SCADA Data-Driven Wind Turbine Main Bearing Fault Prognosis Based on Principal Component Analysis

SCADA Data-Driven Wind Turbine Main Bearing Fault Prognosis Based on Principal Component Analysis

Lorena Campoverde[1], Christian Tutivén[1], Yolanda Vidal[2,3] and Carlos Benalcázar-Parra[4]

[1]ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, Faculty of Mechanical Engineering and Production Science (FIMCP), Mechatronics Engineering, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
[2]Control, Modeling, Identification and Applications (CoDAlab), Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya (UPC), Campus Diagonal-Besós (CDB), Eduard Maristany, 16, 08019 Barcelona, Spain
[3]Institute of Mathematics (IMTech), Universitat Politècnica de Catalunya (UPC), Pau Gargallo 14, 08028 Barcelona, Spain
[4]Universidad ECOTEC, Km. 13.5 Vía a Samborondón - Guayaquil, Ecuador

E-mail: lscampov@espol.edu.ec, cjtutive@espol.edu.ec, yolanda.vidal@upc.edu.ec, carlosbenalcazarparra@gmail.com

Abstract. Condition monitoring for wind turbines is essential for the further development of wind farms. Currently, many of the works are focused on the installation of new sensors to predict turbine failures, which raises the cost of wind projects. Wind turbines operate in a wide variety of environmental conditions, such as different temperatures and wind speeds that vary throughout the year season. Typically, most or all of the data available in a turbine is healthy data (operation without failure), so data-driven supervised classification methods have data imbalance problems (more data from one class). Also, when historical pre-failure data do not exist, those methods cannot be used. Taking into account the aforementioned difficulties, the stated strategy in this work is based on a principal component analysis anomaly detector for main bearing failure prognosis and its contributions are: i) this methodology is based only on healthy SCADA data, ii) it works under different seasons of the year providing its usefulness, iii) it is based only on external variables and one temperature related to the element under diagnosis, thus avoiding data containing information from other fault types, iv) it accomplishes the main bearing failure prognosis (several months beforehand), and v) the performance of the proposed strategy is validated on a real in production wind turbine.

1. Introduction

Renewable energy is increasingly important in our daily lives. Its greater growth is due to the great interest in minimizing environmental damage in energy production. Wind energy is an excellent option and one of the fastest growing alternatives in recent years because it is a clean and exists over a wide geographical area, unlike fossil-fuel energy sources which are air pollutants and are concentrated in a limited number of countries. This growth can be evidenced in the deploy of 87000 (distributed) wind turbines (WTs) between 2003 and 2020 across all 50 states of United States (U.S.), Puerto Rico, the U.S. Virgin Islands, and Guam, totaling 1,055 MW

in cumulative capacity [1]. Also, the U.S. wind capacity addition equaled 16836 MW in 2020 bringing the cumulative total to 121985 MW at the end of the year [2]. This growth represented $24.6 billion of investment in new wind power project installations in 2020, for a cumulative investment total of roughly $240 billion since the beginning of the 1980s [2]. In fact, the year 2020 was the best year in history for the global wind industry, showing year-over-year growth of 53% and the installation of more than 93 GW wind power [3]. This growth goes hand in hand with the continuous increase in the size of turbines, which now come with average rotor diameters greater than 150 meters and turbine capacity greater than 7.5 MW [4]. But this continuous growth requires better condition monitoring systems.

Condition monitoring systems trace the behavior of different signals to know the status of a machine. In this manner, they monitor when there are deviations from normal operating behavior, which is indicative that a fault is developing. Condition monitoring is the foundation of predictive maintenance, which uses advanced analysis based on the current operating conditions of the assets [5]. Forecast future machine states allows maintenance to be planned before failure occurs (predictive maintenance). Ideally, predictive maintenance allows reducing maintenance frequency, reducing the costs associated with performing too many unnecessary preventive maintenance, in addition to preventing unplanned corrective maintenance that can have a high cost. Consequently, companies are conducting research to develop such reliable methods to predict failures associated with WT components [6]. One way to predict failures is detecting "not normal" conditions. This process is commonly known as anomaly detection or outlier detection, defined by Grubbs in 1969 [7]. One of the anomaly detection setups is the semi-supervised anomaly detection method, that consist in training a model only using data without anomalies. The basic idea is, that a model of the normal class is learned and anomalies can be detected afterwards by deviating from that model [8]. This technique is also called "one-class" classification.

Many of the anomaly detection works are based on the use of expensive sensors. On the other hand, industrial-size WT come equipped with a supervisory control and data acquisition (SCADA) systems allowing to collect data for the correct operation of the equipment. In recent years, there has been a growing interest in using these data not only for the proper control of the turbines but also for fault prognosis since it avoids the increase in costs by not having to buy additional sensors. For example, in, [9] is fitted a support vector machines regression to model gearbox oil temperature using SCADA selected variables as predictors. In [10] is proposed a method that only requires healthy WT SCADA data to be collected and trained an artificial neural network with a Bayesian's regularization to predict main bearing failures. A novel dynamic model sensor method to represent the relationship between the generator temperature, wind speed, and ambient temperature is developed in [11] to detect WT generator faults.

Principal Component Analysis (PCA) is typically used for dimensionality reduction in data analysis [12]. In this article, a condition monitoring system based on a PCA anomaly detector to predict WT main bearing failures in advance is proposed, thus giving maintenance staff time to coordinate a review and maintenance without incurring high costs. The basic idea is to develop an PCA anomaly detection using only normal (healthy) SCADA data and then when an inference is made with future data to be able to detect anomalies when there is a pre-failure in the main bearing. The main advantage of using PCA for anomaly detection, compared to alternative techniques such as a neural autoencoder, is its simplicity as it is based on linear algebra, which is computationally easy to solve by computers. In addition to this, machine learning algorithms converge faster when trained on the principal components rather than the original dataset. Finally, another advantage is that high-dimensional data makes regression-based algorithms easily overfit. By using PCA beforehand to reduce the dimensions of the training dataset, we prevent the predictive algorithms from overfitting. On other hand, PCA algorithm has some

disadvantages. Principal components are linear combinations of the original data's features, although they are more difficult to interpret. It's difficult to discern which features in a dataset are the most relevant after computing main components, for example. Another drawback is that while dimensionality reduction is useful, it comes at a cost. Information loss is a necessary part of PCA.

The remainder of this work is organized as follows: A brief description of the used WT and of the SCADA data are provided in Section 2. How is developed the data split is shown in Section 3. In Section 4 the exploratory analysis is explained. Then, in Section 5 the data pre-processing is carried out. The proposed PCA anomaly detection methodology is described in Section 6. The employed failure indicator is given in Section 7. The obtained results are given and discussed in Section 8. Eventually, conclusions are given in Section 9.

2. Wind turbine and SCADA data description

The WT used in this study is located in a wind farm in Poland. It presents the characteristics described in Table 1. The WT rotor has a diameter of 101 m, a swept area of 8000 m^2 and is capable of producing 2300 kW of power. This turbine comes with a power regulation system that allows it to operate near its maximum aerodynamic efficiency for most of the time. Figure 1 details the main components of the WT.

Table 1. Main WT technical characteristic.

Number of blades	3
Nominal power	2300 kW
Rotor diameter	101 m
Wind class	IEC IIb
Swept area	8000 m^2
Rotor speed	6-16 rpm
Cut-in wind speed	3-4 m/s
Rated wind speed	12-13 m/s
Cut-out wind speed	25 m/s
Gearbox type	3-stage planetary/helical
Gearbox ratio	1:91
Power regulation	Pitch regulation with variable speed

A wind power SCADA (WPS) system is part of the WT, which allows remote access to turbine SCADA information in real time. Different electrical and mechanical measures, operating and fault status, and different data affecting and related to the turbine and the environment are collected by the WPS system. This information is the one used to predict WT main bearing failures. The SCADA data were obtained from January 01, 2014, to December 12, 2019, where the mean, maximum, minimum, and standard deviation values of the average period of 10 minutes from the electrical, hydraulic, environmental, control and component temperature measures collected by the WPS system are collected.

When it is desired to study a specific failure, experts must be skillful in choosing the most important variables for the physical system to be studied. In this work, only the mean values of the environment SCADA measurements and the mean of the most related temperature (main shaft bearing temperature) to the studied failures (bearing failures) are used. This is because if variables strongly related to other components of the turbine are employed instead of only the component to be studied, this methodology will not only discover the fault of interest, but also faults in the components that have a strong relationship with the used variables. For example, if one of the input variables is the blade position signal, the model might detect pitch-related

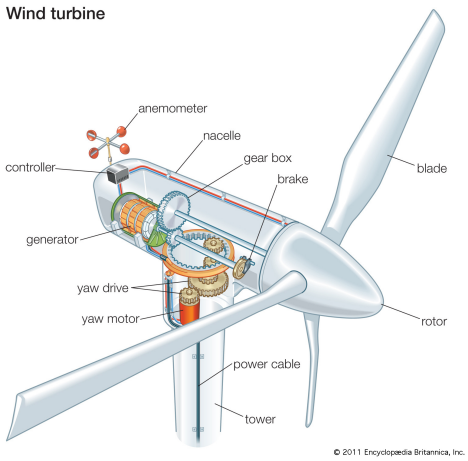

Figure 1. Main components of the wind turbine [13]

.

errors. The selected variables to use in this work are detailed in Table 2. The environment variables can affect the behavior of the WT due to the seasons. For example, the ambient temperature influences all subsystems temperatures (bearings temperature change from winter to summer). The wind speed, which determines the distinct operating regions of the WT, is the most important exogenous variable associated to the WT because of its direct effect on the WT's operation region [14].

Table 2. Selected SCADA measurements.

Variable	Description
MainBTmp	Main shaft bearing temperature
AmbieTmp	Ambient temperature
A1ExtTmp	Ambient temperature of another sensor
SeWindSp	Related to wind speed
AcWindSp	Related to wind speed
PrWindSp	Related to wind speed
PriAnemo	Related to anemometer measurements
SecAnemo	Related to anemometer measurements

In addition, it is relevant that there is WT extra information detailing maintenance and repair dates, which allows having a notion of failures and repair or replacement of elements in the dataset (work orders).

3. Data split

The information from the work orders is used to separate the data. According to this information, the studied WT has a main bearing corrective reparation on May 21, 2018. When bearing failure initiates (e.g., initial crack), it is usually accompanied by a momentary release of frictional heat, but then the bearing temperature goes back to normal (crack is stabilized and not growing). The importance of this methodology is to detect this heat release months before the bearing is completely damaged. For more detailed information about the mentioned failure modes, see [10] and, [15].

It's vital to note that this research is based on a PCA anomaly detection methodology, which implies the model is only trained using one-class data, in this case, healthy data (Section 6 gives more information). Also, considering that it is not desired that seasonality or any environmental change affects the training of the model and therefore its failure prediction, a division of training and test data is made, where each of them has more than one year of information. Given these two requirements, the date range of the training data is from January 01, 2014 to December 11, 2017 (six months before a main bearing failure in work orders, so healthy data) and the test data are selected from December 11, 2017 to December 12, 2019 (pre-failure data), as seen in Figure 2.

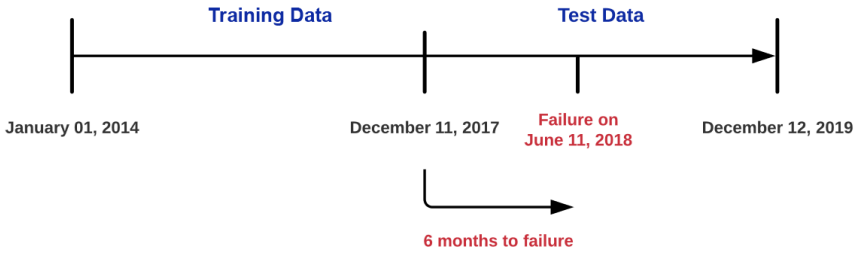

Figure 2. Data split.

4. Exploratory data analysis

An exploratory data analysis (EDA) is a procedure for studying and analyzing data to select important variables, to correct outliers and developing models. In short, it goes into the study of the data to obtain relevant characteristics by means of different graphing methods, ensuring the effectiveness of the subsequent process [16]. Figure 3 shows the behavior of the selected variables. It is observed that MainBTmp presents seasonalities as do those of the A1ExtTmp and AmbieTmp variables, showing that the main bearing temperature is affected by the ambient seasonalities. On the other hand, the SeWindSp and SecAnemo variables present very similar behaviors between them, as do the AcWindSp and PrWindSp variables. All these variables are characterized by an increase in wind speed changes in equal stages, making it known that the samples can be altered by wind gusts. In the graph of PriAnemo, it is reported that it's a constant value of 1.2, deciding to eliminate it from the selected variables because it does not provide relevant information for the study.

Given the possibility of missing values or outliers, which can occur due to equipment failures or inconsistent data collection by the WPS system, different pre-processing techniques are developed to transform the data into effective and quality data.

5. Data pre-processing

The development of a good predictive model strongly depends on the data pre-process, ensuring the effectiveness of the model. Among the different data preprocessing techniques, there is data cleaning, integration, transformation, decrementing and variable selection, which can be used jointly or individually [17]. In the present work, the first and the last preprocessing techniques are used and detailed below.

5.1. Data cleaning

Data cleaning is a process in which the missing or detected outliers are treated with the aim of correcting them to generate a robust and quality dataset [18]. One outlier detection method is

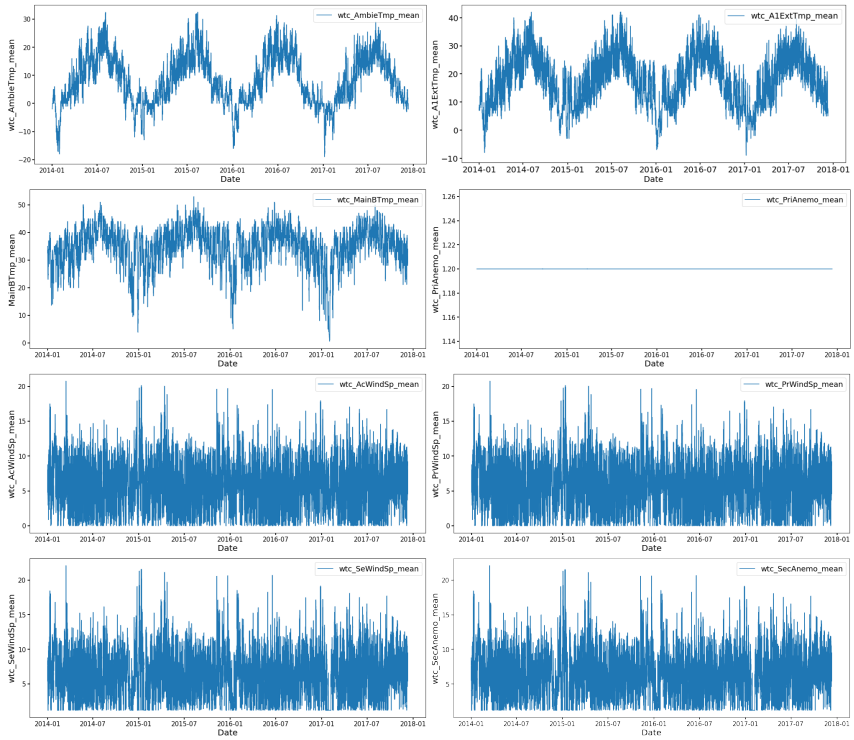

Figure 3. Variables plots.

the interquartile range method. Using the interquartile range method, it was determined that there are no outliers in the data, which increases the predictive efficiency of the model, since it is not contaminated. On the other hand, The number of empty values for each variable is obtained and they are imputed using two techniques as indicated in [10]. So, the interpolating polynomial Piecewise Cubic Hermite imputation method is used to fill the missing values between two samples. The filled new points guarantee monotonic behavior of the function and that the first derivative is continuous. In addition, given the possibility of missing values at the edges, the closest values after or before the missing values are used, respectively.

Figure 4 shows the curve with the original and imputed data, where it is shown how the missing data are imputed using the techniques mentioned before.

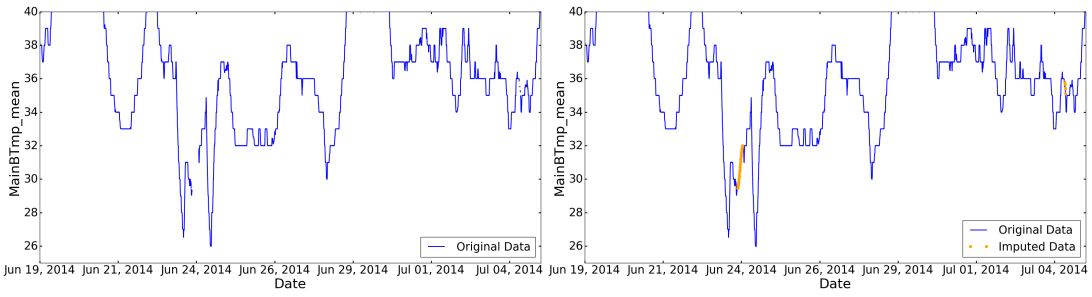

Figure 4. Data imputation strategy.

5.2. Variable selection

As part of the preprocessing, it is necessary to define a variable selection to keep only those variables that contain relevant information and to ensure that the training of the model is optimal [19]. This will help to perform a faster model training and a faster data inference, but at the same time efficient. To select the model inputs, a correlation analysis is carried out. The correlation analysis is a method used to denote the association or relationship between two or more quantitative variables [20].

In this study, the analysis was performed using Pearson's correlation coefficient, which uses a measurement of linear dependence between variables, which is between -1 and +1, being positive when the decrease of one generates the increase of the other. For it to be negative, it is the opposite of the above and zero when there is no relationship between them [21]. Figure 5 shows the correlation analysis of all the predictor variables, where if two variables have a correlation greater than 0.8, one of them is selected and the other is discarded. Therefore, it was concluded that there are highly correlated variables, which implies that they provide the same information, resulting in the final selection of MainBTmp, SecAnemo and AmbieTmp as the input model variables for training and for inferences.

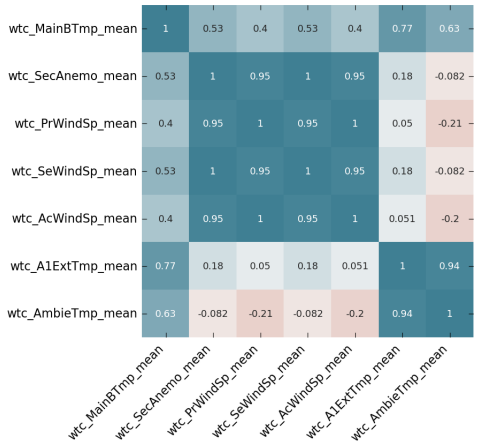

Figure 5. Pearsons correlation heatmap.

After the variable selection stage, the variable mean ambient temperature is subtracted from the mean main shaft bearing temperature to avoid the problem of seasonality variation. With the pre-processed data, the training of the anomaly detection model can already be carried out, as indicated in the next section.

6. Principal component analysis anomaly detector

Principal component analysis is a statistical technique that breaks down a data matrix into vectors called principal components. It is often used to reduce the dimension of data for easy exploratory data analysis. The principal components can be used for several purposes. For example, it can be used to compress information so that users require less space to store it, to reduce high-dimensional data to 2 or 3 dimensions so it can be plotted and explored, obfuscating datasets so they can be shared with others without revealing the nature or meaning of the data. But one of the most novelty uses and for which it is used in this work is to detect anomalies. Anomaly detection is a type of machine learning that looks for anomalies in data sets. The goal of anomaly detection is to identify outliers in data – samples that aren't "normal" when compared to others.

As stated in the basic PCA theory, it can be used to reduce data from m dimensions to n (in this work from three variables to two principal components), and that a PCA transform can be inverted to restore the original m dimensions. You also know that inverting the transform doesn't recover the data lost when the transform was applied. The essence of PCA-based anomaly detection is that an anomalous sample should have more losses or reconstruction errors than a normal sample. In other words, the loss incurred when an abnormal sample is passed through the PCA anomaly detector algorithm must be greater than the loss incurred when the same operation is applied to a normal sample. Figure 6, show an example. Here two data points are passed through the trained PCA model. As can be seen the point 1 has a higher reconstruction error than the point 2. This allows the point 1 to be identified as a possible anomaly.

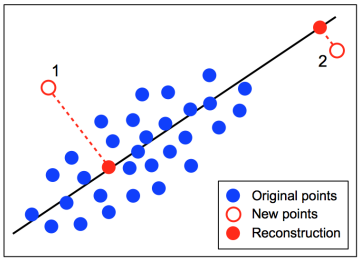

Figure 6. Anomaly detection using PCA reconstruction error.

In this work, the PCA technique is used to detect observations that are different from the majority of the data, in this case training data (healthy samples). Taking into account this approach, it is assumed that the anomalies (test data or pre-failure samples) are qualitatively different from normal samples (training data or healthy samples). The general idea is that the first few principal components explain the largest cumulative proportion of the total sample variance [22]. Consequently, the observations that are outliers (pre-failure data) with respect to the first few components usually correspond to outliers. For this, the PCA anomaly detector model is trained only with one-class data (healthy data) and then imputations to detect anomalies are carried out with the test data set.

With the PCA anomaly detector algorithm we defined, it is possible to resume all the described previous pre-process stages in a diagram, as it is shown in Figure 7. The diagram explains data cleaning, the variable selection, data normalization and finally the training stage of the PCA anomaly detector model.

7. Indicator

In this section, a failure indicator is proposed to trigger an alarm when the number of detected anomalies is higher than a prescribed threshold and this occurs when a main bearing failure exists. To define the threshold, first an anomaly weekly grouping (with only training data set) is carried out to avoid having false alarms due to specific events. After this process takes part, to smooth the historical data and also to reduce the number of false alarms, an exponential weighted moving average (EWMA) is used. The EWMA gives less importance to older information than to more recent data, which has a greater weight assigned. The EWMA is based on a recursive function, where the current value is obtained through the previous one, causing the weights to be reduced as one goes back in history [23].

Finally, to define the prescribed threshold, the mean (μ) and the standard deviation (σ) of

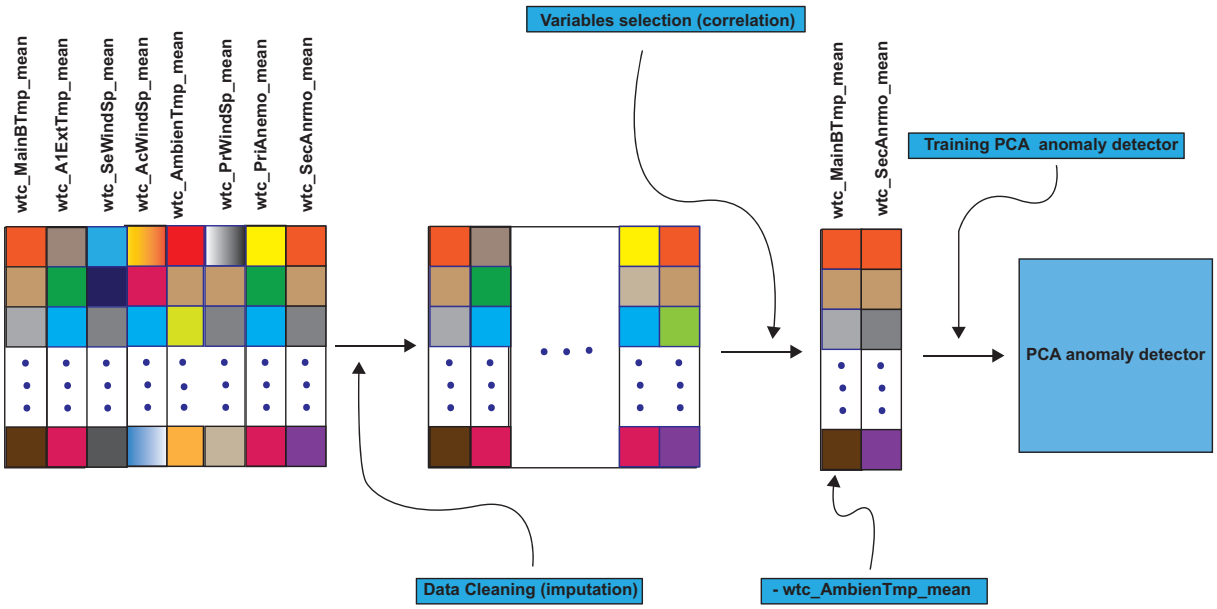

Figure 7. First stage of the proposed strategy: training the PCA anomaly detector.

the obtained output are then calculated. And then, the threshold is defined as:

$$\text{threshold} = \mu + 3\sigma. \tag{1}$$

Once the failure indicator is defined, the performance (of the algorithm) is validated through the test data set, through a similar process. First, generating an anomalies count, then grouping them by weeks and, finally, if any value in the test data set exceeds the predefined threshold, this will generate an alert, as can be seen in Figure 8.

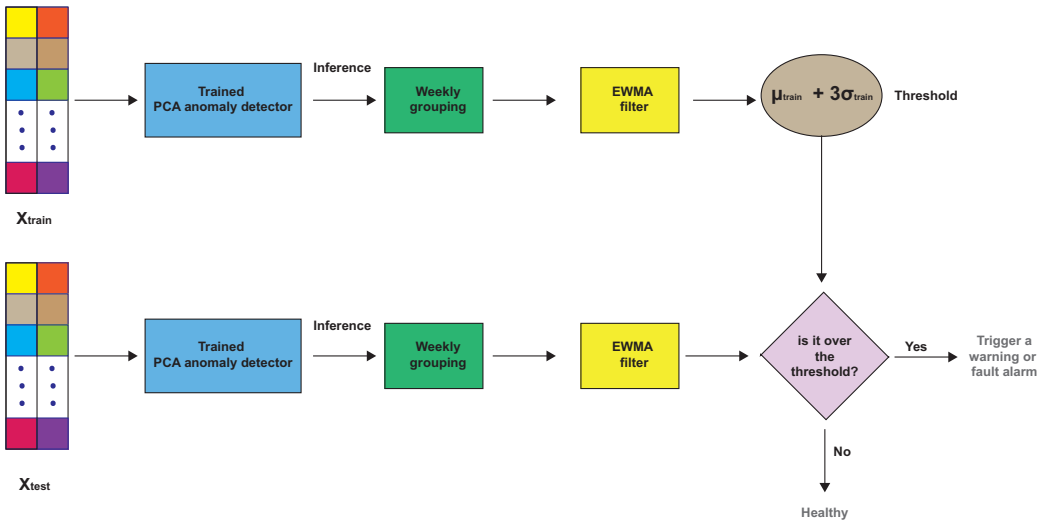

Figure 8. Second stage of the proposed strategy. Given test data, the PCA trained anomaly detector is used, and postprocess (with weekly grouping filtered with EWMA) ensures a reduced number of false positives.

8. Results

According to the work orders, this WT has a main bearing failure on June 11, 2018. Figure 9 shows the smoothed (EWMA filter) weekly grouping detected anomalies from the test data set. As can be seen the failure indicator exceeds the predefined threshold around two months before the failure occurs, presenting three alerts in 2018: the first one, on April 9, the next one on April 16 and the last one on June 11. Therefore, the system guarantees the failure alarm months before this occurs, which will provide time to prepare a preventive maintenance. As previously stated, when a bearing failure initiates, frictional heat is typically released briefly and then the bearing temperature stabilizes. This is why the error falls below the threshold before maintenance. In addition, once the main bearing is repaired, the anomalies no longer exceeded the threshold.

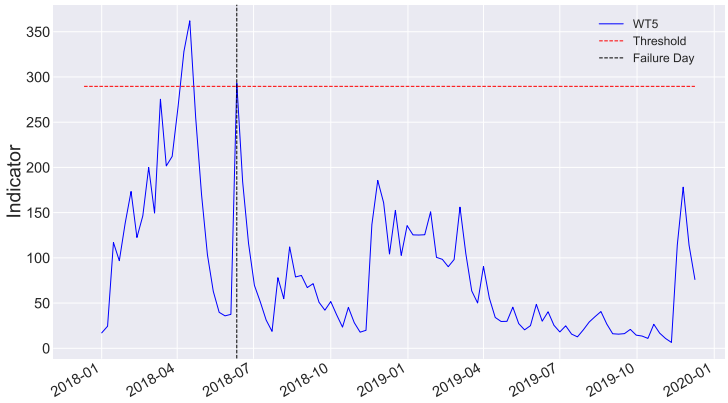

Figure 9. Weekly indicator for WT test data.

9. Conclusions

This study presents a model trained to detect anomalies through principal component analysis when a main bearing failure is about to occur. It is demonstrated that the alerts are presented two months before the failure occurs, which offers benefits to the industry by allowing the development of a preventive maintenance plan. It must be taken into account that this model has been used only for the turbine under study. In case it is desired to deploy main bearing failure prognosis in other turbines, it is recommended that each one should have its own model. But the advantage of this methodology is that no failure data (which are difficult or impossible to obtain in real applications) is needed to train the model. Therefore, the strategy validated with this turbine is applicable to any other turbine.

Acknowledgments

This work has been partially funded by the Spanish Agencia Estatal de Investigación (AEI)—Ministerio de Economía, Industria y Competitividad (MINECO), and the Fondo Europeo de Desarrollo Regional (FEDER) through the research project DPI2017-82930-C2-1-R; and by the Generalitat de Catalunya through the research project 2017 SGR 388. The authors thank a lot to Smartive company, as this work would not have been possible without their support in the ceding of wind farm data.

References

[1] A. C. Orrell, K. Kazimierczuk, and L. M. Sheridan, "Distributed wind market report: 2021 edition," Pacific Northwest National Lab.(PNNL), Richland, WA (United States), Tech. Rep., 2021.

[2] R. H. Wiser, M. Bolinger, B. Hoen, D. Millstein, J. Rand, G. L. Barbose, N. R. Darghouth, W. Gorman, S. Jeong, A. D. Mills *et al.*, "Land-based wind market report: 2021 edition," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2021.
[3] G. W. E. Council, "Gwec— global wind report 2021," 2021.
[4] P. Enevoldsen and G. Xydis, "Examining the trends of 35 years growth of key wind turbine components," *Energy for sustainable development*, vol. 50, pp. 18–26, 2019.
[5] J. Daily and J. Peterson, "Predictive maintenance: How big data analysis can improve maintenance," in *Supply chain integration challenges in commercial aerospace*. Springer, 2017, pp. 267–278.
[6] P. Mazidi, M. Du, L. B. Tjernberg, and M. A. S. Bobi, "A performance and maintenance evaluation framework for wind turbines," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2016, pp. 1–8.
[7] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.
[8] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.
[9] D. Zhang and Z. Qian, "Probability warning for wind turbine gearbox incipient faults based on scada data," in *2017 Chinese Automation Congress (CAC)*. IEEE, 2017, pp. 3684–3688.
[10] Á. Encalada-Dávila, B. Puruncajas, C. Tutivén, and Y. Vidal, "Wind turbine main bearing fault prognosis based solely on scada data," *Sensors*, vol. 21, no. 6, p. 2228, 2021.
[11] S. Zhang and Z.-Q. Lang, "Scada-data-based wind turbine fault detection: A dynamic model sensor method," *Control Engineering Practice*, vol. 102, p. 104546, 2020.
[12] S. Roweis, "Em algorithms for pca and spca," *Advances in neural information processing systems*, pp. 626–632, 1998.
[13] Britannica, "Wind turbine," 2011. [Online]. Available: https://www.pm365.ga/ProductDetail.aspx?iid=65620147 & pr=
[14] P. Tavner, C. Edwards, A. Brinkman, and F. Spinato, "Influence of wind speed on wind turbine reliability," *Wind Engineering*, vol. 30, no. 1, pp. 55–72, 2006.
[15] "Bearing damage and failure analysis," https://www.skf.com/binaries/pub12/Images/0901d1968064c148-Bearing-failures—14219_2-EN_tcm_12-297619.pdf, 2017, accessed: 2021-07-08.
[16] T. Shin, "An extensive step by step guide to exploratory data analysis," jan 2020. [Online]. Available: https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e
[17] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.
[18] J. Brownlee, *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.
[19] K. J. Max Kuhn, *Applied Predictive Modeling*, 1st ed. Springer, 2013.
[20] N. J. Gogtay and U. M. Thatte, "Principles of correlation analysis," *Journal of the Association of Physicians of India*, vol. 65, no. 3, pp. 78–81, 2017.
[21] S. Kumar and I. Chong, "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states," *International journal of environmental research and public health*, vol. 15, no. 12, p. 2907, 2018.
[22] V. Rokhlin, A. Szlam, and M. Tygert, "A randomized algorithm for principal component analysis," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1100–1124, 2010.
[23] D. Radecic, "Time series from scratch — exponentially weighted moving averages (ewma) theory and implementation," 2021. [Online]. Available: https://towardsdatascience.com/time-series-from-scratch-exponentially-weighted-moving-averages-ewma-theory-and-implementation-607661d574fe