Performance Analysis and Anomaly Detection in Wind Turbines based on Neural Networks and Principal Component Analysis

Peyman Mazidi, *Student Member, IEEE,* Lina Bertling Tjernberg, *Senior Member, IEEE,* and Miguel A. Sanz-Bobi, *Senior Member, IEEE*

***Abstract*—This paper proposes an approach for maintenance management of wind turbines based on their life. The proposed approach uses performance analysis and anomaly detection (PAAD) which can detect anomalies and point out the origin of the detected anomalies. This PAAD algorithm utilizes neural network (NN) technique in order to detect anomalies in the performance of the wind turbine (system layer), and then applies principal component analysis (PCA) technique to uncover the root of the detected anomalies (component layer). To validate the accuracy of the proposed algorithm, SCADA data obtained from online condition monitoring of a wind turbine are utilized. The results demonstrate that the proposed PAAD algorithm has the capability of exposing the cause of the anomalies. Reducing time and cost of maintenance and increasing availability and in return profits in form of savings are some of the benefits of the proposed PAAD algorithm.**

***Index Terms*—Condition Monitoring, Fault Detection, Maintenance, Neural Networks, Performance Analysis, Principal Component Analysis, Wind Power Generation**

I. INTRODUCTION

ENVIRONMENT friendly and cost effectiveness are two of the factors that have driven societies towards vast utilization of wind energy by large investment and deployment of wind turbines and wind farms in the past few decades. With being able to effectively operate and produce power for up to 25 years [1], some of the early installed wind turbines are approaching their end of life. This aging matter is bringing up more frequent incidents (failures, malfunctions etc.) and longer unavailability which raise the alarm that a more developed monitoring system is required to address these issues.

Condition monitoring systems (CMS) on modern turbines use the data from supervisory control and data acquisition (SCADA) systems where they record the data of turbines through different sensors installed on them. An accurate CMS does not necessarily require large numbers of sensors (hence a costlier system), rather a system that provides the owner of the asset with enough information to avoid anomalies and failures with an acceptable accuracy [2], [3]. Moreover, including operational conditions in the analysis can help improve the maintenance management [4].

Peyman Mazidi is with KTH Royal Institute of Technology, Stockholm, SE-10044 Sweden and Comillas Pontifical University, Madrid, 28015 Spain (e-mail: mazidi@kth.se).

Lina B. Tjernberg is with KTH Royal Institute of Technology, Stockholm, SE-10044 Sweden (e-mail: linab@kth.se).

Miguel A. Sanz-Bobi is with Comillas Pontifical University, Madrid, 28015 Spain (e-mail: masanz@comillas.edu).

Recently, there have been several studies that benefit from SCADA data and they design CMSs for wind turbines. Yang et al. [5] investigated the processing of the SCADA data and raised some concerns on the analysis of these data as they had large variations and this makes it difficult to detect incipient failures. There have also been a few research works where they review applied techniques on condition monitoring and fault diagnosis of wind turbines [3], [6]–[11]. Ultimately, all these works suggest that the need for developing an algorithm that can manage a broad area is yet to be achieved. Furthermore, most of the approaches and techniques applied are focused on a particular subassembly (a component, e.g. generator, pitch system etc.) in the wind turbine.

There have been several research works focusing on reliability and maintenance of wind turbines considering them as a whole or in subassemblies [12]–[19] and point out the importance of standardized data and their possible impacts which provided the basis for this work. For instance, some works focus particularly on pitch system in the wind turbine [20]–[24].

The techniques utilized must be able to model normal behavior of the component in order to detect anomalies. In reality, this often requires modelling nonlinear relationships among several variables (internal and external) that characterize the behavior of an industrial component. Neural network (NN) is one of the techniques widely used in the analysis of wind turbines mainly due to its flexibility and adaptability to various conditions and it offers the possibility of modelling these nonlinear relationships. Its applications range from behavior prediction to classification and fault diagnosis and have been applied to different subassemblies in a wind turbine [25]–[28], e.g. gearbox [29]–[31], bearings [32], [33] and generators [34]. Adaptive neuro fuzzy inference system was applied to find malfunctions in the wind turbine [35], [36]. Since the results are the outcomes of pure statistics without interference of human knowledge and information, an expert manually implements a number of rules to narrow down the possible origin of the failure. Lie et al. [37] presented an improved fuzzy-synthetic assessment method and applied it to generator of a wind turbine.

There are several areas to improve in order to obtain a thorough CMS. The areas include sensors and data acquisition [38], data processing [39], behavior and performance modelling [40], abnormality and anomaly extraction and evaluation, diagnosis and prognosis, and interactive and integrated proce-

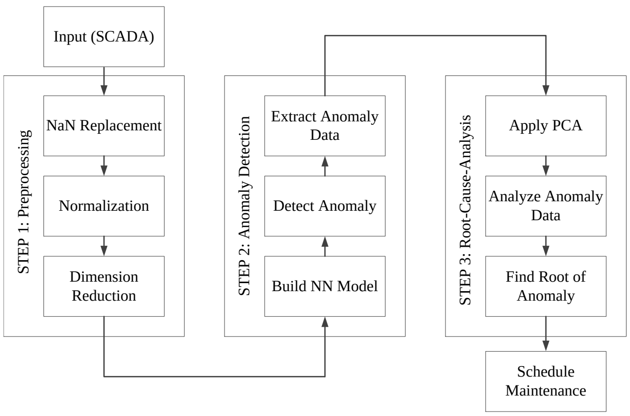

Fig. 1. The proposed PAAD algorithm

dures etc. Since they cover a wide area, each individual work can focus on specific parts. This work provides a link between system layer and component layer which makes it an overall condition monitoring technique for wind turbines.

The work in this paper presents a data-driven generic performance analysis and anomaly detection (PAAD) algorithm (Figure 1) with automation goal. One of the resultant advantages of this approach is decrease in maintenance downtime. Since this approach is able to point out the component (subassembly) which is responsible for the abnormality, the maintenance crew will not need to search the entire system (e.g. wind turbine) to find the faulty component; rather, they can directly hover over the defective part and perform the maintenance [41]. The other connected advantage is increase in availability of the system and hence, higher savings and efficiency.

II. PAAD ALGORITHM

A. Step1: Preprocessing

Measurements recorded by the SCADA system create a large raw dataset. Normally, these data are not in a format to be directly utilized for any purpose. The measurements are recorded for each parameter at frequencies less than 10 minute; however, one average number for these values is provided as the final value, over the 10-minute period.

Analyzing the large recorded data sets can be complex, time-consuming, inefficient and in cases, impossible due to limitations of computational power. For these reasons, several techniques such as Clustering, Scatter plot, Confidence intervals and Pearson Correlation (PC) [42], equation (1). Since statistical techniques lack physical knowledge, consideration of cause-and-effect relationships will improve the analysis.

The information provided by PC technique only show whether two variables are linearly related. Although this is limited, it can present elementary information which at the preliminary stage of the analysis is advantageous. The types of information that PC technique offers are as follows [43]:

- Sign of the coefficient: While + sign in the output coefficient means a linearly positive relationship among the two variables, - sign means a linearly negative relationship. 0 shows that there is no linear relationship.
- Magnitude of the coefficient: While values in the range of [0.1, 0.3] mean small correlation between the variables, values in the range of [0.3, 0.5] mean medium correlation level exist. Values larger than 0.5 show that there is a significant correlation.

$$pc_{x,y} = \frac{\sum_{i=1}^{n}(x_i - x')(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2}\sqrt{\sum_{i=1}^{n}(y_i - y')^2}}, \qquad (1)$$

where $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$ are the two of input parameter vectors to be compared, x' and y' are mean values of their corresponding vectors and n is the number of observations.

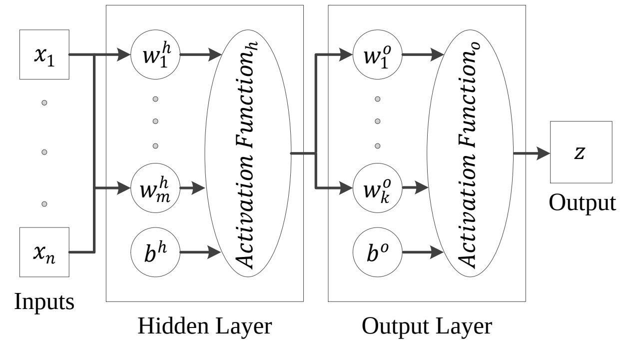

Fig. 2. Multilayer Perceptron Feed-Forward Neural Network Structure

B. Step2: Anomaly Detection

Analysis of the performance of the monitored wind turbine is performed through its power-curve. For this purpose, a normal behavior model is created that the inputs of the model are real-time inputs of the wind turbine (e.g. wind speed, ambient temperature etc.) and information about subassemblies (e.g. gearbox temperature, oil temperature etc.). The output of the model, which is an expected power-curve, is created. Then, the output of the model is compared with the real-time measured output power of the wind turbine. Whenever there are discrepancies between these two values, an anomaly is flagged. It should be noted that a tolerance level needs to be defined for the deviations of the model's output from the measured power. This threshold can be obtained through clustering techniques, defining a confidence interval or from experience and through several trials with the model. In this paper, NN technique (multilayer perceptron feed-forward, Figure 2) is utilized to build the normal behavior model for the wind turbine.

The first phase in building the NN model is to find the optimum configuration for the NN. To find the best configuration for the NN, various possible combinations should be tested. Different performance evaluation methods can be used in the training of the NN and in this study these four methods were applied: mean absolute error (MAE) equation (2), mean squared error (MSE) equation (3), sum absolute error (SAE) equation (4) and sum squared error (SSE) equation (5).

$$MAE = \frac{1}{k}\sum_{j=1}^{k} \mid z_i - z'_i \mid, \tag{2}$$

$$MSE = \frac{1}{k}\sum_{j=1}^{k} (z_j - z'_j)^2, \tag{3}$$

$$SAE = \sum_{j=1}^{k} \mid z_j - z'_j \mid, \tag{4}$$

$$SSE = \sum_{j=1}^{k} (z_j - z'_j)^2, \tag{5}$$

where z_j and z'_j are NN output and target values and k is the number of data points.

As shown in Figure 2, structure of the NN utilized in this work constitutes one hidden layer and one output layer where each has a different activation function.

$$ActivationFunction_1 = f_1(w_i^h x_i + b^h), \tag{6}$$

$$ActivationFunction_2 = f_2(w_i^o v_i^h + b^o), \tag{7}$$

where w_i^h are weights of the neurons in hidden layer, b^h is the bias value of hidden layer, w_i^o are weights of the neurons in output layer, b^h is the bias value of output layer and v_i are outputs of hidden layer to output layer.

Activation Function1 (f_1) calculates outputs of the hidden layer based on Sigmoid function with Hyperbolic Tangent form as shown in equation (8) [44]:

$$f_1(w_i^h x_i + b^h) = \frac{2}{1 + e^{-2(w_i^h x_i + b^h)}} - 1, \tag{8}$$

and Activation Function2 (f_2) estimates outcome of the output layer based on a linear function as shown in equation (9):

$$f_2(w_i^o v_i^h + b^o) = w_i^o v_i^h + b^o. \tag{9}$$

A powerful supervised learning algorithm, Scaled Conjugate Gradient is applied in order to train the NN [45]. Based on the chosen performance function, the algorithm's objective is to minimize the function by tuning biases and weights of the NN.

C. Step3: Root Cause Analysis

In multivariate statistical analysis, principal component analysis (PCA) is one of the most important techniques to deal with correlated data (in the wind turbine case, generated active power, wind speed, rotor speed and generators oil temperature are some of the correlated measured variables). PCA transforms these correlated data into uncorrelated linear data by defining a new coordination system (in this case, the values of generated active power, wind speed, rotor speed and other variables are transferred to a new coordination system so that these variables lose their correlation and become linearly uncorrelated). These new resultant variables are called Principal Components and they aim to capture highest levels of change and variability in the data; consequently, first principal component bears the highest impact in the data [43], [46].

After the anomalies are detected in the system, the data are sent to be further analyzed by PCA. PCA primarily says how many of the input variables are sufficient to model the system without losing any characteristic. The inputs of PCA are the extracted data points that were considered to be anomalies by the NN model in the previous step.

At first, PCA calculates the covariance matrix for the data. For an n dimensional data, the covariance matrix is as shown in equation (10) where $cov(x, y)$ is the covariance between x and y calculated as equation (11) with n data points.

$$[C]_{n\times n} = [cov(x, y)], \tag{10}$$

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - X')(Y_i - Y'). \tag{11}$$

Then PCA extracts the eigenvectors and eigenvalues of the covariance matrix. Afterwards, principal component is defined as the eigenvector that has the largest eigenvalue. A feature vector, a matrix of vectors, is then created from these calculated eigenvectors. One could consider only those vectors that correspond to highest eigenvalues, or all in case would like to maintain complete data. By multiplying the feature vector into the transposed of original data, final PCA outcomes in terms of new principal components (eigenvectors) [47].

Here, four outputs of PCA that can be used are component coefficients, component scores, latent and explained. Components coefficients give the coefficients of the principal components and component score gives the original data in the new coordinate system. The latent output gives the variance of the newly defined principal components and the explained output provides information on the percentage of these variances.

III. Application of PAAD Algorithm

The case study comprises recorded data of 62 signals from SCADA system of a wind turbine. The data consist of measurements the period of approximately 22 months, henceforward named WindMSP1 and WindMSP2 data sets. The only difference between WindMSP1 and WindMSP2 is that in WindMSP2, the wind turbine experienced some abnormalities in its performance and WindMSP2 holds the measurements when the abnormalities occurred. Thus, WindMSP1 is a smaller set in WindMSP2 when no anomaly was observed during operating time.

A. Step1: Preprocessing

1) NaNs Replacement: Since dismissing the measurements can have significant impacts in modeling and analysis, the NaNs were replaced. Available values before and after the NaN point were located and the NaN value was replaced with an average value of these two available values. In this way, if an abnormality happened for some period of time, the resultant variation can be reflected in the data by the difference between the values that were measured before and after that abnormality time.

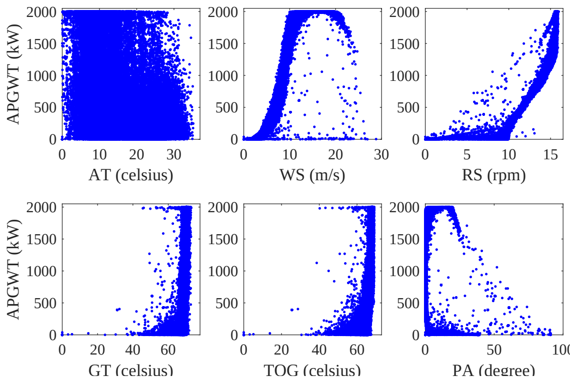

Fig. 3. Mutual behavior of seven parameters in WindMSP1

2) Dimensionality Reduction: Table I presents a complete list of the recorded signals by the SCADA system. Since some parameters do not change based on the power or their change does not cause any specific variation to the output power of wind turbine, there will be a reduction in the number of parameters for modeling. It should also be noted that the cause and effect relationship should be taken into account. The seven primarily chosen parameters are active power generated by wind turbine (APGWT), pitch angle (PA), rotor speed (RS), ambient temperature (AT), gearbox temperature (GT), temperature of oil in gearbox (TOG) and wind speed (WS). These parameters were selected solely from experience.

TABLE I
LIST OF RECORDED SIGNALS BY THE SCADA SYSTEM

#	Parameter (unit)	#	Parameter (unit)
1	Year	32	GROUND RFC 1 Version
2	Month	33	GROUND RFC 2 Version
3	Day	34	GROUND RFC 3 Version
4	Hour	35	GROUND ILC Version
5	Minute	36	GROUND RFC Version
6	Second	37	Hydraulic Group Pressure (bar)
7	Wind Direction	38	Hydraulic Group Temp
8	Ambient Temp. (C)	39	Hydraulic Valve Output Voltage
9	Ice Detection Temp.	40	Gearbox Temp. (C)
10	Wind Speed (m/s)	41	Gearbox Oil Temp. (C)
11	Elect. Generator Ring Temp. (C)	42	Year ProdPower
12	Elect. Generator LA Bearing Temp. (C)	43	Day ProdPower
13	Elect. Generator LOA Bearing Temp. (C)	44	Hour ProdPower
14	Power Factor Set-Point	45	Month ProdPower
15	Elect. Generator Alarm Temp. (C)	46	Producible Power (W)
16	Generated Active Power (kW)	47	Total ProdPower 1
17	Generated Total Power (kW)	48	Network Frequency
18	Elect. Generator Stator Power (kW)	49	Network Power
19	Generated Reactive Power (kVAR)	50	Network Reactive Power
20	Elect. Generator Rotor Power (kW)	51	Network Voltage (V)
21	Elect. Generator Winding 1 Temp. (C)	52	Pitch Angle (deg)
22	Elect. Generator Winding 2 Temp. (C)	53	Rotor Iced Possibility
23	Elect. Generator Winding 3 Temp. (C)	54	Rotor Speed (rpm)
24	Elect. Generator Speed TOP (rpm)	55	Stopped By Tool
25	Mechanical Over Speed (rpm)	56	Ambient Temp TOP (C)
26	Nacelle Position	57	Elect. Transformer Max Temp.
27	Yaw Brake Pressure	58	Elect. Transformer 1 Temp.
28	Nacelle Temp. (C)	59	Elect. Transformer 2 Temp.
29	Tower Height	60	Elect. Transformer 3 Temp.
30	Elect. Generator Speed GROUND (rpm)	61	High Speed Detection
31	GROUND Version	62	Wind Turbine State

One technique is scatter plot of parameters. The idea is simply to look at the graphs and see if some particular relationship (linear or nonlinear) can be observed with naked eye. If no pattern seems to exist, it would be advisable to dismiss that parameter. Figure 3 displays the relationship between the primarily selected parameters and the APGWT.

Table II displays the Pearson Correlation between a few variables and APGWT. WS variable shows the highest linear relationship to APGWT. As shown, Nacelle Temperature does not represent any linear relation.

TABLE II
RESULTS OF PEARSON CORRELATION ANALYSIS FOR WINDMSP1

Parameter	APGWT
Ambient Temperature	- 0.416
Nacelle Temperature	- 0.081
Pitch Angle	- 0.207
Gearbox Temperature	+ 0.363
Temperature of Oil in Gearbox	+ 0.392
Rotor Speed	+ 0.728
Wind Speed	+ 0.876

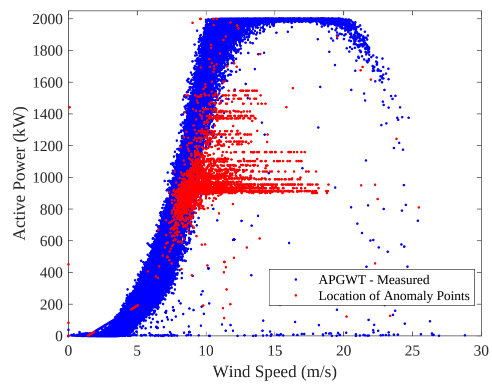

Fig. 4. Detected anomaly points in WindMSP2

3) Normalization: After parameters in building the model are selected and the NaNs are substituted, the data are scaled and normalized. Normalization converts original data into a newly defined range. This indeed matters when industry would like to maintain privacy of its asset's data. The normalization range for each parameter is: APGWT:[0,2000], PA:[0,90], RS:[0,16], AT:[0,35], GT:[0,80], TOG:[0, 70] and WS:[0,30]. The unit of each parameter is [kilowatt], [degree], [rpm], [celsius], [celsius], [celsius] and [celsius], respectively.

B. Step2: Anomaly Detection

1) Building NN: Inputs of the NN are PA, RS, AT, GT, TOG and WS, and APGWT is the target input. Testing various configurations has resulted in creation of 60 different NNs. Two layers, one hidden layer and one output layer, were selected for the NN. It should be mentioned the data in Training, Validation and Testing are divided randomly and Training Ratio of 70%, Validation Ratio of 15% and Test Ratio of 15% have been applied.

2) NN Training Outcomes: Table III shows the training results of the NNs. The PA parameter exhibits low level of accuracy when it is used as a single input, with the highest level of accuracy at 74%. This is due to the fact that the training of the NN tries to familiarize the model with all various available behaviors in the data and large range of variations in PA makes such task very challenging. After changing other parameters, considering PA as a single input, the maximum achieved improvement in the accuracy was

TABLE III
VARIOUS TESTED NEURAL NETWORKS FOR BUILDING THE MODEL USING WINDMSP1

Input					Output						
								Accuracy (%)			
NN	Input	Target Output	Hidden Layer	Performance Evaluation Method	Performance	Iteration	Duration (MM:SS)	Training	Validation	Test	All
1	PA	APGWT	10	MAE	226	212	00:11	67.728	67.813	65.358	67.391
2				MSE	199696	92	00:06	71.290	73.254	70.705	71.491
3				SAE	20515095	370	00:18	69.837	71.592	71.168	70.301
4				SSE	18524265462	56	00:10	71.401	72.023	71.819	71.559
5			20	MAE	214	435	01:04	70.128	70.527	69.936	70.161
6				MSE	191101	65	00:06	72.989	72.301	73.407	72.947
7				SAE	19921847	195	00:12	70.639	71.097	71.128	70.782
8				SSE	17729273653	55	00:05	72.964	73.193	73.035	73.007
9			50	MAE	209	692	02:01	70.825	70.098	69.998	70.595
10				MSE	186981	89	00:12	73.915	72.862	73.112	73.636
11				SAE	18365149	952	02:35	72.017	71.675	72.297	72.007
12				SSE	17345756035	46	00:07	73.423	74.382	74.311	73.701
13	TOG	APGWT	10	MAE	216	414	00:20	0.85383	84.810	85.192	85.270
14				MSE	108156	73	00:13	85.650	85.801	86.126	85.745
15				SAE	19860487	68	00:09	85.529	85.490	85.203	85.472
16				SSE	10046781727	21	00:03	85.867	85.881	85.080	85.753
17			20	MAE	203	273	00:17	86.596	86.589	86.198	86.536
18				MSE	97885	78	00:13	87.143	87.047	87.606	87.199
19				SAE	18284890	111	00:24	87.066	87.066	87.126	87.075
20				SSE	9049578417	156	00:29	87.331	87.009	87.243	87.271
21			50	MAE	198	550	01:50	86.573	86.822	86.347	86.576
22				MSE	100630	18	00:04	86.875	86.652	86.738	86.820
23				SAE	18243110	207	00:52	86.929	86.592	86.493	86.814
24				SSE	9065135049	65	00:27	87.305	86.737	87.481	87.248
25	WS	APGWT	10	MAE	69	74	00:14	94.292	94.269	95.011	94.395
26				MSE	44137	127	00:14	94.337	94.499	94.873	94.443
27				SAE	6382392	7	00:01	94.387	93.841	94.969	94.395
28				SSE	4099235357	16	00:02	94.465	94.309	94.500	94.447
29			20	MAE	69	207	00:44	94.283	94.978	94.434	94.408
30				MSE	44156	59	00:10	94.520	94.185	94.323	94.440
31				SAE	6365863	369	01:01	94.376	94.229	94.728	94.406
32				SSE	4100660843	11	00:02	94.630	93.638	94.385	94.446
33			50	MAE	69	189	01:02	94.375	94.314	94.640	94.404
34				MSE	44270	126	00:23	94.428	94.656	94.176	94.425
35				SAE	6374251	11	00:05	94.315	94.788	94.433	94.404
36				SSE	4103518408	22	00:04	94.617	93.807	94.251	94.441
37	WS + PA	APGWT	10	MAE	50	58	00:09	99.109	99.079	99.079	99.100
38				MSE	7230	120	00:07	99.123	99.047	99.118	99.111
39				SAE	4604124	6	00:01	99.101	99.085	99.107	99.100
40				SSE	638644505	79	00:05	99.155	99.167	99.146	99.155
41			20	MAE	49	499	00:40	98.970	99.084	98.894	98.976
42				MSE	5764	313	00:20	99.287	99.296	99.310	99.292
43				SAE	4094790	34	00:05	99.302	99.258	99.273	99.291
44				SSE	534570291	18	00:03	99.302	99.252	99.288	99.293
45			50	MAE	76	347	01:29	96.439	96.500	96.449	96.449
46				MSE	5757	232	01:17	99.289	99.309	99.296	99.293
47				SAE	4071580	77	00:19	99.306	99.258	99.258	99.291
48				SSE	533523088	21	00:06	99.293	99.339	99.259	99.295
49	WS + PA + TOG + RS + AT + GT	APGWT	10	MAE	38	862	01:40	99.487	99.495	99.467	99.485
50				MSE	2885	164	00:10	99.647	99.636	99.652	99.646
51				SAE	3356199	55	00:07	99.645	99.642	99.646	99.645
52				SSE	254290172	79	00:11	99.664	99.674	99.660	99.665
53			20	MAE	40	498	00:38	99.443	99.428	99.483	99.447
54				MSE	3517	81	00:11	99.574	99.572	99.538	99.569
55				SAE	3556626	103	00:27	99.568	99.539	99.545	99.561
56				SSE	258631853	64	00:17	99.662	99.649	99.656	99.659
57			50	MAE	47	463	02:18	99.194	99.229	99.187	99.198
58				MSE	2763	164	00:40	99.661	99.660	99.666	99.662
59				SAE	2914994	784	02:49	99.717	99.721	99.712	99.717
60				SSE	202990237	43	00:17	99.733	99.732	99.730	99.732

7% (74%-67%), thus, it is preferable to use more than one parameter to account for relation among parameters as well.

From performance point of view, NN49 obtained the lowest value with MAE method (38), NN58 obtained the lowest value with MSE method (2763), NN59 obtained the lowest value with SAE method (2914994) and NN60 obtained the lowest value with SSE method (202990237). Considering iterations, NN39 achieved the lowest iteration number of 6 while NN11 has the highest iteration number of 952. Interestingly, none of these two NNs has the lowest error in performance. On the time duration of the operation, while NN39 has the lowest running time (00:01), NN59 has the largest time (02:49). It can be seen that MAE and SAE resulted in high number of iterations and large running times more frequently than other performance evaluation methods. From the accuracy perspective, NN60 outperformed all other networks. After considering all the discussed points, NN60 was chosen as the optimum NN.

3) Verification of NN Outcomes: WindMSP2 dataset is utilized in order to verify the accuracy of the model in anomaly operating times. Figure 4 plots anomaly data points detected by NN after applying a threshold in the model for WindMSP2. It can be seen while the WS is at range [10:20], the APGWT does not behave optimally and produces power below the expected value [1000:1500]. In this work the threshold was obtained through several trial-and-errors as well as cross-validating the input data and the outcome. The threshold level was defined as 200 units which means only those points with higher values showed significant impact and therefore were stored. This threshold is applied on the difference between predicted and measurement data. Figure 4 verifies that the level of the defined threshold is acceptable as it detects anomaly points with high accuracy. In additional, it is expected that the threshold level for each wind turbine (and each failure mode) to be different.

C. Step3: Root Cause Analysis

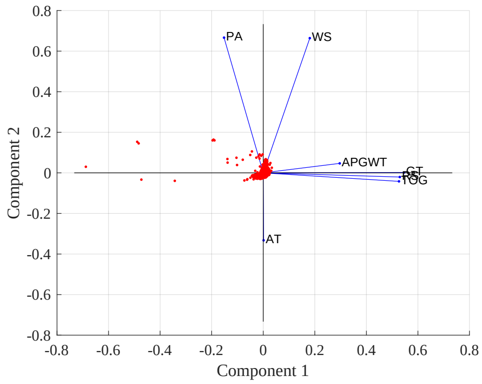

Fig. 5. Anomaly data points analyzed by PCA from WindMSP2

1) PCA Analysis: It should be remembered that after WindMSP2 is analyzed, a new behavior, which origins from the abnormality, is observed. This abnormal behavior is the consequence of irregular performance of one (or more) of the input variables. As the first principal component in the PCA contains the highest variance in the data, by comparing these outputs with the outputs resulting from WindMSP1, the defective subassembly with abnormal behavior can be uncovered.

2) PCA Outcomes: Table IV displays PCA results for the anomaly data points extracted by the NN model from WindMSP2. As it can be seen, PCA has created seven principal components. The most important observation from Table IV is the behavior of the anomaly data points which shows that the first principal component covers 41% of the variances in the anomaly data. Thus, this principal component could be the major reason for the anomalies.

Figure 5 visualizes the variables in a plot for 3000 anomaly data points extracted by the NN in WindMSP2. The plot is drawn from the Coefficients and Scores that are calculated for each of the variables and their corresponding principal components. They were scaled to the maximum Score value and maximum Coefficient length. For example, the largest coefficients in the second principal component (Component2) correspond to the PA and WS variables; and for the first principal component (Component1), PA has the lowest score in the negative direction. GT, TOG, RS and APGWT variables are closely grouped together because they show a coherent behavior which is a significant observation. AT also obtains the lowest value by the second principal component (Component2).

Knowing that principal components 1 and 2 together account for more than 65% of the variance in the data (Table IV), the reference for analyzing the results should be based on these two principal components. From statistical perspective, after analyzing WindMSP2, the final conclusion is that WS, PA and AT are the three variables causing irregularity in the performance of the wind turbine.

D. Results

NN accurately detected anomaly data points in WindMSP2 and statistical results concluded that PA, WS and AT are the main factors for the observed abnormality. Figure 6 displays the histogram of the WS values during the anomaly time. It is evident that the anomalies occurred when the wind speed had a normal profile with average WS of about 10 (m/s). Similarly, the AT behavior during this time is plotted in Figure 7 which shows a normal behavior as well. It should also be remembered that these data points are not sequential. As a conclusion, WS and AT can be disregarded as the causes of detected anomalies.

The next variable flagged by PCA is PA. To verify this as the final conclusion, further investigation is carried out. Figure 6 shows that the anomalies happened when the WS varied mainly between the values of 9 and 16. Figure 8 displays the performance of the wind turbine for the normal data (WindMSP1) and anomaly data in (WindMSP2) over the WS range of 9 and 16.

TABLE IV
RESULTS OF PCA ANALYSIS FOR ANOMALY POINTS BY NN IN WINDMSP2

	Latent	Explained
PC1	2,873	41,041
PC2	1,720	24,568
PC3	1,060	15,141
PC4	0,878	12,545
PC5	0,367	5,236
PC6	0,056	0,802
PC7	0,047	0,666

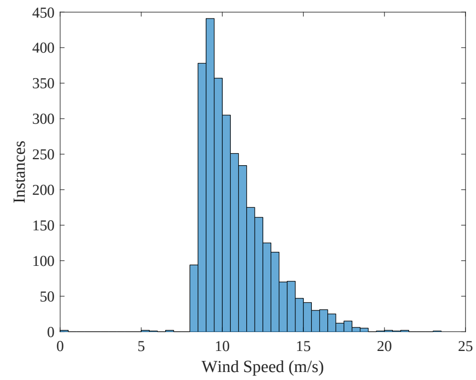

Fig. 6. WS for anomaly flagged data points in WindMSP2

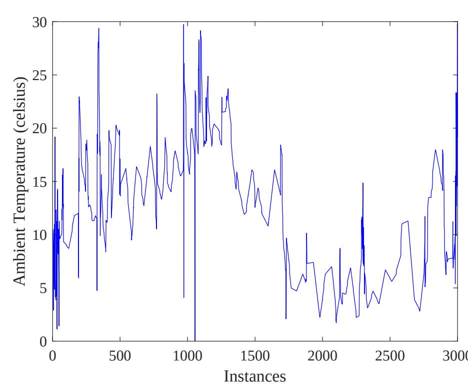

Fig. 7. AT for anomaly flagged data points in WindMSP2

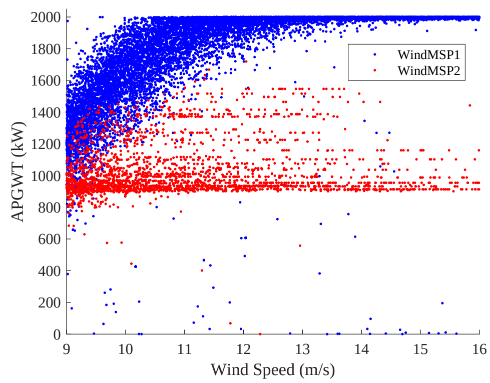

Fig. 8. Power-curve for a particular WS range

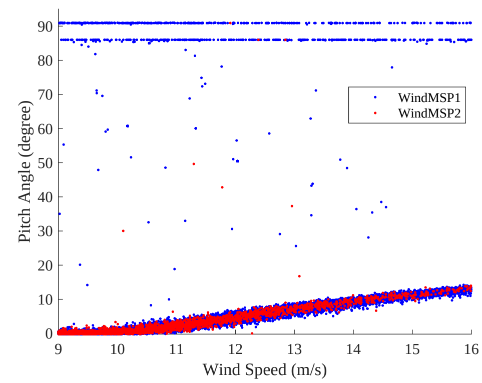

Fig. 9. PA versus WS

One point to mention here is that Figure 8 demonstrates how the model accurately detected the anomaly points. The anomaly detection part has managed to correctly discover an under performance situation in the wind turbine which is very beneficial for the asset owner and the maintenance management. From here, they can decide on a preventive maintenance action prior to an actual failure which brings about a significant impact for the system.

To review the analysis so far, Figure 8 shows for wind speeds above 11 (m/s), the expected power should be higher than the recorded power and the anomaly has been detected accurately. Then, PCA suggested the reason for this underperformance is related to probably WS, AT and PA. Analyzing the data of WS and AT confirmed these two parameters could not be the cause of the observed situation.

Pitch system in the wind turbine operates through a controller. This controller fits the angle of the pitch to the wind conditions based on the orientation of the wind turbine. Therefore, in analysis of PA variable, it must be remembered not only the pitch angle is considered, but also other components in the pitch system (e.g. the controller) are accounted for. This is an extremely important point to be taken into account. Considering the PA parameter can physically be origin of the detected anomaly and PCA analysis has also flagged PA as the major parameter, behavior of the PA during normal operation and abnormality (in both data sets) is compared in Figure 9.

The abnormality shown in Figure 8 and Figure 9 can be interpreted as follows: Although the wind speed is in normal condition and follows expected behavior, the measured output power is not following the correct path (many PA points with 86 and 90 degrees). Considering the structure of the pitch system, the only remained diagnosis is that a component (e.g. controller) in the pitch system is causing these anomalies (e.g. reading/sending incorrect signals). This is one of the points in designing a control algorithm in active pitch control as well [48]. Final practical verification of the result was indeed achieved after discussing the results with the wind turbine owner and the owner claimed this abnormal behavior was due to the testing of new controllers in the pitch system which is in conformance with the result of the proposed PAAD algorithm.

IV. Conclusion

This paper proposes an algorithm for performance analysis and anomaly detection. The algorithm utilizes the data obtained from SCADA system of a wind turbine and is divided into three steps: preprocessing of the raw input data, anomaly detection and root-cause analysis. In the first part of the algorithm, raw input data from SCADA system is processed and the dimensionality of the data is reduced. A NN model is built in the second step. Detailed phases in building the NN are described by evaluating several important factors. The created NN model detects anomalies through monitoring the performance of the wind turbine from its power-curve and extracts the detected anomaly data points. In the final step, PCA technique analyzes the data in order to discover the origin of the anomalies by exploring the information carried through the data. To verify the model, real data have been examined and the steps to apply and utilize the algorithm are explained in detail. The results show that the algorithm is capable of evaluating the performance of a wind turbine very accurately and can also assist in discovering the root of anomalies. The benefits of applying this algorithm are reductions in maintenance time and cost and increase in availability and in return profit of the wind turbine.

Acknowledgment

The authors would like to acknowledge the Erasmus Mundus Joint Doctorate (EMJD) Fellowship in Sustainable Energy Technologies and Strategies (SETS) program funded by the European Commission's Directorate-General for Education and Culture where provided the opportunity for carrying out this work.

References

[1] I. Staffell and R. Green, "How does wind farm performance decline with age?" *Renewable Energy*, vol. 66, p. 775786, June 2014.

[2] R. de Andrade Vieira, M. A. S. Bobi, and S. Kato, "Wind turbine condition assessment based on changes observed in its power curve," in *International Conference on Renewable Energy Research and Applications (ICRERA)*, Madrid, Spain, 2013.

[3] T. Pierre, W. Rene, O. Mohand, S.-H. Fouad, A.-T. Tommy, and E. Gabriel, "Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges," *Energies*, vol. 7, pp. 2595–2630, April 2014.

[4] J. Hu, Z. Jiang, and H. Wang, “Preventive maintenance for a single-machine system under variable operational conditions,” *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 230, pp. 391–404, 2016.

[5] W. X. Yang, R. Court, and J. S. Jiang, “Wind turbine condition monitoring by the approach of scada data analysis,” *Renewable Energy*, vol. 53, p. 365376, May 2013.

[6] Y. Amirat, M. Benbouzid, S. Member, B. Bensaker, and R. Wamkeue, “Condition monitoring and fault diagnosis in wind energy conversion systems: A review,” in *IEEE International Electric Machines & Drives Conference(IEMDC)*, vol. 2, Antalya, Turkey, 2007, pp. 1434–1439.

[7] Z. Hameed, Y. Hong, Y. Cho, S. Ahn, and C. Song, “Condition monitoring and fault detection of wind turbines and related algorithms: A review,” *Renewable and Sustainable Energy Reviews*, vol. 13, pp. 1–39, 2009.

[8] B. Lu, Y. Li, X. Wu, and Z. Yang, “A review of recent advances in wind turbine condition monitoring and fault diagnosis,” in *IEEE Power Electronics and Machines in Wind Applications (PEMWA)*, Lincoln, USA, 2009.

[9] M. Aziz, H. Noura, and A. Fardoun, “General review of fault diagnostic in wind turbines,” in *18th Mediterranean Conference on Control & Automation (MED)*, Marrakech, Morocco, 2010.

[10] F. Garcia, A. Tobias, J. Pinar, and M. Papaelias, “Condition monitoring of wind turbines: Techniques and methods,” *Renewable Energy*, vol. 46, pp. 169–178, 2012.

[11] P. Tchakoua, R. Wamkeue, T. Tameghe, and G. Ekemb, “A review of concepts and methods for wind turbines condition monitoring,” in *World Congress on Computer and Information Technology (WCCIT)*, Sousse, Tunisia, 2013, pp. 1–9.

[12] L. B. Tjernberg, R. Allan, and R. Eriksson, “A reliability-centered asset maintenance method for assessing the impact of maintenance in power distribution systems,” *IEEE Transactions on Power Systems*, vol. 20, pp. 75 – 82, 2005.

[13] F. Spinato, P. Tavner, G. V. Bussel, and E. Koutoulakos, “Reliability of wind turbine subassemblies,” *IET Renewable Power Generation*, vol. 3, pp. 387 – 401, 2009.

[14] M. Wilkinson, B. Hendriks, F. Spinato, E. Gomez, H. Bulacio, J. Roca, P. Tavner, Y. Feng, and H. Long, “Methodology and results of the reliawind reliability field study,” in *European Wind Energy Conference (EWEC)*, Warsaw, Poland, April 2010.

[15] K. Fischer, F. Besnard, and L. B. Tjernberg, “A limited-scope reliability-centred maintenance analysis of wind turbines,” in *European Wind Energy Conference & Exhibition(EWEA)*, Brussels, Belgium, 2011, pp. 89–93.

[16] K. Fischer, F. Besnard, and L. B. Tjernberg, “Reliability centered maintenance for wind turbines based on statistical analysis and practical experience,” *IEEE Transactions on Energy Conversion*, vol. 27, pp. 184 – 195, 2011.

[17] P. Tavner, *Offshore Wind Turbines: Reliability, Availability & Maintenance*. The Institution of Engineering and Technology (IET), 2012.

[18] P. Tavner, D. Greenwood, M. Whittle, R. Gindele, S. Faulstich, and B. Hahn, “Study of weather and location effects on wind turbine failure rates,” *Wind Energy*, vol. 16, pp. 175–187, March 2013.

[19] B. Chen, “Automated on-line fault prognosis for wind turbine condition monitoring using scada data,” School of Engineering and Computing Science, Durham University, PhD Thesis, 2014.

[20] J. Moorse, “Analysis of scada data from large wind farms to provide incipient fault detection,” School of Engineering and Computing Science, Durham University, MEng Thesis, 2010.

[21] B. Chen, Y. Qiu, Y. Feng, P. Tavner, and W. Song, “Wind turbine scada alarm pattern recognition,” in *IET Conference Renewable Power Generation (RPG)*, Edinburgh, Scotland, 2011, pp. 1–6.

[22] A. Kusiak and A. Verma, “A data-driven approach for monitoring blade pitch faults in wind turbines,” *IEEE Transactions on Sustainable Energy*, vol. 2, pp. 87–96, 2010.

[23] B. Chen, P. Tavner, Y. Feng, W. Song, and Y. Qiu, “Bayesian networks for wind turbine fault diagnosis,” in *European Wind Energy Association Conference*, Copenhagen, Denmark, 2012.

[24] B. Chen, P. C. Matthews, and P. J. Tavner, “Automated on-line fault prognosis for wind turbine pitch systems using supervisory control and data acquisition,” *IET Renewable Power Generation*, vol. 9, pp. 503 – 513, 2015.

[25] F. Besnard and L. B. Tjernberg, “An approach for condition-based maintenance optimization applied to wind turbine blades,” *IEEE Transactions on Sustainable Energy*, vol. 1, pp. 77 – 83, 2010.

[26] W. Qiao and D. Lu, “A survey on wind turbine condition monitoring and fault diagnosis - part ii: Signals and signal processing methods,” *IEEE Transactions on Industrial Electronics*, vol. PP, April 2015.

[27] P. Mazidi, L. T. Bertling, and M. A. Sanz-Bobi, “Wind turbine prognostics and maintenance management based on a hybrid approach of neural networks and proportional hazards model,” *Journal of Risk and Reliability*, 2016.

[28] P. Mazidi, M. Du, L. Bertling-Tjernberg, and M. A. Sanz-Bobi, “A health condition model for wind turbine monitoring through neural networks and proportional hazard models,” *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 2017.

[29] Q. Huang, D. Jiang, L. Hong, and Y. Ding, “Application of wavelet neural networks on vibration fault diagnosis for wind turbine gearbox,” in *Advances in Neural Networks - International Symposium on Neural Networks*, September 2008, pp. 313–320.

[30] S. Yang, W. Li, and C. Wang, “The intelligent fault diagnosis of wind turbine gearbox based on artificial neural network,” in *International Conference on Condition Monitoring and Diagnosis*, Beijing, China, 2008, pp. 1327 – 1330.

[31] Y. Wang and D. Infield, “Neural network modelling with autoregressive inputs for wind turbine condition monitoring,” in *International Conference on Sustainable Power Generation and Supply (SUPERGEN)*, Hangzhou, China, 2012, pp. 1–6.

[32] D. J. Lekou, F. Mouzakis, A. A. Anastasopoulos, and K. D., “Emerging techniques for health monitoring of wind turbine gearboxes,” in *European Wind Energy Conference: Scientific Track-Operation and Maintenance*, 2009.

[33] P. Bangalore and L. B. Tjernberg, “An artificial neural network approach for early fault detection of gearbox bearings,” *IEEE Transactions on Smart Grid*, vol. 6, pp. 980 – 987, 2015.

[34] A. Zaher, S. McArthur, D. Infield, and Y. Patel, “Online wind turbine fault detection through automated scada data analysis,” *Wind Energy*, vol. 12, pp. 574–593, September 2009.

[35] M. Schlechtingen and I. Santos, “Condition monitoring with ordinary wind turbine scada data a neuro-fuzzy approach,” in *European Wind Energy Association (EWEA) Offshore*, Amsterdam, Netherlands, 2011, p. 9.

[36] M. Schlechtingen, I. Santos, and S. Achiche, “Wind turbine condition monitoring based on scada data using normal behavior models. part 1: System description,” *Applied Soft Computing*, vol. 13, p. 259270, January 2013.

[37] H. Li, Y. Hu, C. Yang, Z. Chen, H. Ji, and B. Zhao, “An improved fuzzy synthetic condition assessment of a wind turbine generator system,” *International Journal of Electrical Power & Energy Systems*, vol. 45, p. 468476, February 2013.

[38] P. Mazidi, M. Du, and M. A. Sanz-Bobi, “A comparative study of techniques utilized in analysis of wind turbine data,” in *China International Conference on Electricity Distribution (CICED)*, 2016.

[39] M. Du, J. Yi, P. Mazidi, L. Cheng, and J. Guo, “A parameter selection method for wind turbine health management through scada data,” *Energies*, vol. 10, 2017.

[40] P. Mazidi, M. Du, L. Bertling-Tjernberg, and M. A. Sanz-Bobi, “A performance and maintenance evaluation framework for wind turbines,” in *International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, October 2016.

[41] P. Mazidi, Y. Tohidi, and M. A. Sanz-Bobi, “Strategic maintenance scheduling of an offshore wind farm in a deregulated power system,” *Energies*, vol. 10, 2017.

[42] R. Fisher, *Statistical Methods for Research Workers*. Hafner, 1958.

[43] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*. Wiley, 2012.

[44] P. de B. Harrington, “Sigmoid transfer functions in backpropagation neural networks,” *Analytical Chemistry*, vol. 65, p. 21672168, 1993.

[45] M. F. Moller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural Networks*, vol. 6, pp. 525–533, 1993.

[46] D. F. Morrison, *Multivariate Statistical Methods*. Duxbury Press, 2004.

[47] L. I. Smith, “A tutorial on principal components analysis,” 2002. [Online]. Available: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[48] A. Staino and B. Basu, “Emerging trends in vibration control of wind turbines- a focus on a dual control strategy,” *Philosophical Transactions of the Royal Society A*, vol. 373, January 2015.