BioGecko

Vol 12 Issue 03 2023
ISSN NO: 2230-5807

Fault Detection in Wind Turbines using K-Nearest Neighbor Regression and K-Means Clustering: An Analysis of Rotor Bearing Temperature Data

Gautham Chakrawarthy A K[1], Vinoth T[2], Shubham Dadaso Patil[3], Bhukya Ramdas[4]
Veena Sharma[5]
[1,3,5] Dept. of Electrical Engineering NIT, Hamirpur, Hamirpur, HP, India
[2,4] Measurement & Testing Division National Institute of Wind Energy, Chennai, India
Email: [1]akgc5087@gmail.com, [2]tvinoth.official@gmail.com, [3]shubhampatil6399@gmail.com
[4]bhukyaramdas@niwe.res.in, [5]veena@nith.ac.in

ABSTRACT

Faults in wind turbines can lead to reduced energy generation and increased maintenance costs. This research focuses on detecting faults in wind turbines by analyzing the rotor bearing temperature using machine learning algorithms such as K-nearest neighbor (KNN) regression and K-means clustering. SCADA dataset with 15 features was selected, pre-processed, and used to evaluate the KNN regression using various metrics. The results obtained depict a high accuracy of 98.002% with a mean absolute error of 0.9300 and a mean squared error of 1.7776, while the residuals plot indicated a normal distribution with most of the points near zero, which suggests that the model is suitable for predicting rotor bearing temperature. The KNN regression results were used to identify potential gearbox faults through K-means clustering, with a silhouette score of 0.6221. This study demonstrates the potential use of KNN regression and K-means clustering for detecting faults in wind turbine gearbox by analyzing rotor bearing temperature data.

Keywords: Wind Turbines, Faults, Rotor Bearing Temper- ature, K-Nearest Neighbor (KNN) Regression, K-Means Cluster- ing, Fault Prediction

INTRODUCTION

The increasing use of wind energy as a clean and sustainable source of electricity has led to the growth of the wind energy industry. Offshore wind turbines, in particular, are becoming an important source of renewable energy with a projected global capacity of over 1 TW by 2030[1]. However, the harsh marine environment in which these turbines operate can result in increased wear and tear which further results in a higher maintenance cost along with various other risks. In such a situation it becomes essential to predict faults early to ensure the safe and efficient operation of offshore wind turbines.

One of the key areas of research in this field is the use of condition monitoring and diagnosis systems, with the help of AI-based techniques[2-4], this method has the potential to reduce the operation and maintenance costs of offshore wind

farms by up to 50%. The gearbox, being a critical component of a wind turbine, plays a crucial role in its overall efficiency and reliability. Early detection of faults in the gearbox can thereby ensure the reliable operation of wind turbines [7].

SCADA (Supervisory Control and Data Acquisition) sys- tems, which use sensors and other devices to collect data from various points within the process, including the gearbox, and transmit it to a central computer for analysis and control. The use of SCADA data for fault detection in wind turbines has the potential to improve their reliability and efficiency by enabling earlier and more accurate fault detection. This paper provides an overview of the use of SCADA data for wind turbine gearbox fault detection using rotor bearing temperature and the potential benefits of this approach.

In this study, 10 minute averaged SCADA data containing data collected from 15 wind turbines of capacity 2 MW each has been utilized to train a generic machine learning model to predict the rotor bearing temperature, thereby predicting gearbox faults.

LITERATURE REVIEW

Gearbox failures in wind turbines can result in significant downtime and repair costs for wind turbine operators [8]. To minimize these costs, it is important to predict gearbox faults as early as possible. Rotor bearing temperature is a direct and accurate measurement of gearbox health and can provide valuable information about the location and nature of faults within the gearbox [11].

Several data collection approaches like SCADA data, vi- bration analysis, and thermal imaging cameras have been developed to predict gearbox faults in wind turbines. SCADA data provides information on bearing temperature, oil temper- ature, vibration levels, and other sensor readings [5][11][27].

Vibration analysis involves collecting data on the vibration levels of the gearbox and other components [10][12], while thermal imaging cameras use infrared technology to measure the temperature of the bearings and other components in the turbine [28].

After the data is collected and finalized, different techniques like statistical analysis and machine learning analysis can be used to predict gearbox faults. Some common machine learning algorithms used for prediction of gearbox oil temper- ature from SCADA data are Deep Learning [15], Multilayer Feedforward Networks [16], Deep Neural Networks [17], Neu- ral Networks [18], Artificial Neural Networks [19], Support Vector Machines [20][21], Synthetic Minority Over-Sampling Technique [22], Manifold Learning and Shannon Wavelet Support Vector Machine [23], Multiscale Convolutional Neural Networks [24], Functional Trees Algorithm [25].

Machine learning algorithms have been shown to be effec- tive methods for predicting gearbox faults. The performance of different methods for predicting gearbox faults can vary depending on the quality and quantity of the data being analyzed, the complexity of the fault being predicted, and the accuracy and reliability of the analysis method being used. The use of rotor bearing temperature as a factor in detecting wind turbine gearbox faults is significant because it provides a clearer indication of the gearbox's overall performance. Unlike oil temperature, which can be affected by a variety of external factors, rotor bearing temperature is directly linked to the gearbox's functionality [30]

By utilizing rotor bearing temperature as a key factor in detecting wind turbine gearbox faults, a clearer understanding of gearbox health is obtained. To ensure accurate and reli- able predictions, a thorough and well-rounded approach that includes gathering quality data, utilizing strong analysis meth- ods, and rigorously testing and validating prediction models is necessary.

METHODOLOGY

The flowchart in figure 1 below provides an overview of the procedures followed in this research.

A. Dataset Selection

Dataset selected for this study contains 15 features of data along with a target (Rotor bearing temperature) and with 909604 data samples each. It is necessary to mention that the data contained in the dataset are 10 minutes averaged data. The parameters contained in the dataset along with its units is described in table I.

B. Data Pre-Processing

Data pre-processing is an important step in the analysis of data which depends on the specific data sources and the intended analysis, pre-processing steps used for this study has been mentioned below:

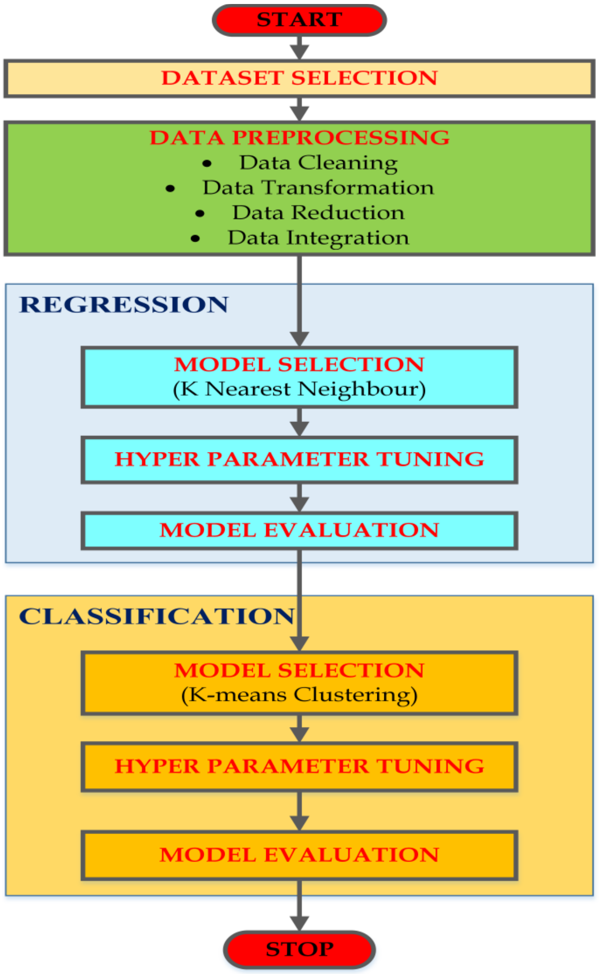

Fig. 1. Flow Chart of Proposed Approach

1) ***Data Cleaning****:* Data cleaning involves identifying and correcting errors or inconsistencies in the data, such as missing values, duplicates, out of range, or incorrect data types. It is important to ensure that the data is complete and accurate before proceeding with the analysis.

2) ***Data Transformation****:* Data transformation involves ap- plying various transformations to the data, such as scaling, normalization, or aggregation, in order to make the data more suitable for the intended analysis. It is necessary to scale the data to a common range in order to compare different features or to normalize the data to account for different units or scales.

3) ***Data Reduction****:* Data reduction involves selecting a subset of the data that is relevant to the analysis which helps to reduce the complexity of the data and improve the efficiency of the analysis.

4) ***Data Integration****:* Data integration involves combining data from multiple sources or sources with different formats, such as merging data from multiple SCADA systems or inte- grating data from different sensors. This can help to provide a more comprehensive view of the data and enable more advanced analyses.

TABLE I:DATASET DESCRIPTION

Sr.No	Parameter	Unit
1	Time Stamp	Sec
2	Active Power calculated by converter	KW
3	Active Power Raw	KW
4	Ambient Temperature	Degree Celsius
5	Generator Speed	RPM
6	Generator Winding Temp Max	Degree Celsius
7	Grid Power 10min Average	KW
8	nc1 inside Temp (Temperature inside Na- celle)	Degree Celsius
9	Nacelle Temp (Temperature outside Na- celle)	Degree Celsius
10	Reactive Power calculated by converter (Secondary Power generated by wind tur- bines at output source)	KVAR
11	Reactive Power (Secondary Power gener- ated by wind turbines at input source)	KVAR
12	Wind Direction Raw	Degree
13	Wind Speed Raw	Meter/Second
14	Wind Speed Turbulence	Meter/Second
15	Turbine ID	NA
16	Rotor Bearing Temperature	Degree Celsius

Overall, data pre-processing is important to carefully con- sider the pre-processing steps that are appropriate for the intended analysis in order to obtain accurate and meaningful results.

C. *Model Selection*

Model selection is the process of choosing best model for a particular prediction task. Generally, model selection will be done by training different algorithms or configurations and then validating its performance using metrics like accuracy, precision or recall. Model selected for our work is K-nearest neighbour (KNN) regression and K-means clustering.

D. *Regression*

1) ***K-nearest neighbour (KNN) regression:*** KNN regres- sion is a simple and effective non-parametric approach which works by identifying the K nearest data points to a given input point and using the mean or median of those points as the prediction. KNN regression will be more beneficial while working with complex or heterogeneous data like wind turbine SCADA data as it does not make assumptions about the underlying data distribution. Also, KNN regression is relatively easy to implement and does not require extensive pre-processing or feature engineering when working with large and noisy datasets, making it best tool for identifying potential issues and optimizing maintenance schedules in a wind turbine.

2) ***Hyper parameter tuning***: Hyper-parameter tuning is the process of adjusting the hyper-parameters of a machine- learning model to improve the model performance and to ensure the efficiency and accuracy of the prediction. KNN regression has variety of hyper-parameters like number ofneighbours (K), distance metrics, weighting function and methods to tune hyper-parameters like grid search.

Number of neighbours (K) is a hyper-parameter which determines the number of nearest points to the query point are used to make a prediction. Greater value of K makes the model smoother but increase the risk of over fitting. Similarly, lesser value of K makes the model more complex and increase the risk of under fitting. So it is necessary to tune the value of K to an appropriate value to ensure the best results out of the KNN model. The apt value for K was chosen by plotting an elbow curve. The idea is to pick the lowest point in the curve, the point on the graph where we see an elbow shape, as the optimal value of K. Distance metrics is a hyper-parameter which determines the distance between points. Weighting function is a hyper-parameter used to assign weights to the neighbours, and any parameters associated with the weighting function.

3) ***Model Evaluation***: Model evaluation is the process of measuring the performance of a machine learning model on a set of test data. This is typically done by comparing the predicted output of the model to the true output, and computing metrics such as Accuracy, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Median Ab- solute Error, Mean Squared Logarithmic Error, Mean Absolute Percentage Error. These metrics help to determine the effec- tiveness of the model and identify areas where it may need to be improved. Another way of model evaluation can be done by using histogram of residuals, which is a plot that shows the distribution of the residuals of a machine learning model. The residuals are the differences between the observed values and the predicted values of the model. The plot is used to assess whether the residuals are approximately normally distributed, which is an assumption of many machine learning models. If the residuals are normally distributed, the histogram should be roughly bell-shaped, with most of the residuals near zero and fewer residuals farther away from zero. If the residuals are not normally distributed, the histogram may have a different shape, such as a positive or negative skew, or multiple peaks.

E. *Classification*

1) ***K-means Clustering***: K-means clustering is a widely used unsupervised learning algorithm for partitioning a set of data points into K clusters based on their proximity to the K centroids. The algorithm works by randomly selecting K initial centroids, and then iteratively updating the centroids and re-assigning the data points to their closest centroid until convergence. The objective of the K-means clustering is to minimize the sum of squared distances between each data point and its assigned centroid. K-means clustering is useful for SCADA data because it allows the data to be grouped into similar clusters based on their proximity to the centroids. These clusters can then be analyzed to identify correlations between variables and identify areas for improvement in the system. Additionally, the clustering results can be used for predictive maintenance and to detect anomalies or outliers in the data.

2) ***Hyper parameter tuning***: The main hyper-parameter in K-means clustering is the number of clusters (K) to be formed. The value of K determines the number of centroids and, therefore, the number of clusters in the data. Silhouette analysis, is one method used to determine the appropriate value of K, where we study the separation distance between the resulting clusters, to measure how close each point in one cluster is to points in the neighbouring clusters. Basically, we measure how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Higher the silhouette value more well matched the object is to its own cluster and poorly matched to its neighbouring cluster.

3) ***Model evaluation***: Model evaluation for K-means clus- tering can be performed using the metrics specified in KNN regression’s model evaluation and also by using silhouette score. Additionally, visualizing the clusters using scatter plots can also provide insights into the quality of the clustering results.

RESULTS AND DISCUSSION

In this study, a dataset with 15 features was selected for the prediction of rotor bearing temperature. In data pre-processing stage, few features which is not necessary for further processing is eliminated like Turbine ID resulting in 14 features of data for further processes. A correlation heat map is shown in the figure 2 to visualize the relationships between the parameters.

TABLE IIHYPER-PARAMETERS VALUES OR METHOD USED

Sr. No.	**Hyper-Parameter**	**Value / Method Used**
1	Number of Neighbours (K)	7
2	Distance Metric	Euclidean Distance
3	Hyper-parameter Tuning Method	Grid Search

The table II shows the values chosen for the respective hyper-parameters. The values have been carefully chosen based on various factors and after performing various tests. The number of neighbours (K) was selected as 7 based on the elbow curve shown in the figure 3.

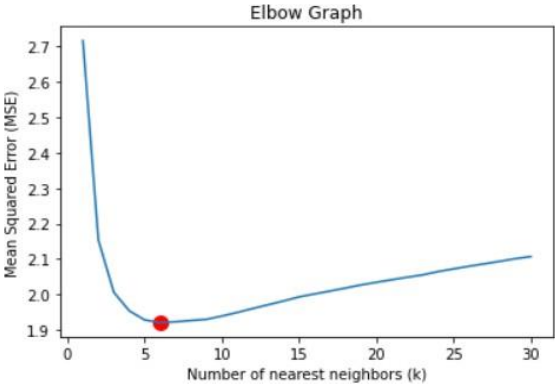

Fig. 3. Elbow Graph

The distance metric was chosen to be Euclidean Distance, which is a measure of the true straight-line distance between two points in Euclidean space, being the default distance metric and most frequently used. To arrive at the best hyper- parameter metrics, while also performing various tests to determine the best values individually, we also used a hyper- parameter tuning method namely Grid Search to programmati- cally arrive at the best metrics. The above chosen metrics have been arrived by combining the results of both elbow curve and grid search methods. Various error metrics obtained by comparing the predicted value and true value has been listed in the table III. Based on the error values obtained, it is evident that the error values are low, acceptable and well within the range.

TABLE III:ERROR METRICS TABLE

Sr. No.	**Error Metrics**	**Error Value**
1	Mean Absolute Error	0.9300
2	Mean Squared Error	1.7776
3	Root Mean Squared Error	1.3333
4	Median Absolute Error	0.6970
5	Mean Squared Logarithmic Error	0.0007
6	Mean Absolute Percentage Error	0.01998

7	R-squared score	0.7397
8	explained variance score	0.7424

From the figure 4, it is clear that the histogram of residuals for the applied KNN regression is normally distributed with most of the points lying near to zero and following bell shaped curve.

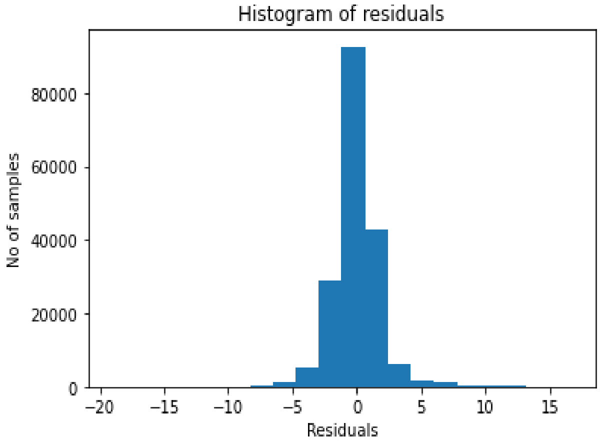

Fig. 4. Histogram of Residuals

Rotor bearing temperature prediction with a accuracy of 98.002% has been acheived using K-nearest neighbour regres- sor.

The next step after finding the rotor bearing temperature is to predict the fault using K-means clustering classifier. Here we have used Silhouette Analysis for programmatically identi- fying the best set of hyperparameters. The optimal number of

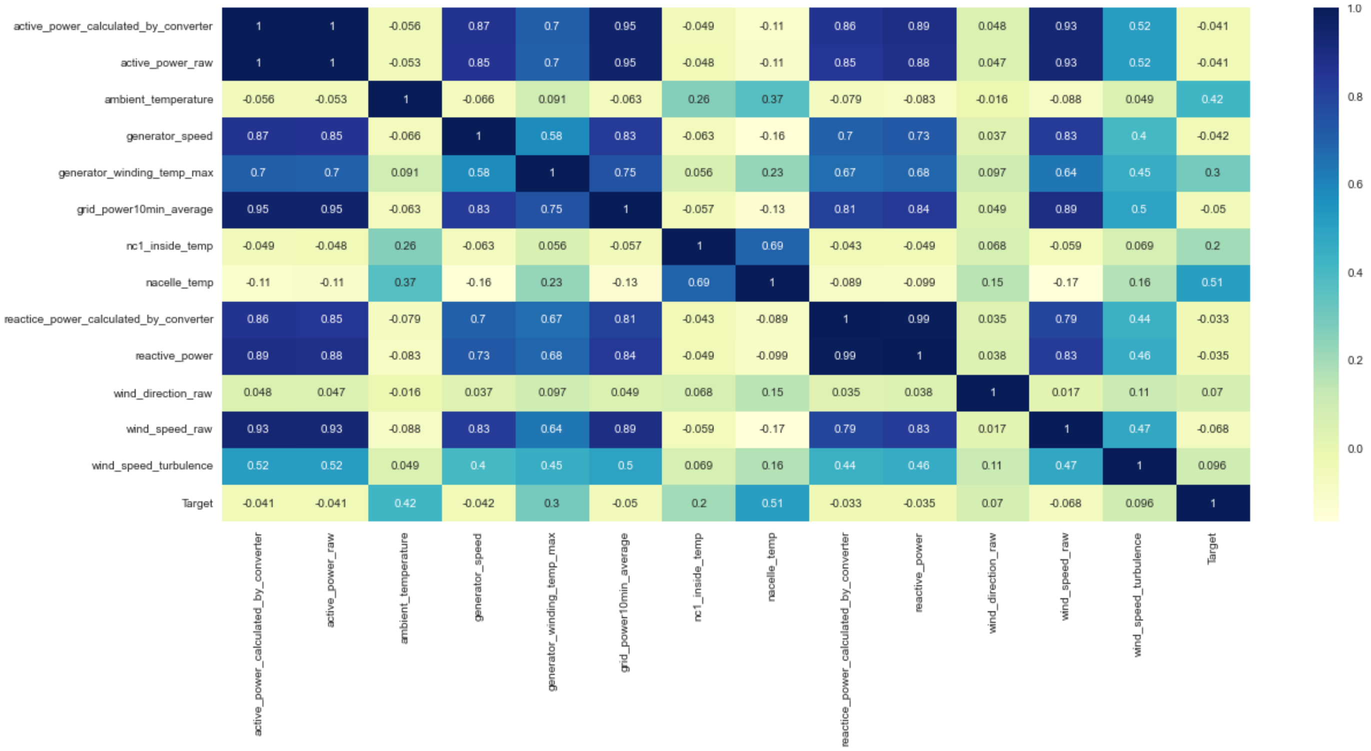

Fig. 2. Heatmap

clusters (K) for the K-means clustering was determined based on the above method.

TABLE IV:HYPER-PARAMETERS VALUES OR METHOD USED

Sr.No	Hyper-Parameter	Value / Method Used
1	Number of clusters (K)	2
3	Hyper-parameter Tuning Method	Silhouette Analysis

The number of cluster(K) was selected as 2 based on the graph shown below in figure 5

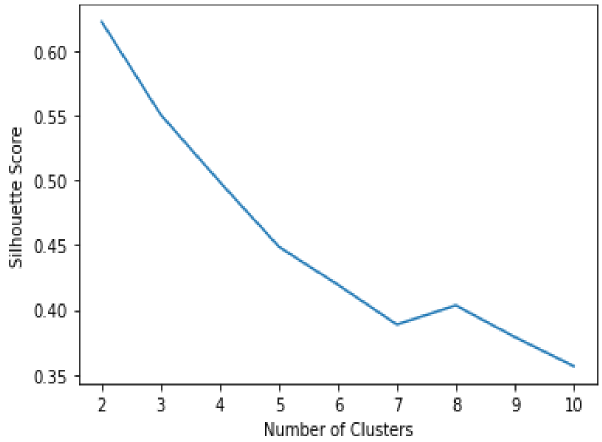

Fig. 5. K means Graph

Based on the silhouette score obtained, K-means clustering classifier has divided the whole data into 2 clusters and the cluster distribution has been indicated in figure 6.

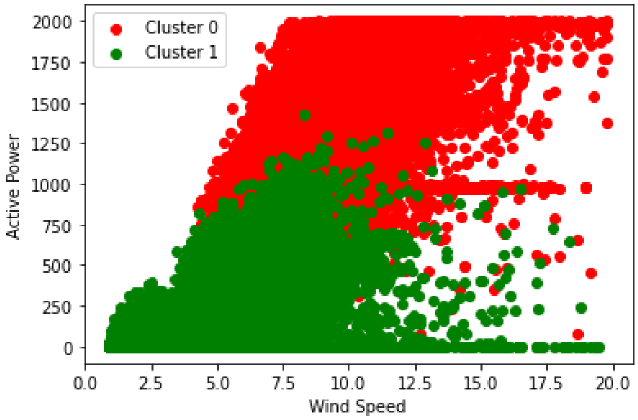

Fig. 6. Cluster Distribution

The classification of the temperature data into healthy point and faulty point has been done and shown in the figure 7.

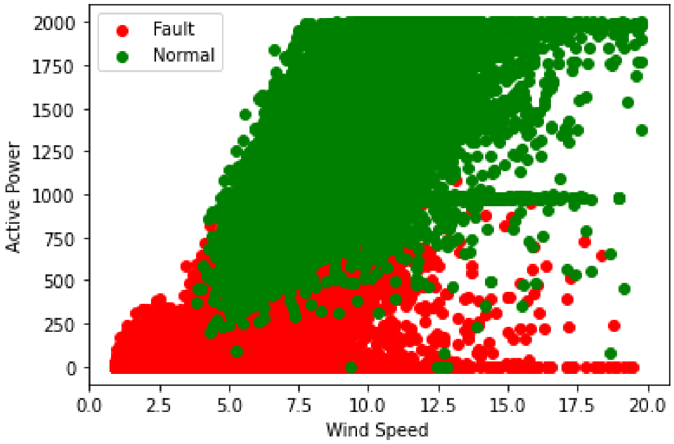

Fig. 7. Fault results

In this study, we have achieved the classification based on the predicted rotor bearing temperature into normal condition and fault condition with a silhouette score of 0.6221.

CONCLUSION

We have established that the analysis of rotor bearing temperature data can play an important role in detecting faults in wind turbines. This study demonstrates the usefulness of K- nearest neighbor (KNN) regression for predicting rotor bearing temperature, the analysis performed and the model built has a high accuracy of 98.002% and low mean absolute error of 0.9300, with the residuals plot showing a normal distribution, thereby suggesting that the model is suitable for predicting rotor bearing temperature. The KNN regression results were used to identify potential gearbox faults through K-means clustering, with a Silhouette Score of 0.6221. This research highlights the potential of using KNN regression and K-means clustering for fault detection in wind turbine gearbox and provides a valuable tool for optimizing maintenance schedules and reducing the risk of unplanned downtime.

REFERENCES

IEA (2022), Wind Electricity, IEA, Paris https://www.iea.org/reports/wind-electricity, License: CC BY 4.0

A´ . M. Costa, J. A. Orosa, D. Vergara, and P. Ferna´ndez-Arias, “New Tendencies in Wind Energy Operation and Maintenance,” Applied Sciences, vol. 11, no. 4, p. 1386, Feb. 2021, doi: 10.3390/app11041386.

Maples, Ben, et al. Installation, operation, and maintenance strategies to reduce the cost of offshore wind energy. No. NREL/TP-5000-57403. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2013.

Crabtree CJ, Zappala´ D, Hogg SI. Wind energy: UK experiences and off- shore operational challenges. Proceedings of the Institution of Mechan- ical Engineers, Part A: Journal of Power and Energy. 2015;229(7):727- 746. doi:10.1177/0957650915597560

Helsen, Jan, et al. ”Condition monitoring by means of scada analysis.” Proceedings of European wind energy association international confer- ence Paris. 2015.

Tautz-Weinert, Jannis, and Simon J. Watson. ”Using SCADA data for wind turbine condition monitoring–a review.” IET Renewable Power Generation 11.4 (2017): 382-394.

Maldonado-Correa, Jorge, et al. ”Using SCADA data for wind turbine condition monitoring: A systematic literature review.” Energies 13.12 (2020): 3132.

A. Desai, Y. Guo, S. Sheng, S. Sheng, C. Phillips, and L. Williams, “Prognosis of Wind Turbine Gearbox Bearing Failures using SCADA and Modeled Data”, PHM CONF, vol. 12, no. 1, p. 10, Nov. 2020.

Wenfeng Hu, Hong Chang, Xingsheng Gu,A novel fault diagnosis technique for wind turbine gearbox,Applied Soft Computing,Volume 82,2019,105556,ISSN 1568-4946,

Koukoura, Sofia, et al. ”Comparison of wind turbine gearbox vibration analysis algorithms based on feature extraction and classification.” IET Renewable Power Generation 13.14 (2019): 2549-2557.

Salameh, Jack P., et al. ”Gearbox condition monitoring in wind turbines: A review.” Mechanical Systems and Signal Processing 111 (2018): 251- 264.

Selwyn, T. Sunder, and S. Hemalatha. ”Experimental analysis of me- chanical vibration in 225 kW wind turbine gear box.” Materials Today: Proceedings 46 (2021): 3292-3296.

Carroll, James, Alasdair McDonald, and David McMillan. ”Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines.” Wind Energy 19.6 (2016): 1107-1119.

de Azevedo, Henrique Dias Machado, Alex Maur´ıcio Arau´jo, and Nade`ge Bouchonneau. ”A review of wind turbine bearing condition monitoring: State of the art and challenges.” Renewable and Sustainable Energy Reviews 56 (2016): 368-379.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. ”Multilayer feedforward networks are universal

approximators.” Neural networks 2.5 (1989): 359-366.
Jia, Feng, et al. ”Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data.” Mechanical systems and signal processing 72 (2016): 303-315.
Ibrahim, Raed, Jannis Weinert, and Simon Watson. ”Neural networks for wind turbine fault detection via current signature analysis.” (2016).
Kalogirou, Soteris A. ”Artificial neural networks in renewable energy systems applications: a review.” Renewable and sustainable energy reviews 5.4 (2001): 373-401.
Burges, Christopher JC. ”A tutorial on support vector machines for pattern recognition.” Data mining and knowledge discovery 2.2 (1998): 121-167.
Stetco, Adrian, et al. ”Machine learning methods for wind turbine condition monitoring: A review.” Renewable energy 133 (2019): 620- 635.
Chawla, Nitesh V., et al. ”SMOTE: synthetic minority over-sampling technique.” Journal of artificial intelligence research 16 (2002): 321- 357.
Tang, Baoping, et al. ”Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine.” Renewable Energy 62 (2014): 1-9.
Jiang, Guoqian, et al. ”Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox.” IEEE Transactions on Industrial Electronics 66.4 (2018): 3196-3207.
Joshuva, A., and V. Sugumaran. ”A data driven approach for condition monitoring of wind turbine blade using vibration signals through best- first tree algorithm and functional trees algorithm: A comparative study.” ISA transactions 67 (2017): 160-172.
Faulstich, Stefan, Berthold Hahn, and Peter J. Tavner. ”Wind turbine downtime and its importance for offshore deployment.” Wind energy 14.3 (2011): 327-337.
Kavaz, Ayse Gokcen, and Burak Barutcu. ”Fault detection of wind turbine sensors using artificial neural networks.” Journal of Sensors 2018 (2018): 1-11.
Peng, Yayu, et al. ”Wind turbine drivetrain gearbox fault diagnosis using information fusion on vibration and current signals.” IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-11.