An Insight into Wind Turbine Planet Bearing Fault Prediction Using SCADA Data

Sofia Koukoura[1], James Carroll[2], and Alasdair McDonald[3]

[1,2,3] *University of Strathclyde, Glasgow, G1 1XQ,UK*
sofia.koukoura@strath.ac.uk

ABSTRACT

Condition based maintenance is being adopted into the decision making process of wind farms, in order to reduce operation costs. SCADA systems are integrated in wind turbines, providing low frequency operational data and are increasingly being used in condition monitoring. The aim of this paper is to explore how can wind turbine gearbox components be monitored using SCADA data. The case study presented utilises 10-minute averaged SCADA data from three operating wind turbines that have a double planetary stage gearbox. Historic data is collected for more than a year at sparse time periods before the occurrence of a bearing failure on a planet of the first planetary stage. Data pre-processing is applied using a clustering filter in order to improve prediction confidence. An insight into the data is given which indicates the potential importance of generator speed estimation for planet bearing faults. Normal behaviour models are thus proposed to predict this type of fault. A classification model is also presented, which uses different time periods before the component failure are used for wind turbine health state determination. A successful prediction of the bearing health state can be performed through the suggested models and some insight is given into into the optimal SCADA sensors utilization for this type of failure mode.

1. INTRODUCTION

Wind energy is one of the most rapidly developing renewable energy source for electrical power generation worldwide. With a total net installed capacity of more than 160 GW, wind energy remains the second largest form of power generation capacity in Europe, closely approaching gas installations. As the fleet of turbines is constantly increasing, the need to optimise maintenance actions becomes vital. Therefore, the industry has now moved to a maintenance regime that is more predictive and proactive. Utility scale wind turbines have a Supervisory Control and Data Acquisition (SCADA) system which was originally installed for performance monitoring. SCADA systems provide numerous data at usually 10-minute resolution and are a low cost solution for condition monitoring, requiring no additional sensors like the vibration or oil sensors usually found in traditional condition monitoring systems.

A wide range of approaches that use SCADA for early failure detection has been developed over the past years. A recent comprehensive review of how SCADA data are used for condition monitoring of wind turbines is given by (Tautz-Weinert & Watson, 2016). The main categories of approaches taken using SCADA data for fault detection are trending, clustering and normal behaviour modelling. A trending technique using correlations among relevant SCADA data is investigated in (Yang, Jiang, et al., 2013). The first law of thermodynamics is used to derive the relationship between temperature, efficiency and power output in (Feng, Qiu, Crabtree, Long, & Tavner, 2013), (Feng, Qiu, Crabtree, Long, & Tavner, 2011) and it shows the temperature trend rises, while the efficiency decreases a few months before a planetary gear failure. Clustering can be applied in the form of self organising maps, which have the ability to represent the shape of datasets with complex relations between variables and to visualise of high dimensional datasets. Some work on SCADA data using this method has been performed by (Catmull, 2011) and (Kim et al., 2011). Trending and clustering have both shown limitations for online monitoring due to challenges in interpreting the results and changes and setting thresholds.

Normal behaviour models are used in a wide range of condition monitoring applications including transformers and gas turbines so that anomalies are detected from normal operation (Tarassenko, Nairac, Townsend, Buxton, & Cowley, 2000), (McArthur, Catterson, & McDonald, 2005). A model of the measured parameter is trained based on various operating examples and the residual of the measured minus the modelled signal acts as an indicator of a possible fault. These models have been introduced in SCADA data analysis using either linear and polynomial approaches or artificial neural networks (ANNs). A linear model that can detect generator bear-

ing failures based on the bearing temperature can be found in (Garlick, Dixon, & Watson, 2009). A higher order polynomial full signal reconstruction method for monitoring drive train temperatures is shown in (Wilkinson, Darnell, van Delft, & Harman, 2014) and is tested successfully on real wind turbine gearbox and main bearing failures. ANNs have the ability to determine non-linear relationships between observations, which makes them suitable for SCADA data. Gearbox bearing and cooling oil temperatures have been modelled in (Garcia, Sanz-Bobi, & del Pico, 2006), (Zaher, McArthur, Infield, & Patel, 2009) using ANNs. The advantages of artificial neural networks over linear models are presented in (Schlechtingen & Santos, 2011) and demonstrated on a bearing damage events of offshore wind turbines.

The aforementioned research shows that SCADA data can be successfully used for prediction of incipient wind turbine faults. However, gearbox component failures and their signatures on SCADA data has not be thoroughly researched. It is indeed challenging to correlate faults to specific components in the gearbox and for that reason other sources of data -such as vibration- have been more successfully utilised. There are, nevertheless, cases where diagnosis using vibration signals can be more challenging, or cases where there are no vibration signals available at all. Planet bearing faults in wind turbines is one example that has been proven difficult to diagnose though vibrations, since they are located on the low speed stage and their characteristic frequencies can often be masked by other more dominant components in the gearbox. It should also be noted that the size of the planetary stage determines its overall weight, so in order to control the gearbox weight, the components must be sized close to the margins of calculated design life.

This paper aims to examine which SCADA sensors are more useful to monitor for planetary stage fault detection. Normal behaviour models are built in order to predict speeds across the gearbox and a classification methodology is presented in order to perform wind turbine gearbox prognostics. A case study using historical data from three operating wind turbines, leading up to a planet bearing failure is used to demonstrate the proposed models.

The rest of the paper is organised as follows: the case study of the planet bearing failure is presented along with a SCADA data pre-processing methodology in Section 2. Some data insight at different time periods before the component failure is also given. Based on this insight, a normal behaviour model is developed and validated in Section 3. A classification model that predicts the different times before the component failure is given in Section 4. Finally, conclusions are given in Section 5.

2. DATA INSIGHT

This section gives an overview of the SCADA data used for this case study. A data pre-processing methodology is presented in order to remove outliers. Some insight into the variables is given at different time steps before the wind turbine gearbox component failure.

2.1. Wind Turbine SCADA Data

SCADA systems initially provided measurements for a wind turbine's energy production and to confirm that the turbine was operational through 5-10 minute averaged values transmitted to a central database. However, SCADA systems can also provide warning of impending malfunctions.

Usually the data available through SCADA systems include various operational parameters and temperatures inside the turbine. The most common are active power output, anemometer measured wind speed, rotor speed and generator speed. Regarding temperature sensors, ambient, nacelle, gearbox oil, gearbox bearing and generator wind temperatures are usually considered, but that can differ depending on the commercially available system installed. Often parameters that do not have any obvious relationship with environmental conditions are measured, such as yaw angle error and pitch angle error, but these are out of the scope of this paper.

The three wind turbines considered in this study are offshore wind turbines located in two different wind farms. The turbines are from the same make models, rated between 2.5 and 3.5 MW [1] There are in total almost 4000 samples from each turbine of 10 minute averaged measurements throughout the course of 34 months. The data collected at the start of the period are considered to be from a healthy turbine system, according to the maintenance logs of the operator. A failure on the planetary stage occurred at the end of the 34th month.

The gearbox examined has a structure commonly found in offshore wind turbines, where high step-up ratios and compactness are required. It consists of two planetary stages and one parallel stage. The main shaft is connected to the planet carrier of the first planetary stage and the high speed stage of the gearbox is coupled to the generator. The ring gears of the planetary stages are fixed. The failure mode studied occurred on a first planetary stage planet bearing. It started on the inner race way with debris eventually effecting the outer raceway. The failure mode is the same for all three wind turbines.

The SCADA sensors used in this case study mainly consider the power output, the speed at different stages of the wind turbine and various temperatures inside the gearbox. These are given in Table 1.

The dataset is quite sparse since there are not available measurements for every 10 minutes within the 34 months. A

[1] Ranges are provided for confidentiality purposes.

SCADA Sensors
Gearbox Oil Temperature Bottom
Gearbox Oil Temperature Higher Level
Bearing Temperature High Speed Rotor End
Bearing Temperature High Speed Generator End
Bearing Temperature Intermediate Stage
Bearing Temperature Shaft Low Speed Stage
Bearing Temperature Shaft Generator Stage
Nacelle Temperature
Ambient Temperature
Generator Speed
Rotor Speed
Wind Speed
Electrical Power

Table 1. List of gearbox SCADA sensors.

grouping of the dataset that leads to 4 main time periods of balanced amount of data is performed as follows:

- Healthy
- 1 Year before failure
- 6 months before failure
- 1 month before failure

2.2. Data Pre-Processing

SCADA systems can experience sensor errors and maintenance actions can lead to missing data. The process followed for preprocessing the training data is similar to the one described in (Bangalore, Letzgus, Karlsson, & Patriksson, 2017).

Firstly, samples with missing values or no power production are filtered out. Moreover, the aim of this paper is to understand and model normal behaviour, so curtailment should not be considered, even though it is set manually. Only a few samples of curtailment examples are usually present, which is not sufficient to be used in the training process. Data points where maximum wind speed has reached more than 25 m/s are also filtered out because beyond this wind speed the turbine is stopped. These points will not fit any pattern, thus cannot be taught to the model. In addition, data sampling during frequent startup or stop in the low-wind-speed period may have a different variation. Thus, a lower limit of output power is set at 0 kW for data sample selection.

The cluster filter is applied on the training data and aims to remove outliers depending on the operating conditions of the wind turbine. A simple threshold does not take into account the nonlinear operational characteristics of the system. A multivariate outlier detection approach based on Mahalanobis distance is used in (Kusiak & Verma, 2013). A similar approach is extended and used in (Bangalore et al., 2017), by dividing the data based on operating power and temperatures ranges. This paper utilizes agglomerative hierarchical clustering (Rokach & Maimon, 2005). Essentially, the distance between every pair of objects in a data set is computed. This information is then used in the linkage function which determines how the objects in the data set should be grouped into clusters that form a binary hierarchical cluster tree. The distance in this paper is calculated in the Euclidean space and the inner squared distance is computed using Ward's algorithm.

The distance is calculated for each data vector in the training data set from its cluster centre. The Mahalanobis distance values can be estimated by a loglogistic distribution as elaborated in (Bangalore et al., 2017) and data below the probability threshold of 2.5% are filtered out. The distribution of the Mahalanobis distance values of the training data set for the wind turbine case study is shown in Figure 1. It can be observed that the values can be estimated by a loglogistic probability distribution function.

A probability threshold of 2.5% is chosen. The cluster filter is used only on the training data set, and therefore false alarms due to curtailment could occur in the implementation stage. It is suggested that the condition monitoring process is blocked during power curtailment, which is available in most modern SCADA system logs. This information was not available in this case study though.

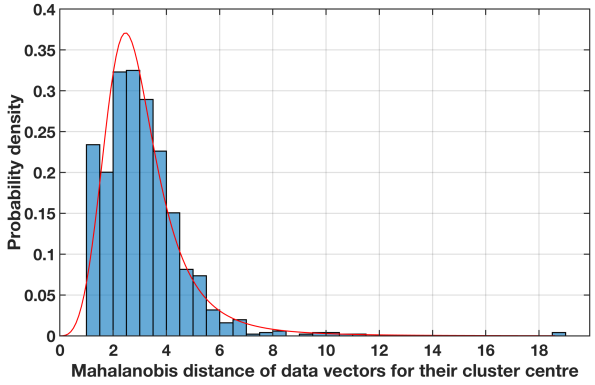


Figure 1. The histogram and probability density function fit for Mahalanobis distance values of data vectors from its cluster centre.

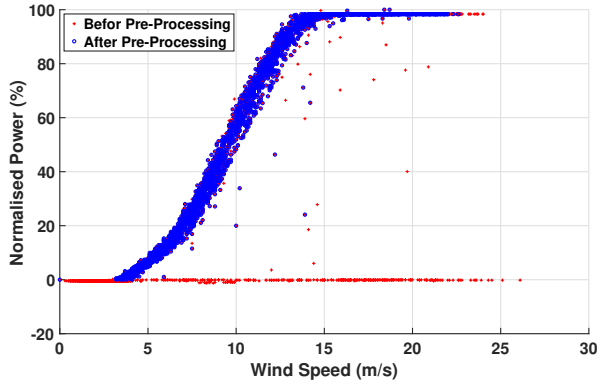


Figure 2. Power curve of wind turbine 1 before and after data pre-processing.

The power curve before and after pre-processing is shown in Figure 2.

2.3. Variable Exploration

This Section gives an insight into the SCADA variables for the different time period groups before the bearing failure, as explained in Section 2. The results presented are for wind turbine 3.

The relationship between the rotor and generator speed is depicted in Figure 3. The relationship is linear in the period that is considered healthy- as expected. However, the relationship between the two variables seems to change within 1 month before the component failure. It also seems that the generator speed exceeds its maximum expected value based on the input rotor speed and the gearbox ratio.

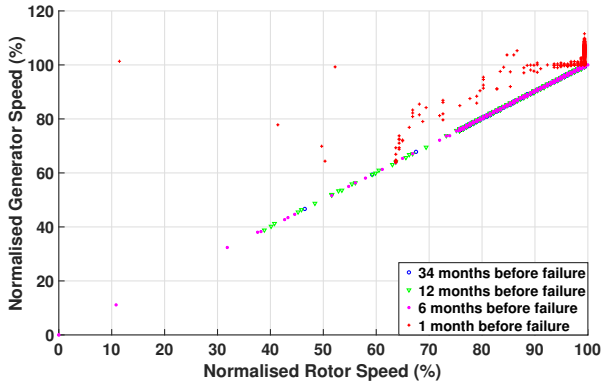


Figure 3. Generator speed as a function of rotor speed. The relationship should be linear, but changes within 1 month before the component failure.

.

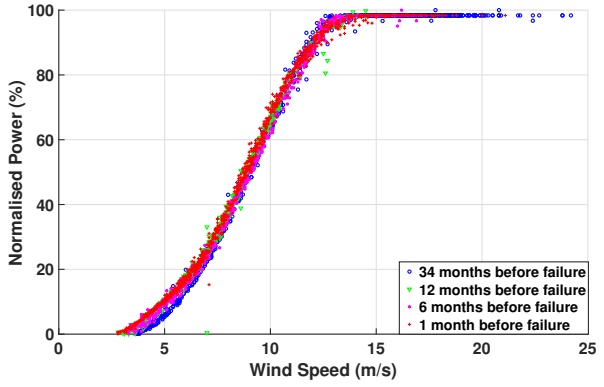


Figure 4. Power curve of wind turbine at different stages before failure. No significant changes are noticed between the different time periods before failure.

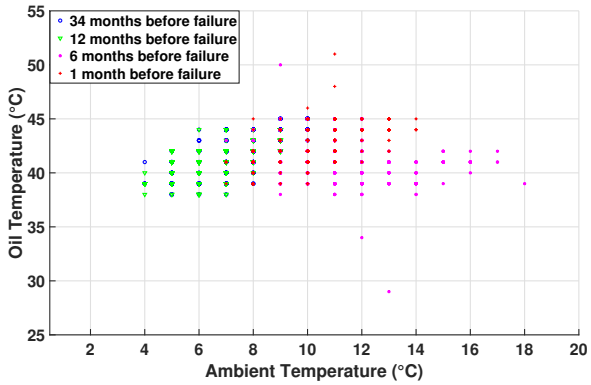


Figure 5. Power curve of wind turbine 1 before and after data pre-processing. No significant changes are noticed between the different time periods before failure.

The wind power curve and the relationship between the ambient and oil temperature are shown in Figures 4, 5. No significant changes are shown between healthy gearbox state and before failure state

3. ANOMALY DETECTION USING ROTOR AND GENERATOR SPEED

3.1. Choosing input parameters

The aim of the methodology is to develop a model that predicts key gearbox variable values and detects an anomaly, based on the error between the actual and the estimated value. The selection of the appropriate model parameters is vital in this process. The gearbox related input parameters can be selected using domain knowledge, as demonstrated in (Garcia et al., 2006), (Zaher et al., 2009).

The gearbox steps up the speed from the rotor to the generator. The speed of the generator F_{gen} is linearly proportional to the speed of the rotor F_{rot}, as shown in Eq. 1, where n_{gb} is the gearbox ratio.

$$F_{\text{gen}} = n_{\text{gb}} F_{\text{rot}} \tag{1}$$

The generator speed reading is calculated based on the frequency output of the generator. It was shown in Section 2 that there is a large deviation of the generator-rotor speed relationship from its expected linear behaviour, within one month before the planet bearing occurrence. A theory to explain this is that the fault can potentially affect the way that the generator speed is estimated, so there is some error introduced to generator speed measurement when the gearbox is unhealthy. The generator speed is calculated based on the electrical frequency of the converter, which is turned into slip and into generator rotational speed. This involves a frequency domain transformation, with the slip frequency changing as the machine changes speed. The explanation given by the

Output	Input
Generator Speed(t)	Rotor Speed(t)

Table 2. Prediction model parameters.

authors is that as the gearbox fault gets worse, the rotor frequency/slip/speed algorithm is incorrectly picking up a faulty harmonic, rather than the slip frequency. Further work is required to confirm this theory and to determine the cause of the generator rpm measurement error. Another less likely explanation is that there are some transmission errors, unbalanced load sharing at the planetary stage, or distorted frictional forces due to the fault, that lead to irregular transmissions between the planet and sun gears.

The inputs and outputs of the simple normal behaviour model are given in Table 2.

3.2. Regression Normal Behaviour Model

The correlation analysis between the generator and rotor speed indicates a linear relationship between them. Robust linear regression is performed, using the generator speed as dependent and the rotor speed as independent variable.

The training phase involves data that are considered to be in a healthy operating condition (34 months before the component failure). The testing dataset includes data that are in all available time periods before the component failure.

The rotor-generator speed relationship for the three turbines of the case study is shown in Figure 6.

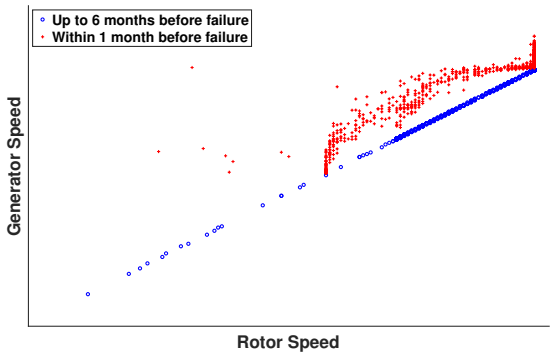
Results are shown in Figure 7. The generator speed is predicted fairly accurately (straight line between actual values and predictions) both in the validation phase and the testing phase up to approximately 6 months before the bearing failure. Within 1 month before failure the speed is underpredicted for the three turbines. The R^2 of the regression models is shown in Table 3 The mean daily absolute error for Turbine 1 is shown in 8 and it seems that in October, just a few weeks before the component failure, the error increases significantly.

3.3. Fault Detection

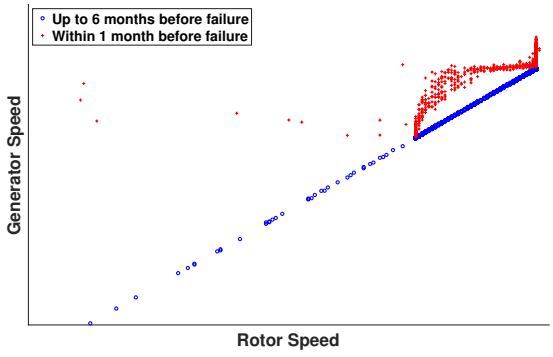
The performance of the normal behaviour model is assessed using the distribution of errors, which in a healthy operation should have a mean around zero. If an abnormality occurs, the behaviour prediction model should yield higher errors and therefore the mean will be shifted.

The two-sample t-test (Snedecor & William, 1989) is used to determine if two population means are equal. The variances of the two samples are assumed to be unequal.

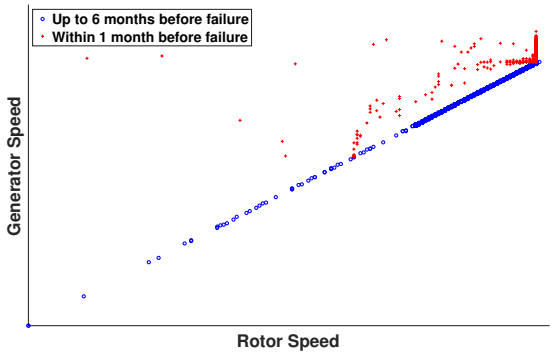
An 8 hour window with a 2 hour time step is used to calculate the mean errors. The t-test is defined as:



(a) Turbine 1

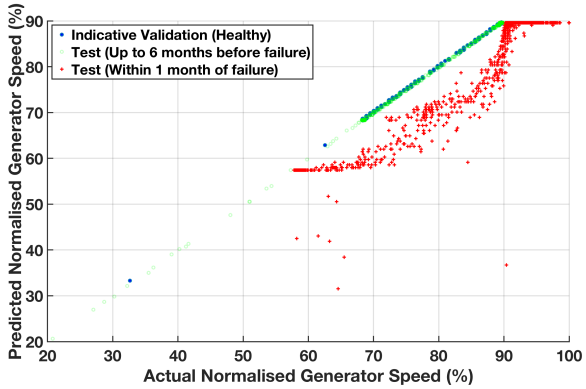


(b) Turbine 2

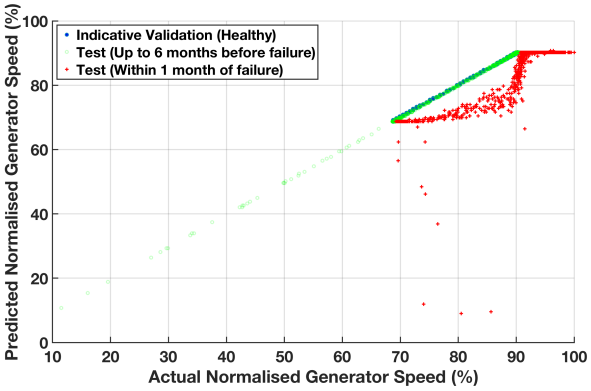


(c) Turbine 3

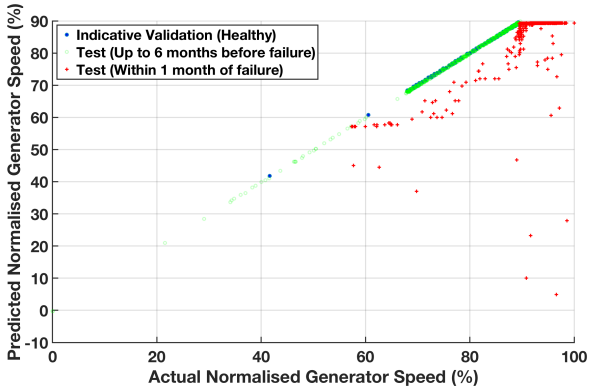
Figure 6. The relationship between generator speed and rotor speed is linear, as expected in normal operation. As the gearbox comes closer to failure the relationship becomes nonlinear.



(a) Turbine 1



(b) Turbine 2



(c) Turbine 3

Figure 7. Actual and Predicted Generator Speed. Within 1 month before failure the speed is underpredicted for the three turbines.

	Validation (Healthy)	Test (Healthy)	Test (Close to Failure)
Turbine 1	0.994	0.993	0.25
Turbine 2	0.992	0.99	0.31
Turbine 3	0.992	0.991	0.28

Table 3. R^2 of the regression models.

- H_0 if the data tested come from independent random samples from normal distributions with equal means.
- $H_{\text{alternative}}$ otherwise

The result h is 1 if the test rejects the null hypothesis and 0 if it accepts it. The significance level chosen is 1%.

The hypothesis test results are shown in Figure 9. The null hypothesis is rejected during the last month of the bearing operation since the regression error means of the training and testing sets differ by a large amount.

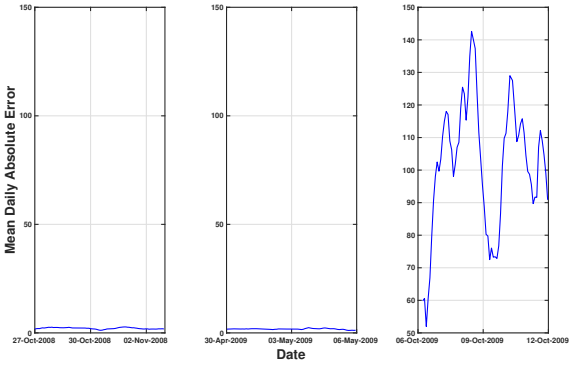


Figure 8. Mean daily absolute error of generator speed estimation. The error increases significantly close to failure.

.

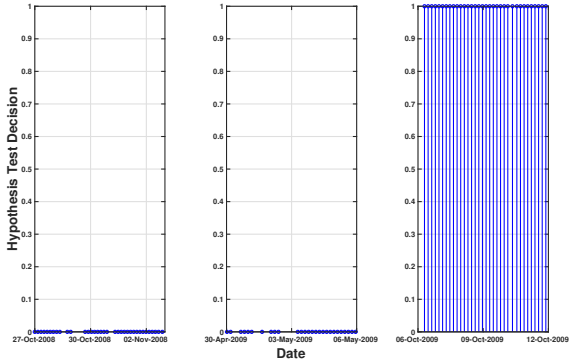


Figure 9. Hypothesis test results.

4. CLASSIFICATION USING ALL VARIABLES

This Section aims to apply supervised learning in order to classify the data into health states, according to the time before the failure. The SCADA data from all the three turbines are combined for this analysis.

4.1. Support Vector Machines

Support Vector Machines (SVMs) aim to create a hyperplane that separates input data in two classes. This can either be done linearly or through a non-linear kernel function. The mathematical formulation of SVMs can be found in (Vapnik, 2013). SVMs can also be used for multi-class classification (Weston, Watkins, et al., 1999).

In this paper a radial basis function kernel is used and one-vs-all for multi-class classification.

4.2. Classification Results

The inputs to the classifier are the SCADA measurements, as shown in Table 1. Multi-class classification is first applied according to the time window before the component failure, as explained in Section 2.1. A training/test ratio of 70%/30% is used. The results are shown in the confusion matrix in Figure 10.

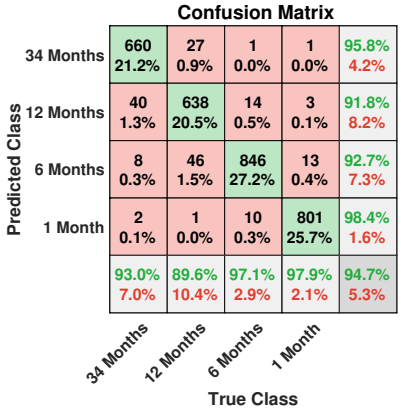


Figure 10. Multi-class classification confusion matrix.

Binary classification is also performed. According to Section 2.3, the behaviour of the generator speed starts to become abnormal within 1 month before the bearing failure.The grouping of the classes is shown in Table 4 and the results are presented in Figure 11.

	Two Class Model
Class 1	34 Months 1 Year 6 Months
Class 2	1 Month

Table 4. Binary model and class allocation.

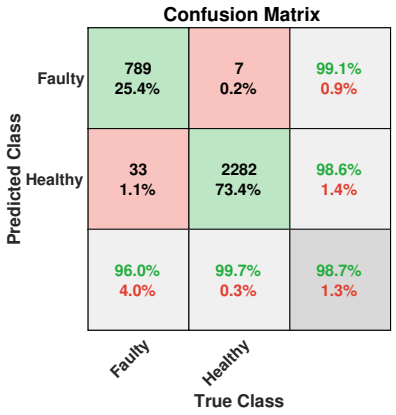


Figure 11. Confusion matrix for binary classifier.

The results indicate that a multi-class classification for the three wind turbines can be achieved and the time before the component failure can be predicted. This shows that the fault developed similarly in the three wind turbines. It also indicates that apart from the generator-rotor speed relationship which only changes within 1 month before the bearing failure, the relationship between all the SCADA variables changes gradually within the course of the 34 months leading up to failure.

A binary classification can indicate if a wind turbine is healthy or faulty, based again on the time before the bearing failure. The binary classification results show that during 1 month before failure the relationship between the SCADA variables is different than the rest of the operating time periods. Even though the samples collected for 1 month are much less than the other class, as depicted in Figure 11, a large percentage of the points within 1 month of the incipient fault are classified correctly. Some collective performance results are shown in Table 5.

	Precision	Recall
Multi-class	0.95	0.94
Two-class	0.98	0.97

Table 5. Performance Metrics of Classifiers

5. CONCLUSION

This paper investigated the use of SCADA data for wind turbine gearbox planet bearing fault prediction. The case study concerned historical data from three operating wind turbines leading up to the same planet bearing failure mode. An insight into the dataset was given for different time steps prior to the component failure. The generator speed and two gearbox temperatures were modelled in normal operating condition. Abnormalities can be detected through the error between the predicted and the actual variables. The results indicate that the relationship between the generator and rotor speed changes in the time period close to the fault. This could be related to measurement procedure of the generator speed. A classifier using all the measured variables is also presented and it indicates that the relationship between the SCADA variables changes within one month before the bearing catastrophic failure, as shown by the normal behaviour model. It is furthermore presented that potential prognosis of the fault can be achieved using enough run-to-failure examples.

ACKNOWLEDGMENTS

The authors would like to acknowledge EPSRC funding number EP/L016680/1 for the funding of this project.

REFERENCES

Bangalore, P., Letzgus, S., Karlsson, D., & Patriksson, M. (2017). An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox. *Wind Energy*, *20*(8), 1421–1438.

Catmull, S. (2011). Self-organising map based condition monitoring of wind turbines. In *Ewea annual conf* (Vol. 2011).

Feng, Y., Qiu, Y., Crabtree, C. J., Long, H., & Tavner, P. J. (2011). Use of scada and cms signals for failure detection and diagnosis of a wind turbine gearbox. In *European wind energy conference and exhibition 2011, ewec 2011* (pp. 17–19).

Feng, Y., Qiu, Y., Crabtree, C. J., Long, H., & Tavner, P. J. (2013). Monitoring wind turbine gearboxes. *Wind Energy*, *16*(5), 728–740.

Garcia, M. C., Sanz-Bobi, M. A., & del Pico, J. (2006). Simap: Intelligent system for predictive maintenance: Application to the health condition monitoring of a windturbine gearbox. *Computers in Industry*, *57*(6), 552–568.

Garlick, W. G., Dixon, R., & Watson, S. J. (2009). A model-based approach to wind turbine condition monitoring using scada data..

Kim, K., Parthasarathy, G., Uluyol, O., Foslien, W., Sheng, S., & Fleming, P. (2011). Use of scada data for failure detection in wind turbines. In *Asme 2011 5th international conference on energy sustainability* (pp. 2071–2079).

Kusiak, A., & Verma, A. (2013). Monitoring wind farms with performance curves. *IEEE Transactions on Sustainable Energy*, *4*(1), 192–199.

McArthur, S., Catterson, V., & McDonald, J. (2005). A multi-agent condition monitoring architecture to support transmission and distribution asset management.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321–352). Springer.

Schlechtingen, M., & Santos, I. F. (2011). Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical systems and signal processing*, *25*(5), 1849–1875.

Snedecor, G. W. C., & William, G. (1989). *Statistical methods/george w. snedecor and william g. cochran.* (Tech. Rep.).

Tarassenko, L., Nairac, A., Townsend, N., Buxton, I., & Cowley, P. (2000). Novelty detection for the identification of abnormalities. *International Journal of Systems Science*, *31*(11), 1427–1439.

Tautz-Weinert, J., & Watson, S. J. (2016). Using scada data for wind turbine condition monitoring–a review. *IET Renewable Power Generation*, *11*(4), 382–394.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Weston, J., Watkins, C., et al. (1999). Support vector machines for multi-class pattern recognition. In *Esann* (Vol. 99, pp. 219–224).

Wilkinson, M., Darnell, B., van Delft, T., & Harman, K. (2014). Comparison of methods for wind turbine condition monitoring with scada data. *IET Renewable Power Generation*, *8*(4), 390–397.

Yang, W., Jiang, J., et al. (2013). Wind turbine condition monitoring by the approach of scada data analysis. *Renewable Energy*, *53*, 365–376.

Zaher, A., McArthur, S., Infield, D., & Patel, Y. (2009). Online wind turbine fault detection through automated scada data analysis. *Wind Energy*, *12*(6), 574–593.

BIOGRAPHIES

Sofia Koukoura received her degree in Mechanical Engineering from the National Technical University of Athens in 2015. She then joined the University of Strathclyde Wind and Marine Energy Systems Centre for Doctoral Training. Her PhD focuses on diagnostics and prognostics of wind turbine gearboxes using signal processing and machine learning techniques.

James Carroll got his PhD in October 2016 from the University of Strathclyde Wind Energy Systems Centre for Doctoral Training. He received the EPSRC doctoral prize for his postdoctoral studies and is now a Lecturer within the department of Electronic and Electrical Engineering at the University of Strathclyde. His research interests include cost of energy modelling, failure rates and condition monitoring.

Alasdair McDonald is a Senior Lecturer at the EPSRC Wind and Marine Energy Systems Centre for Doctoral Training based in the Institute for Energy and Environment, Department of Electronic Electrical Engineering, University of Strathclyde. His research interests are focussed on the modelling and design of electrical generators and powertrains and their application to renewable energy, especially offshore wind turbines.