed. The proposed method is validated by SCADA data collected from a field WT. The results show that the ensemble model can achieve the desired estimation accuracy and improve the operation efficiency by approximately 30%.


INDEX TERMS Wind turbine gearbox, data-driven method, condition monitoring, KNN, bagging.

I. INTRODUCTION

To cope with global climate change, China has announced that it will achieve peak carbon dioxide emissions by 2030 and carbon neutrality by 2060 [1]. Therefore, clean energy power generation technology has broad development potential. As a kind of clean energy, wind energy has been widely utilized worldwide. According to the Global Wind Energy Council (GWEC) report [2], although the global newly installed capacity reached 93 GW in 2020, a large number of wind turbines (WTs) still need to be installed.

With an increase in the number of wind turbines in service and the extension of operation times, the possibility of component failure also increases. According to the statistical data [3], due to the harsh operating environment and other conditions, the gearbox is the component of WTs with a high incidence of faults. Once the gearbox is damaged, problems such as high maintenance costs, complex maintenance processes, and long maintenance times due to structural constraints will ensue [4]. Therefore, it is necessary to carry out online condition monitoring (CM) of gearboxes.

The CM of gearboxes is divided into vibration signal analysis [5], oil quality analysis [6] and supervisory control and data acquisition (SCADA) system data analysis [7]

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

according to different signal sources. However, vibration signal analysis requires the installation of professional sensors to collect high-frequency vibration data, resulting in additional expenses. Oil quality analysis is an invasive method that cannot realize online monitoring. Presently, almost all wind turbines are equipped with the SCADA system [8], which can collect a large amount of operational and record fault data. Therefore, WTCM based on SCADA data has been widely employed by scholars.

Since a large amount of data accumulates in the SCADA system, the data-driven method is an essential method in CM based on SCADA data [9]. Fu *et al.* [10] built a model based on deep learning to process the temperature of a gearbox using historical SCADA data. Jin *et al.* [11] mined health status-related information from SCADA data and established a Mahalanobis space as a reference space for wind turbine condition monitoring. Liu *et al.* [12] constructed a global monitoring statistic based on all temperature variables contained in the SCADA system to monitor the overall health status of the wind turbine. Zhang *et al.* [13] combined the random forest with extreme gradient boosting to establish a wind turbine fault detection framework. Luo *et al.* [14] proposed a SCADA data-based, online monitoring method based on a pair-copula and BP neural network. Dhiman *et al.* [15] applied a SVM to WT gearbox condition monitoring and analyzed the regression residual from a statistical point of

view. Dhiman *et al.* [16] proposed the highly reliable method of applying the TWSVM and adaptive threshold to WT gearbox anomaly detection.

The K-nearest neighbors (KNN) algorithm is a commonly used nonparametric method that was proposed by Cover and Hart in 1968 [17]. The KNN algorithm does not need to train a model in advance and is often selected to solve regression and classification problems. The operating environment and operating parameters of WTs are complex and changeable. The models trained by conventional parametric methods, such as neural networks and Bayesian methods, may have poor flexibility and encounter the problem of model mismatch. The KNN algorithm does not need to train the model in advance and only needs to update the training set in time to obtain the optimal estimation effect; thus, it is suitable for WTCM.

However, when the number of training samples is large, the KNN algorithm is bound to cause a vast amount of time overhead and to reduce the operation efficiency when calculating the Euclidean distance between the test sample and all training samples. Scholars have performed many studies on how to reduce the loss of operation accuracy while improving the computational efficiency of the KNN algorithm [18].

Ensemble learning constructs and combines multiple learners to complete learning tasks, which can often obtain generalization performance with significant advantages over a single learner [19]. Ensemble systems have proved to be very effective and extremely versatile in a broad spectrum of problem domains and real-world applications. Bagging is the most famous representative of the parallel ensemble learning strategy [20]. Bagging can be combined with almost any learning algorithm to form an ensemble learning system, such as a neural network [21] or decision tree [22].

To solve the problem of slow operation caused by an excessively large training set of KNNs, a condition monitoring method based on the bagging ensemble strategy and the KNN regression method is proposed in this paper. The training set is randomly sampled based on bagging to construct multiple KNN individual learners. With an increase in the number of individual learners, the ensemble system will become more complex, the estimation accuracy will increase, and the calculation time will remain at the same order of magnitude.

The SCADA data collected from a WT are used to validate the feasibility of the industrial application of the proposed approach. The results show that the proposed method can realize gearbox CM and provide health rate indicators.

The remainder of this paper is organized as follows: Section II presents the framework of the ensemble KNN method. Section III gives a detailed description of the KNN algorithm, bagging ensemble strategy and SPC technology. Section IV shows the results of experiments to validate the proposed method. The experimental results are summarized, and the conclusions are given in Section V.

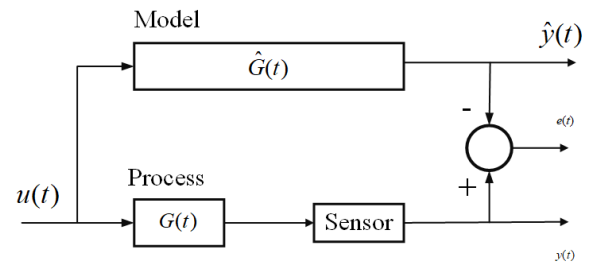


FIGURE 1. NBM monitoring with the input $u(t)$ for both the process $G(t)$ and its model $\hat{G}(t)$, their outputs $y(t)$ and $\hat{y}(t)$, respectively, and the final error or residual $e(t)$.

II. THE FRAMEWORK OF PROPOSED METHOD

A. NORMAL BEHAVIOR MODELING STRATEGY

Normal behavior modeling (NBM) is a condition monitoring method based on the SCADA data-driven method [23]. Because the collected field data often have no distinct label, we cannot use the classification algorithm to detect the specific fault of the equipment. The NBM method based on SCADA data can use the unlabeled operation data to monitor the statuses of the target variables in which we are interested. The framework of the NBM is shown in Fig. 1.

We will describe the NBM in detail. First, aiming at the target variable, a model is established by using the historical, collected, normal SCADA data, which is usually a data-driven algorithm model. Second, the real-time SCADA data are input into the model to obtain the estimated value of the target variable, which represents the real-time value if the equipment is in normal operation. Last, the residual between the estimated value and the observed value is calculated. The residual represents the current deviation of the target variable from the normal operation.

B. THE PROPOSED FRAMEWORK OF ENSEMBLE KNN METHOD

As shown in Fig. 2, the proposed WTCM, ensemble KNN method can be divided into three parts: data preprocessing, offline ensemble model establishment and online monitoring. The specific steps are described as follows:

(1) Data preprocessing: First, to obtain the normal available data, the missing data and abnormal data in the original data are deleted according to the relevant technical documents. Second, the relevant operating parameters are selected as the target variable and auxiliary variables. Among them, the target variable is the variable that we want to monitor, and the auxiliary variables are the operating parameters that are closely related to the target variable. Last, the normal data are determined and normalized.

(2) Offline ensemble model establishment: Since the KNN algorithm does not need to be trained in advance, it only needs to randomly sample the training set n times to establish the ensemble KNN learner. The threshold is calculated based on SPC technology and the verification set.

(3) Online monitoring: The real-time SCADA data are input as the testing sample into the ensemble model to

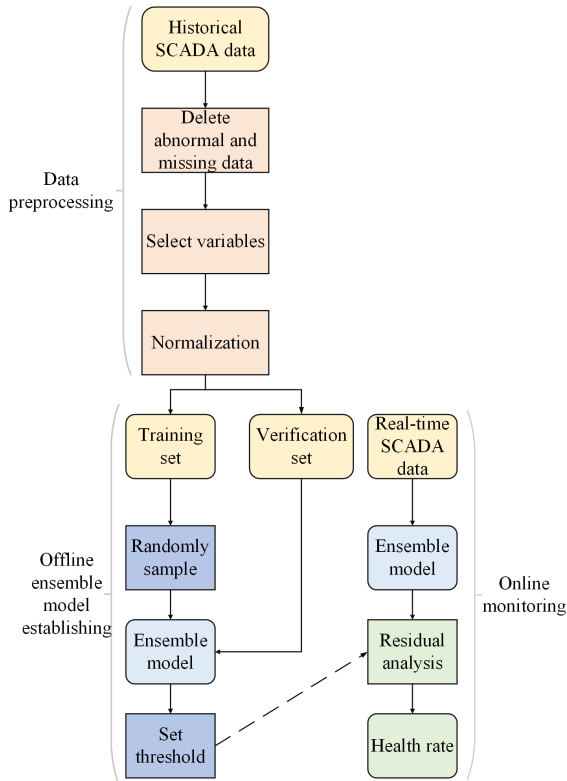


FIGURE 2. The framework of proposed method.

obtain the estimated value $\hat{y}(t)$ of the target variable and to calculate the residual $e(t)$ between $\hat{y}(t)$ and the observed value $y(t)$ of the testing sample. Next, $e(t)$ is compared with the threshold, and the health rate of the gearbox is calculated according to the sliding window method.

III. METHOD

A. KNN REGRESSION ALGORITHM FOR CONDITION MONITORING

The basic principle of the KNN regression algorithm is described as follows: when the target variable of the testing sample is unknown, obtain the K nearest neighbors of the testing sample, and take the average value of the target variable of the K nearest neighbors as the estimated value of the target variable of the testing sample. Therefore, the essence of the regression problem is the prediction problem. The difference is that in the condition monitoring, the target variables of the testing samples can be actually measured, so the distance measurement formula of the KNN regression algorithm can be improved. The calculation is detailed as follows:

For a testing sample $x = (x_1, x_2, \dots, x_m, y)^T$ (y is the target variable)

(1) Calculate the Euclidean distance between x and the training sample $x_j = (x_{1j}, x_{2j}, \dots, x_{mj}, y_j)^T$:

$$d_j(x, x_j) = \left(\sum_{l=1}^m |x_l - x_{lj}|^2 + |y - y_j|^2 \right)^{1/2} \quad (1)$$

where $d_j(x, x_j)$ is the Euclidean distance between x and x_j .

(2) Sort the Euclidean distance in descending order, identify K training samples closest to x and record them as $x_p^{(K)} = (x_{1p}^{(K)}, x_{2p}^{(K)}, \dots, x_{mp}^{(K)}, y_p^{(K)})^T, p \in [1, K]$. The estimated value of the testing sample's target variable is expressed as follows:

$$\hat{y} = \sum_{p=1}^K y_p^{(K)} / K \quad (2)$$

To further improve the accuracy of the algorithm, the "weighted method" can be employed instead of the "average method." In this paper, the nearest neighbor samples closer to the testing sample will be assigned a large weight, and the nearest neighbor samples further from the testing sample will be assigned a smaller weight. The formula of the weighted KNN (wKNN) is expressed as follows:

$$\hat{y} = \sum_{p=1}^K y_p^{(K)} \cdot w_p \quad (3)$$

$$w_p = D_{1+K-p}^{(K)} / D^{(K)} \left(\sum_{p=1}^K w_p = 1 \right) \quad (4)$$

where w_p is the weight of the p -th nearest neighbor, $D_{1+K-p}^{(K)}$ is the Euclidean distance between x and its $1 + K - p$ -th neighbor, and $D^{(K)}$ is the sum of the Euclidean distances between x and its neighbors.

According to the specific steps of the KNN algorithm, the time consumption of the KNN algorithm is only related to the size of the training set; the calculation time is linear with the size of the training set; its time complexity is $O(n)$; and n is the number of training samples.

B. KNN ALGORITHM WITH BAGGING STRATEGY

The general structure of ensemble learning is to generate multiple individual learners and then combine them through a selected strategy. The common strategies are boosting, bagging and stacking [24]. Boosting is a serial iterative structure, with a strong dependence among individual learners. Stacking is a hierarchical structure. To avoid overfitting, individual learners are required to be heterogeneous.

The bagging ensemble strategy is a parallel strategy. All homogeneous learners are interdependent. By randomly changing the distribution of the training sets, new training subsets are generated, and individual learners are trained. The steps of the KNN regression algorithm with the bagging strategy are presented as follows:

Let the number of individual learners be b , and let the number of training samples for each individual learner be n_b , generally $n_b \leq n$.

(1) The original training set is randomly sampled b times to form b KNN-based training subsets.

(2) A KNN individual learner is trained based on each training subset, and then these KNN learners are combined. When combining the individual learners, the simple voting

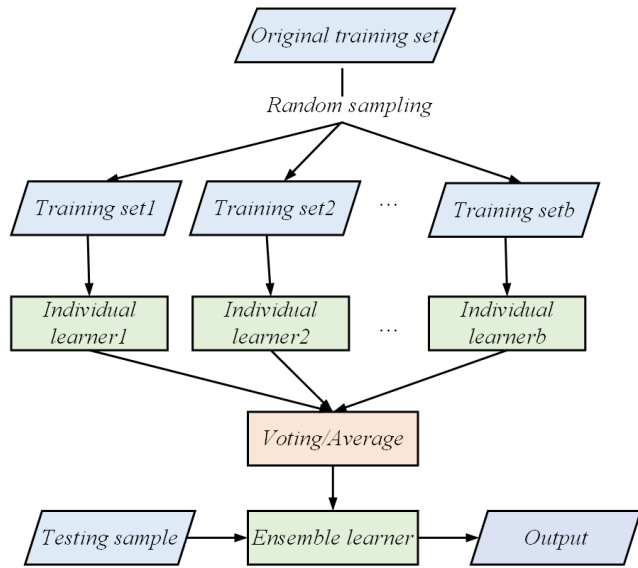


FIGURE 3. Bagging ensemble strategy diagram.

method is utilized for the classification task, and the simple average method is employed for the regression task.

As previously mentioned, the time complexity of the KNN learner is $O(n_b)$, so the time complexity of the bagging ensemble KNN algorithm is $b(O(n_b) + O(s))$. Generally, the voting/average complexity s is very small, and b is generally a small constant, so the complexity of the conventional KNN algorithm and bagging integrated KNN algorithm has the same order of magnitude.

C. THRESHOLD SETTING METHOD

The output of condition monitoring is often a continuous value, but we cannot judge whether the gearbox is faulty according to the continuous output, so we need to convert the condition monitoring result into a binary output by setting the threshold. If the residual exceeds the threshold, the gearbox will have a high probability of abnormality at this time. In this paper, statistical process control (SPC) is selected as the method to use to set the threshold.

SPC is a process monitoring method based on statistical theory that can be applied to involve the alarm threshold in WTCM [24]. Because the gearbox often shows overheating faults, this paper mainly discusses how to set the upper limit of the alarm. The detailed method is expressed as follows:

Assuming random variable $X \sim N(\mu, \sigma^2)$, according to the relevant theory of normal distribution, the probability of random variable X falling in the interval $-\infty, \mu + 1.645\sigma]$ and $-\infty, \mu + 1.282\sigma]$ are expressed as follows:

$$P(-\infty < X \leq \mu + 1.645\sigma) \approx 0.95 \quad (5)$$

$$P(-\infty < X \leq \mu + 1.282\sigma) \approx 0.90 \quad (6)$$

If the value of X continually exceeds the above range, the operation process will be affected by abnormal factors and fails. Therefore, the overheating and warming threshold can be designed according to the normal distribution of μ and σ .

In practical applications, the sample mean \bar{X} and sample standard deviation S are employed to replace μ and σ of the normal distribution. The formula is presented as follows:

$$\bar{X} = \frac{1}{n} \sum_{l=1}^n e_l \quad (7)$$

$$S = \left(\frac{1}{n-1} \sum_{l=1}^n (e_l - \bar{X})^2 \right)^{1/2} \quad (8)$$

where e_l is the residual of the observed value and estimated value and n is the number of testing samples.

The formula of the threshold can be defined as follows:

$$T_1 = \bar{X} + 1.282S \quad (9)$$

$$T_2 = \bar{X} + 1.645S \quad (10)$$

where T_1 is the first alarm threshold and T_2 is the second alarm threshold.

If the gearbox oil temp continuously exceeds the threshold, a significant failure of the gearbox will occur at this time. Hierarchical thresholds can realize hierarchical alarms in industrial applications and provide different fault tolerances for industrial applications.

IV. CASE ANALYSIS

A. DATA DESCRIPTION

The SCADA data used in this paper to verify the effectiveness of the proposed method in this paper are obtained from an onshore WT in Hebei Province, China. The main characteristic parameters of the WT are listed as follows: the related power is 2 MW; the related wind speed is 12 m/s; the cut-in wind speed is 4 m/s; the cut-off wind speed is 25 m/s and the SCADA system sample interval is 10 mins. The fault type is gearbox bearing overheating.

B. DATA PREPROCESSING

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 speeds less than the cut-in speed, and data with wind speeds greater than the cut-off wind speed. The samples with abnormal operating parameters are removed based on the Laida criterion; 14,000 samples remain.

There are dozens of characteristic parameters in SCADA system. Firstly, eight variables are roughly selected and their Pearson correlation coefficients with gearbox bearing temperature are calculated: wind speed, generator speed, ambient temp, active power, impeller speed, wind direction angle and reactive power. The range and Pearson correlation coefficient of selected variables is shown in Tab. 1.

It can be seen from table 1 that there is a strong positive correlation between bearing temperature and wind speed, generator speed and active power, which is in line with the normal characteristics that the increase of wind speed leads to the increase of rotating speed and load of impeller, generator and other equipment, and then the increase of gearbox bearing

TABLE 1. The range of selected variables.

Variable	Range	Correlation coefficient
Wind speed $/(m \cdot s^{-1})$	[4.0,19.31]	0.846
Generator speed $/(r \cdot \text{min}^{-1})$	[1001.2,1694.3]	0.928
Ambient temp $/^{\circ}\text{C}$	[4.0,31.0]	-0.214
Active power / kW	[54.8,2018.2]	0.829
Impeller speed $/(r \cdot \text{min}^{-1})$	[9.00,15.02]	0.927
Wind direction angle $/^{\circ}$	[13.5,338.0]	-0.151
Reactive power / var	[-1.54,3.92]	-0.093
Gearbox bearing temp $/^{\circ}\text{C}$	[43.65,69.00]	1

temperature. Wind direction angle is closely related to pitch and yaw system, reactive power and grid connection process. They have little impact on gearbox, so the correlation is poor. Although the correlation between bearing temperature and ambient temperature is general, when the power and wind speed are the same, the difference of ambient temperature will also lead to great differences in the operation state of the unit, which needs to be referred to in condition division and condition monitoring. Combined with the above analysis, it is determined that the auxiliary variables are wind speed, generator speed, impeller speed, ambient temperature and active power.

Normalize the samples to avoid dimensional influence. The formula is presented as follows:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (11)$$

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

We select No. 1-6,000 samples as the training samples to establish the ensemble model and designate No. 6,001-7,000 samples as the verification samples to select various parameters of the ensemble model and to design the alarm threshold. Samples No. 6,001-14,000 are taken as testing samples to monitor the condition of the gearbox.

C. PERFORMANCE ANALYSIS OF MODEL

This section will compare the performance of the conventional KNN model and ensemble model. The running environment of the program is MATLAB R2019a; the CPU model is Intel i7-10710U; and the RAM is 16G.

The performance of the conventional KNN model is analyzed, and the influence of the number of training samples on the conventional KNN model is discussed ($K = 10$).

Starting from the last 500 samples as the training set, expand the training set by increasing 500 samples each time from back to front according to the chronological relationship; construct the training set of conventional KNN; and test based on the verification set. In this paper, the root mean square error (RMSE) is utilized as the error function to evaluate the estimation accuracy of the model; its formula is

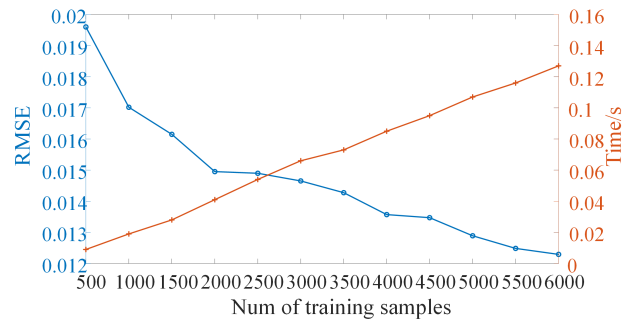


FIGURE 4. The performance of conventional KNN model.

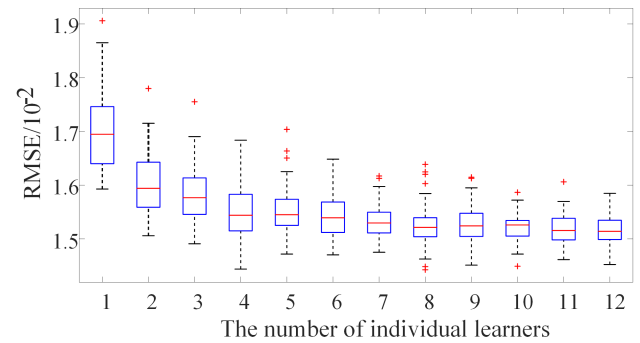


FIGURE 5. The performance of ensemble KNN model.

shown as follows:

$$RMSE = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

where y_i is the observed value of the target variable and \hat{y}_i is the estimated value of the target variable. The RMSE and operation time are shown in Fig. 4.

First, the above figure is analyzed in terms of estimation accuracy. With an increase in the number of training samples, the RMSE shows a downward trend. When the number of training samples is less than 2,000, the decline rate of the RMSE increases. When the number of training samples exceeds 2,000, the decline rate of the RMSE decreases. From the aspect of running time, although the time fluctuates slightly during the running process, it increases linearly with an increase in the number of training samples, which is consistent with the previously mentioned law of the KNN algorithm.

Second, the performance of the ensemble KNN model is investigated. Set the number of training samples of individual learners to 500, set the number of individual learners to 1-12 and establish 12 ensemble KNN models. Because the bagging ensemble model exhibits randomness in sampling, 50 repeated experiments are carried out for each model. The RMSE obtained from the experimental results is shown in the box diagram of Fig. 5.

As shown in Fig. 5, with an increase in the number of individual learners, the overall box of the ensemble KNN model shows a tightening trend, indicating that the larger

TABLE 2. Parameters of conventional KNN model and ensemble KNN model.

		500	1,000	1,500	2,000	2,500	3,000	
Conventional	Num of training samples	500	1,000	1,500	2,000	2,500	3,000	
	RMSE/10 ⁻²	1.96	1.70	1.61	1.52	1.49	1.47	
	t ₁ /s	0.009	0.019	0.028	0.041	0.054	0.066	
Ensemble	Num of individual learner	1	2	3	4	5	6	
	RMSE/10 ⁻²	Q3	1.74	1.64	1.61	1.58	1.57	1.56
		Q2	1.69	1.59	1.56	1.54	1.54	1.53
		Q1	1.64	1.55	1.54	1.51	1.52	1.51
	t ₂ /s	0.007	0.018	0.027	0.039	0.051	0.062	

is the number of individual learners, the more complex the ensemble model, and the better its stability.

Simultaneously, the median (Q2) of the RMSE tends to decrease with an increase in the number of individual learners. When the number of individual learners is less than 4, Q2 decreases at a faster rate. When the number of individual learners is greater than 4, Q2 fluctuates, but the change tends to be flat as a whole.

We compare the performance of the conventional KNN model and ensemble KNN model. The estimation accuracy and calculation time of the conventional KNN model and ensemble KNN model under different training set scales are shown in the following table. In the table, t₁ of the conventional KNN model is the time required to calculate 1,000 verification samples; t₂ of the integrated KNN model is the average time of 50 experiments; and Q3, Q2 and Q1 are the upper quartile, median and lower quartile, respectively, of the RMSE of the ensemble KNN model in 50 repeated experiments.

As shown in Tab. 2, with an increase in the number of individual learners, the training time of the ensemble KNN model shows a linear upward trend. When there are 3 individual learners, Q3 is 1.56 × 10⁻², and the average training time is 0.027 s. Compared with 2,000 training samples of the conventional KNN model, the estimation accuracy decreases by 2.6% and the training time increases by 34.14%, indicating that the ensemble KNN model has advantages with regard to training time.

Next, we will use the proposed KNN method, conventional KNN method, CNN and LSTM to perform an experiment on the verification set and to compare the program running time and estimation accuracy. The specific parameter settings of the above methods are listed as follows (the training sample of conventional KNN, CNN and LSTM is the last 1500 samples in the training set):

(1) CNN: The size of the 2DCNN convolution kernel is set to 3 × 3; the number is set to 16; the ReLU activation function is applied the number of epochs is set to 20; and the learning rate is 0.005.

(2) LSTM: The input layer time step of LSTM is set to 6; the output is set to 6; the sigmoid activation function is employed; the learning rate is set to 0.005; epochs are set to 20; the loss function is the average absolute error MAE;

TABLE 3. Comparative experimental results.

Method	Ensemble KNN	Conventional KNN	CNN	LSTM	
Num of training samples	3*500	1500	1500	1500	
Training time/s	-	-	26.69	28.71	
Estimation Time/s	0.027 (Average)	0.028	1.21	1.03	
RMSE/10 ⁻²	Q3	1.61			
	Q2	1.56	1.61	3.12	1.59
	Q1	1.54			

and the Adam optimizer is utilized to update the network weight.

As shown in Tab. 3, when the total number of training samples is constant, the time consumed by the conventional KNN and ensemble KNN is basically constant, the CNN and LSTM consumed a substantial amount of time in the training process, and the estimation time is longer than that of the two KNN methods. In terms of estimation accuracy, the ensemble KNN has advantages over the other three methods.

D. CONDITION MONITORING OF GEARBOX

This section will monitor the condition of the gearbox using the previously verified ensemble model and 8,000 testing samples. The number of individual learners is 3, and the training sample of each individual learner is 500.

First, the alarm threshold is set according to the verification set. The \bar{X} of the residuals of the ensemble model on the verification set is -0.0023, and S is 0.0154. According to (9) and (10), the first alarm threshold is 0.0176, and the second alarm threshold is 0.0232. Fig. 6 shows the condition monitoring results and two-level alarm threshold of the testing samples. It is worth mentioning that the difference between the two-level thresholds in this paper is small, which can explain why the estimation accuracy of the method proposed in this paper is high.

The figure shows that the residuals of the first 2,500 samples do not continually or greatly exceed the threshold,

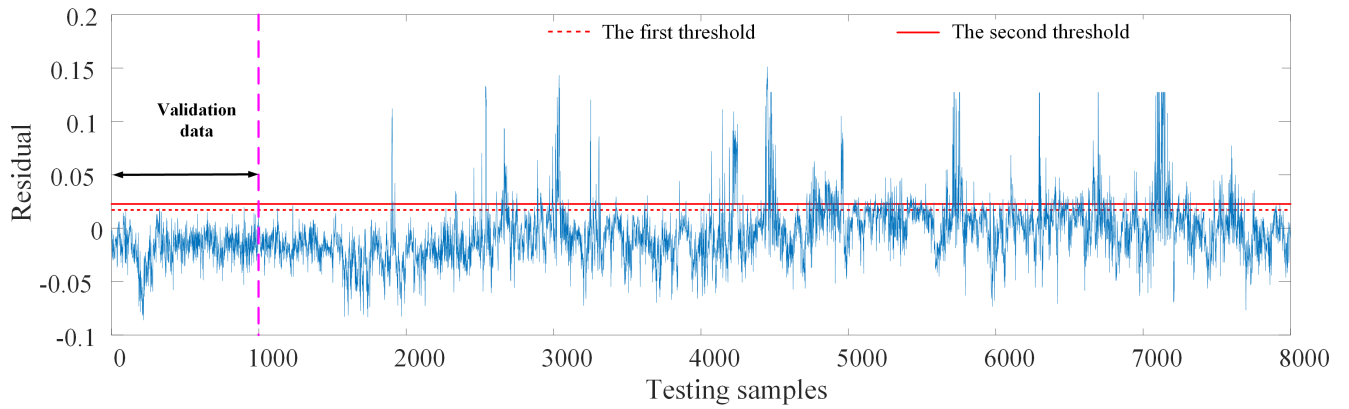


FIGURE 6. The residual and two-level alarm threshold.

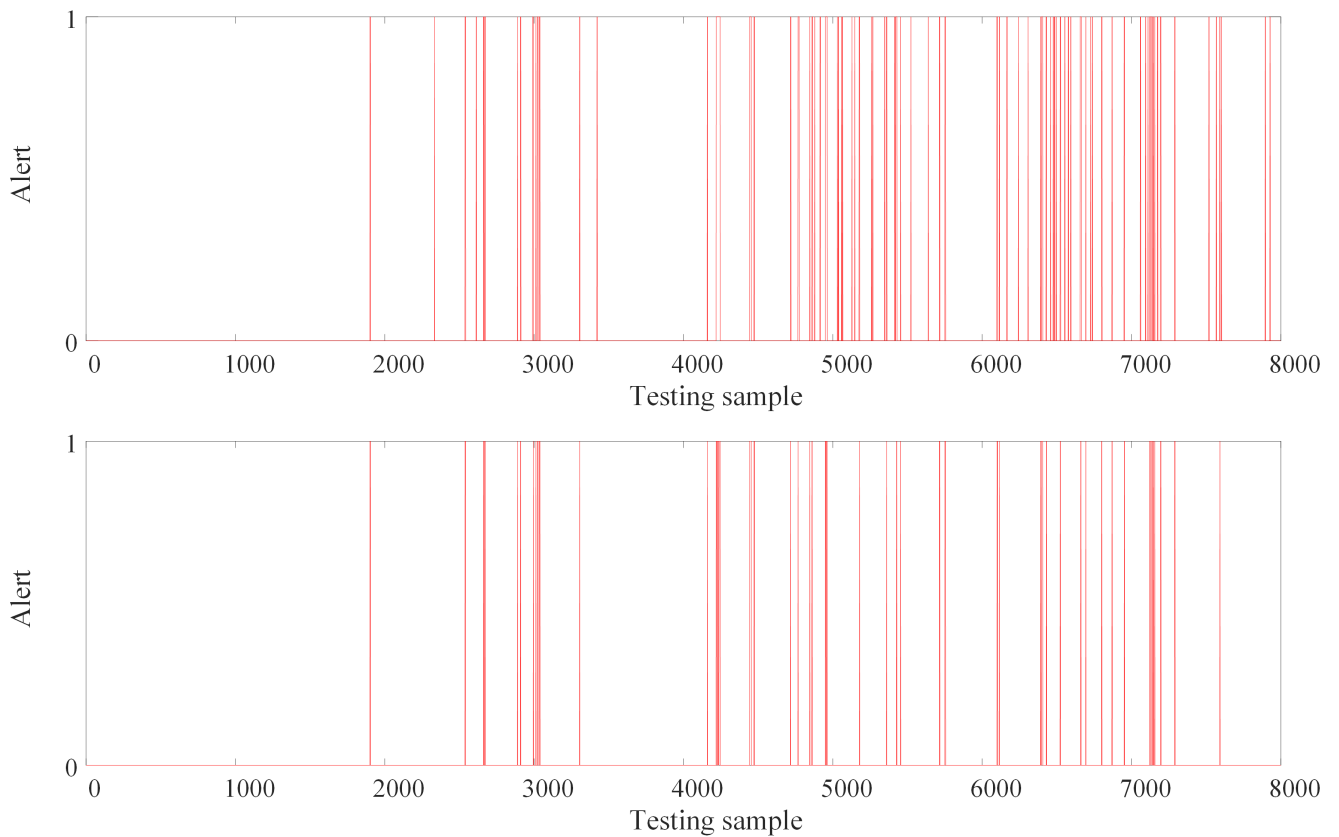


FIGURE 7. Real-time alert. (a) First alarm threshold. (b) Second alarm threshold.

indicating that the gearbox is still in normal operation at this time. After approximately 2,500 points, the residual of the sample significantly and continuously exceeds the threshold value, indicating that the gearbox state is abnormal at this time.

Due to the harsh environment or sensor failure, the residual may exceed the alarm threshold under the normal condition of the gearbox. To avoid false alarms, we stipulate that when 4 consecutive intervals (40 mins) are detected to exceed the

threshold, an alarm will be triggered. The alarm condition is shown in Fig. 7.

However, the residual sequence has the characteristics of frequent fluctuation, which is not conducive to the intuitive judgment of the operation state of the gearbox. Therefore, we define a health rate based on the sliding window method to improve the condition monitoring process of the gearbox.

If the length of the sliding window is M and the number of samples below the alarm threshold in one window is N , then

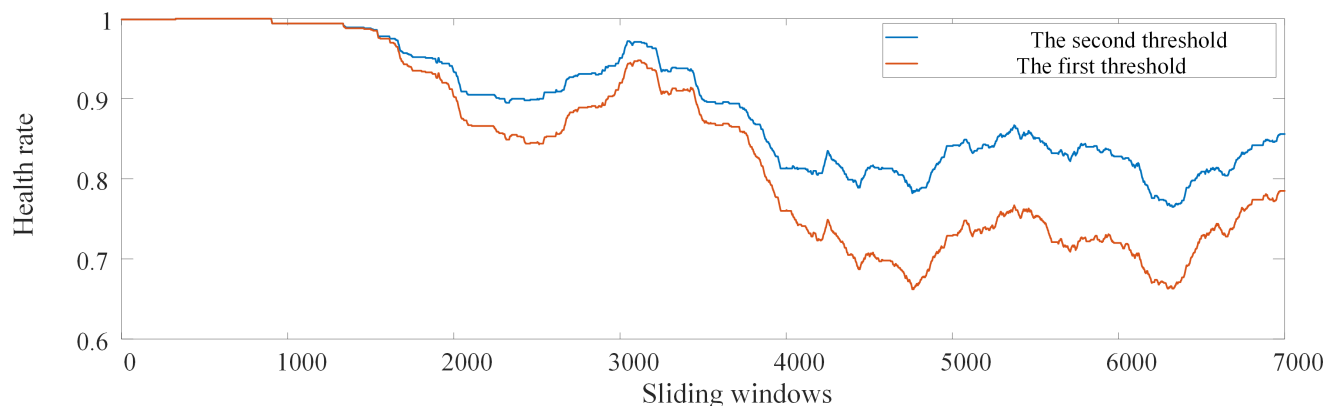


FIGURE 8. The health rate of gearbox.

for the first alarm threshold T_1 and second alarm threshold T_2 , the first health rate R_1 and second health rate R_2 can be defined as:

$$R_1 = \frac{N_1}{M} \quad (13)$$

$$R_2 = \frac{N_2}{M} \quad (14)$$

In this section, $M = 1,000$. The health rate of the gearbox is shown in Fig. 8.

The health rate curve shows that at approximately the first 1,500 windows, the health rate of the gearbox remains stable and high, and the two-level health rates remain approximately 100%. After the 1,500th window, the health rate began to decline, and at approximately the 2,500th window, the health rate fell to the first trough. After the health rate fluctuates, it reaches the lowest value after the 4,500th sliding window. The first health rate is below 70%, and the minimum second health rate is approximately 80%, indicating that the operation state of the gearbox is unstable and has failed. Thus, the health rate of the gearbox is always low. The health rate curve can provide a more intuitive operation status of the gearbox, reduce false alarms, and provide an important reference for staff.

V. CONCLUSION

In this paper, we propose a gearbox condition monitoring method based on the KNN algorithm and ensemble strategy. The method is validated by SCADA data collected from a WT. With the results of the experiments, we can safely reach the following conclusions:

(1) With an increase in the number of training samples, the estimation accuracy of the KNN algorithm is improved, but when it is expanded to a certain scale, its estimation accuracy is improved slowly.

(2) With an increase in the number of individual learners, the estimation accuracy of the ensemble learning model is improved, and the time only increases linearly. When the estimation accuracy is similar, the integrated learning model may have higher operational efficiency.

(3) The ensemble KNN model combined with SPC technology can realize the condition monitoring of WT gearboxes, and the health rate calculation method based on the sliding window method is more suitable for the actual site.

REFERENCES

- [1] C. Zou, B. Xiong, H. Xue, D. Zheng, Z. Ge, Y. Wang, L. Jiang, S. Pan, and S. Wu, "The role of new energy in carbon neutral," *Petroleum Explor. Develop.*, vol. 48, no. 2, pp. 480–491, Apr. 2021.
- [2] *Global Wind Report 2021: Wind Energy's Role on the Road to Net Zero*, Global Wind Energy Council. Accessed: Jan. 20, 2021. [Online]. Available: <https://gwec.net/global-offshore-wind-report-2021/>
- [3] L. Jiang, D. Xiang, Y. F. Tan, Y. H. Nie, H. J. Cao, Y. Z. Wei, D. Zeng, Y. H. Shen, and G. Shen, "Analysis of wind turbine Gearbox's environmental impact considering its reliability," *J. Cleaner Prod.*, vol. 180, pp. 846–857, Apr. 2018.
- [4] J. Zhou, S. Roshanmanesh, F. Hayati, V. J. Junior, T. Wang, S. Hajiabady, X. Y. Li, H. Basoalto, H. Dong, and M. Papaalias, "Improving the reliability of industrial multi-MW wind turbines," *Insight-Non-Destructive Test. Condition Monitor.*, vol. 59, no. 4, pp. 189–195, Apr. 2017.
- [5] S. Koukoura, J. Carroll, A. McDonald, and S. Weiss, "Comparison of wind turbine gearbox vibration analysis algorithms based on feature extraction and classification," *IET Renew. Power Gener.*, vol. 13, no. 14, pp. 2549–2557, 2019.
- [6] S. Sheng, "Monitoring of wind turbine gearbox condition through oil and wear debris analysis: A full-scale testing perspective," *Tribol. Trans.*, vol. 59, no. 1, pp. 149–162, 2016.
- [7] A. Heydari, D. A. Garcia, A. Fekih, F. Keynia, L. B. Tjernberg, and L. De Santoli, "A hybrid intelligent model for the condition monitoring and diagnostics of wind turbines gearbox," *IEEE Access*, vol. 9, pp. 89878–89890, 2021.
- [8] J. P. Salameh, S. Cauet, E. Etien, and A. Sakout, "Gearbox condition monitoring in wind turbines: A review," *Mech. Syst. Signal Process.*, vol. 111, pp. 251–264, Oct. 2018.
- [9] 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.
- [10] J. Fu, J. Chu, P. Guo, and Z. Chen, "Condition monitoring of wind turbine gearbox bearing based on deep learning model," *IEEE Access*, vol. 7, pp. 57078–57087, 2019.
- [11] X. Jin, Z. Xu, and W. Qiao, "Condition monitoring of wind turbine generators using SCADA data analysis," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 202–210, Jan. 2021.
- [12] X. Liu, J. Du, and Z. S. Ye, "A condition monitoring and fault isolation system for wind turbine based on SCADA data," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 986–995, Feb. 2022.
- [13] 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.

- [14] 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.
- [15] H. S. Dhiman, D. Deb, J. Carroll, V. Muresan, and M.-L. Unguresan, "Wind turbine gearbox condition monitoring based on class of support vector regression models and residual analysis," *Sensors*, vol. 20, no. 23, p. 6742, Nov. 2020.
- [16] H. Dhiman, D. Deb, S. M. Muyeen, and I. Kamwa, "Wind turbine gearbox anomaly detection based on adaptive threshold and twin support vector machines," *IEEE Trans. Energy Convers.*, vol. 36, no. 4, pp. 3462–3469, Dec. 2021.
- [17] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
- [18] C. J. Ma and Z. S. Ding, "Improvement of k-nearest neighbor algorithm based on double filtering," in *Proc. 5th Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Dec. 2020, pp. 1567–1570.
- [19] X. Dong, Z. Yu, and W. Cao, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, pp. 241–258, Apr. 2020.
- [20] Breiman, "Bagging predictors, machine learning research: Four current directions," *AIM Mag.*, vol. 6, no. 18, pp. 97–136, 1997.
- [21] M. Park, S. Lee, S. Hwang, and D. Kim, "Additive ensemble neural networks," *IEEE Access*, vol. 8, pp. 113192–113199, 2020.
- [22] Z.-H. Zhou and Y. Jiang, "NeC4.5: Neural ensemble based C4.5," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 6, pp. 770–773, Jun. 2004.
- [23] T. Xiao, J. Zhu, and T. Liu, "Bagging and boosting statistical machine translation systems," *Artif. Intell.*, vol. 195, pp. 496–527, Feb. 2013.
- [24] D. S. Holmes and A. E. Mergen, "Using SPC in conjunction with APC," *Qual. Eng.*, vol. 23, no. 4, pp. 360–364, Oct. 2011.



HONGMIN ZHANG received the B.E. degree from the Beijing University of Chemical Technology, Beijing, China, in 2018.

She is currently working as an Engineer with the Research and Development Center, Guoneng Zhishen Control Technology Company Ltd., China. Her current research interests include wind turbine condition monitoring and fault prediction.



HAIMING NIU received the B.E. degree in electrical automation from the Northeast China Institute of Electric Power, Heilongjiang, China, in 2003, and the B.S. degree in electrical automation from China Electric Power Research Institute, Beijing, China, in 2013.

He is currently working as an Engineer with Guoneng Zhishen Control Technology Company Ltd., China. His current research interests include wind turbine condition monitoring and fault prediction.



ZENGHUI MA received the Ph.D. degree in control theory and control engineering from North China Electric Power University, Beijing, China, in 2015.

He is currently a Senior Engineer with Hainan Tropical Ocean University, Sanya, China. His current research interests include thermal intelligent control and wind turbine fault prediction.



SHUYAO ZHANG received the B.E. degree in automation from North China Electric Power University, Baoding, China, in 2019. She is currently pursuing the Ph.D. degree in control theory and control engineering with North China Electric Power University, Beijing, China.

...