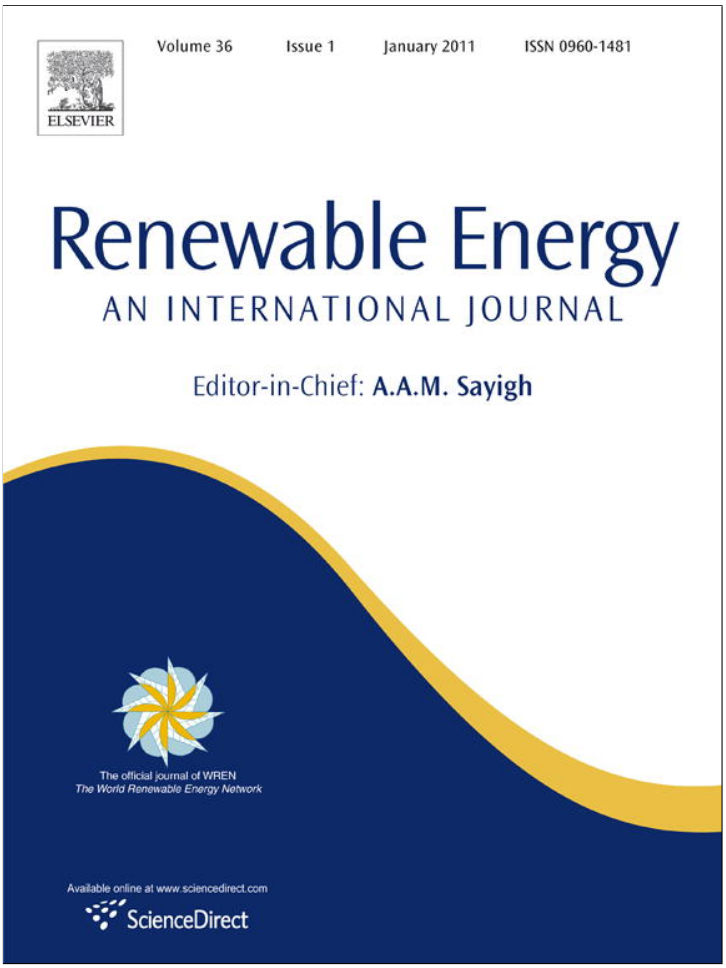

Renewable Energy 36 (2011) 16–23

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

The prediction and diagnosis of wind turbine faults

Andrew Kusiak*, Wenyan Li

Department of Mechanical and Industrial Engineering, 3131 Seamans Center, The University of Iowa, Iowa City, IA 52242-1527, USA

ARTICLE INFO

Article history:
Received 11 October 2009
Accepted 19 May 2010
Available online 9 June 2010

Keywords:
Wind turbine
Fault prediction
Fault identification
Condition monitoring
Predictive modeling
Computational modeling

ABSTRACT

The rapid expansion of wind farms has drawn attention to operations and maintenance issues. Condition monitoring solutions have been developed to detect and diagnose abnormalities of various wind turbine subsystems with the goal of reducing operations and maintenance costs. This paper explores fault data provided by the supervisory control and data acquisition system and offers fault prediction at three levels: (1) fault and no-fault prediction; (2) fault category (severity); and (3) the specific fault prediction. For each level, the emerging faults are predicted 5–60 min before they occur. Various data-mining algorithms have been applied to develop models predicting possible faults. Computational results validating the models are provided. The research limitations are discussed.

1. Introduction

Wind power is regarded as a key source in meeting the planned targets of the carbon emission reductions and the diversity of energy supply sources [1]. The growing interest in wind energy has led to the rapid expansion of wind farms [2,3].

The growth of wind power has increased interest in the operations and maintenance of wind turbines. As wind turbines are located at remote locations that may be difficult to access, their maintenance becomes an issue. As indicated in Ref. [4], a $5000 replacement of a bearing can turn into a $250,000 project involving cranes and a service crew in addition to the loss of power generation. For a turbine with 20 years of operating life, the operations, maintenance, and part replacement costs were estimated in the past to be at least 10–15% of the total income from the generation [5]. Thus, condition monitoring and fault diagnosis of wind turbines are of high priority.

The state-of-the-art research in wind turbine condition monitoring and fault diagnosis has been covered in the past literature [6–10] with more recent updates included in Refs. [11,12]. Modern wind turbines are usually equipped with some form of condition monitoring systems, including system-level or subsystem-level fault detection. Subsystem-level fault detection systems are usually based on monitoring parameters such as the vibration of the wind turbine drive train [13], bearing temperature, oil particulate content, optical strain measurements [14], and so on. Some commercially available solutions include blade monitoring systems [15], Supervisory Control and Data Acquisition (SCADA) interpretation systems [16], and holistic models [17]. The system-level condition monitoring and fault diagnosis offer a challenge that has led to numerous modeling and solution approaches presented in the literature, including Petri Nets [18], physics-based models [19,20], multi-agent framework for fault detection [21], and sensor-based network [22].

Data mining is a promising approach for modeling wind energy, e.g., power prediction and optimization [23–25], wind speed forecasting [26,27], and power curve monitoring [28]. It involves a number of steps including data pre-processing, data sampling, feature selection, and dimension reduction.

This paper proposes a methodology for system-level fault diagnosis in wind turbines using a data-driven approach. The fault-related data is analyzed at three levels. The existence of a status or a fault is predicted (Level 1), the category (severity) of the fault or the status is determined (Level 2), and the specific fault is predicted (Level 3).

The paper is organized into eight sections. Section 2 describes the status/fault data as well as the data routinely collected at a wind farm (referred here as the SCADA data). Section 3 analyzes the power curves generated by the data collected from a randomly selected turbine. Section 4 defines the method to distinguish faults and statuses based on the available status/fault data. Section 5 presents a methodology for fault prediction. Section 6 illustrates the model-building process; the models are validated with two test data sets. Section 7 discusses the computational results from all the models. Finally, the research conclusions and limitations are presented in Section 8.

* Corresponding author.
E-mail address: andrew-kusiak@uiowa.edu (A. Kusiak).

0960-1481/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.renene.2010.05.014

Table 1
Sample status codes.

Status code	Status text	Category
1	Program start PLC	2
2	No errors	4
3	Manual stop	4
4	Remote stop	4
5	Remote start	4
6	System OK	4
9	Under-voltage	4
21	Cable twisting left	4
25	No speed reduction with primary braking	1
28	No speed reduction with secondary braking	1

2. Data description and pre-processing

The data available for the research reported in this paper has been collected by SCADA systems at four wind turbines (Turbine 1, Turbine 2, Turbine 3, and Turbine 4). For each turbine, two separate sets of data were provided: SCADA data and status/fault data. Both data sets were collected at period of three months from 01/04/2009 to 30/06/2009. The details of the data are discussed next.

2.1. SCADA data

The SCADA data for four wind turbines was collected at 5-min intervals. The nearly 25 000 records (instances) collected for each turbine on over 60 parameters have been grouped into four categories.

1) *Wind parameters:* Wind parameters are the direct measurements of the wind (e.g., wind speed, wind direction) and derived values (e.g., wind intensity and turbulence).
2) *Energy conversion parameters:* Parameters in this category are related to the energy conversion process (e.g., power output, blade pitch angle, generator torque, rotor speed) and so on.
3) *Vibration parameters:* Vibration parameters indicate operational conditions of the turbine systems. They usually involve measurements of the drive train acceleration and tower acceleration.
4) *Temperature parameters:* This category of parameters includes the temperature measured at turbine components (e.g., bearing temperature) and the air temperature around turbine components and subsystems (e.g., nacelle interior temperature).

2.2. Status/fault data

Status/fault data provides information on statuses and faults recorded by the SCADA system. A fault, in this paper, refers to a status that with a certain probability results in a severe consequence to the wind turbine system. For example, ignoring the status "Emergency stop nacelle/hub" or "Pitch thyristor 1 fault" might damage the wind turbine components. Other statuses, however, such as "No errors" and "Remote start" may not lead to severe consequences. Examples of status codes are illustrated in Table 1.

Each status code in Table 1 is associated with a specific abnormality of a turbine component or a subsystem. There are nearly 350 different status codes in the data considered in this research. The status text in Table 1 provides a short description of the status, and the category denotes its severity. Category "1" implies the most severe status, and Category "4" corresponds to the least severe status.

The status/fault data has been collected by the SCADA system. Nearly 7000 occurrences of status codes have been observed at each turbine over the three-month period, including the seven parameters illustrated in Table 2.

Table 2
Parameters related to the fault information.

Parameter name	Definition	Unit	Symbol
Fault time	Date and time of the fault occurrence		t_{fault}
Status code	Status code assigned to the fault		
Category	Category of the status code (four categories)		Category
Generator speed	Generator speed at the time the fault occurred	Nm	GS(t_{fault})
Power output	Power production at the time the fault occurred	kW	PO(t_{fault})
Wind speed	Wind speed at the time the fault occurred	m/s	WS(t_{fault})

Table 3
Illustration of data instances with out-of-range values of wind speed.

Date	Time	Status code	Wind speed	Power output	Generator torque
4/9/2009	4:24:10 AM	0	**−42946720**	−1	0
4/9/2009	4:34:10 AM	0	**−42946720**	−1	0
6/25/2009	1:54:54 AM	183	**32509316**	−2	0
6/25/2009	1:54:54 AM	183	**27872676**	−1	0

2.3. Issues with status/fault data

Although the raw data contained over 7000 status/fault instances for each of the four wind turbines, some of the instances could not be considered for the following reasons:

1) Presence of wind speed measurements with unreasonably large values, as illustrated in (bold) Table 3
The wind speed measured by an anemometer should be in the range [0, cut-out speed], here [0 m/s, 21 m/s].
In this case, the negative values of wind speed were assigned status code "0", and the positive out-of-range speed was assigned status code "183" (see Table 3). However, the status code "183" is not unique to the wind speed error, as it is used to label other anomalies when the wind speed is in the feasible range. A possible reason for the multiple meaning of the same status code (here "183") might be due to multiple errors occurring simultaneously. The status code "0" is discussed next.

2) Status code "0"
In Table 3 status code "0" was assigned to the out-of-range negative values of the wind speed. The same status value is assigned for the four instances in Table 4 as illustrated in bold. Status code "0", however, does not offer any useful status or fault information. Based on the data analysis, the meaningful status codes appear to be in the range of [1, 350].

3) Presence of duplicate data
Some of the data entries associated with the status code could be repeated a number of times, as illustrated in Table 5 for the status code "183". The reason behind the repeated values could be in the imperfection of the SCADA software.

Table 4
Fault information for status code "0".

Date	Time	Status code	Wind speed	Power output	Generator torque
4/7/2009	3:02:43 AM	**0**	17	−3	76
4/7/2009	3:18:05 AM	**0**	16	−3	72
4/7/2009	3:18:05 AM	**0**	15	−3	47
4/7/2009	3:22:46 AM	**0**	14	0	77

Table 5
Duplicate fault information.

Date	Time	Status code	Wind speed	Power output	Generator torque
5/24/2009	7:20:34 PM	183	5	−2	998
5/24/2009	7:20:34 PM	183	5	−2	998
5/24/2009	7:20:34 PM	183	5	−2	998
5/24/2009	7:20:34 PM	183	5	−2	998

2.4. Pre-processing status/fault data

As incorrect data would negatively impact the models built, all status/fault data is pre-processed for removal of the data in doubt. The number of status/fault instances after data pre-processing for each of the four turbines is shown in Table 6.

The data set in Table 6 has been significantly reduced. For example, the 7000 instances of the status/fault data initially provided for Turbine 4 have led to 1329 instances covering 66 different status codes. The data collected at Turbine 4 (as indicated in bold in Table 6 and Table 7) has been selected for further analysis.

3. The power curve

3.1. Power curve based on SCADA data

The shape of the power curve determines the health of a wind turbine. A model power curve is portrayed as a sigmoid function representing the relationship between the power produced for the wind speed in the range between cut-in and cut-out speed.

A power curve built from the actual data deviates from an ideal power curve in the following: (1) some power outputs are negative; (2) there are different values of power output for identical wind speeds. The results of analysis of over 25 000 instances of 5-min SCADA data collected for each of the four turbines during a three-month period are summarized in Table 7.

The data in Table 7 has been organized according to the values of the power output. Three categories of power output are considered for each wind turbine: positive values, negative values, and values in error. Positive power implies generation of electrical energy. Negative power implies that the wind turbine is consuming energy likely due to the low wind speed. The erroneous data is due to various status/fault situations. Figs. 1 and 2 illustrate the power curve for Turbine 4 for positive and negative power values, respectively.

The power curve in Fig. 1 includes scattered points providing a basis for fitting into an ideal power curve. There are a number of reasons for the variability reflected in the power curve, including the errors caused by malfunctions of the turbine systems and components.

Most (97.48%) negative values of the power output are in the range of [−10 kW, 0 kW], and the minimum negative power is −30 kW. Nearly 2/3 of the power outputs (63.11%) of Turbine 4 are negative during the period analyzed. There are two main reasons for negative power: the wind speed is lower than the cut-in speed, and there are maintenance issues with the turbine.

Table 6
Reduced status/fault data set.

Turbine No.	Status/fault instances	No of status codes
1	2383	65
2	2619	59
3	817	49
4	**1329**	**66**

Table 7
Summary of SCADA data for four turbines.

Turbine	Number of positive power values	Number of negative power values	Number of erroneous data values	Total number of instances
1	24 892	3415	1031	29 338
2	21 035	3844	1030	25 909
3	3504	21 309	1108	25 921
4	**8466**	**16 359**	**1095**	**25 920**

3.2. Power curve based on status/fault data

In addition to turbine operational data collected by the SCADA system, a turbine status/fault data is generated for each turbine. The status/fault data is time stamped, and therefore it can be linked with the SCADA records. The solid line in Fig. 3 illustrates a model power curve obtained from the data used to map the power curve in Fig. 1. The model power curve in Fig. 3 was built from 8466 instances representing the normal (fully functional) status of Turbine 4 by constructing 30 neural networks.

The neural network with the smallest training error was selected to predict the power curve (solid line) in Fig. 3. A similar approach to generate a power curve was used in previous research [28]. The scattered points in Fig. 3 represent the status/faults instances collected as a separate file. As illustrated in Fig. 3, some status/fault data points present themselves no differently than the points creating a typical power curve. The data points representing zero or negative power consumption also fall in the status/fault category.

The wind speed and power output are used in this paper as input variables to identify faults of a wind turbine.

4. Status codes in the Turbine data

4.1. Frequency for statuses/faults

The frequency of faults in the data set varies. For example, for Turbine 4, the status codes “180” to “184” happen hundreds of times, while the status code “1” or “5” occurs only a few times, or as rarely as once every three months. Fig. 4 illustrates the frequency of statuses/faults for Turbine 4.

Table 8 provides detailed information on five status codes, 181 to status 185.

The most frequent status shown in Table 8 is “Start-up”, which occurs 220 times in the three-month period. The other three statuses occur hundreds of times. These statuses, however, do not seriously impact the wind turbine system. The low impact status codes are of lesser interest to this research. Rather, the focus is on the severer faults.

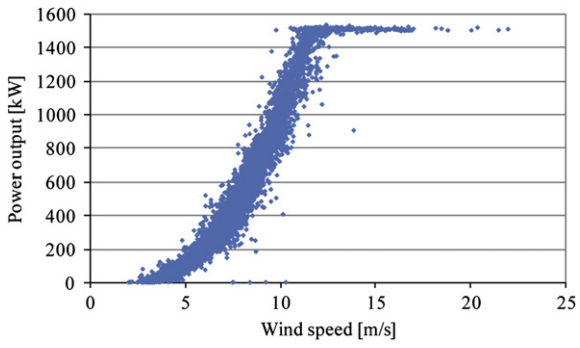

Fig. 1. Turbine 4 curve for positive power values.

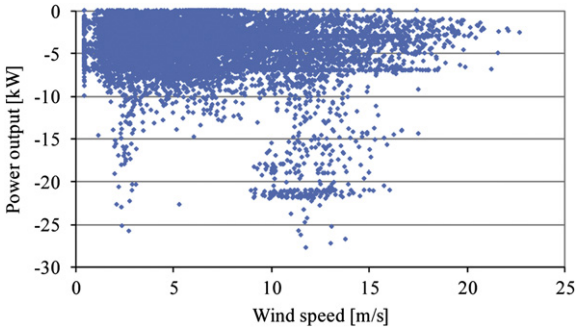

Fig. 2. Turbine 4 curve for negative power values.

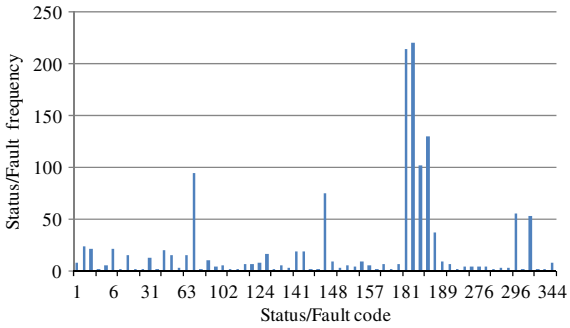

Fig. 4. Fault frequency of Turbine 4.

4.2. Fault versus status

Neither the data available in this research nor the current literature discusses the relationship between "statuses" and "faults" in wind turbines. This paper presents a useful approach for making such a distinction.

Each status/fault code of a wind turbine is assigned one of the four categories according to its severity of impact on the wind turbine system. It is observed from the data provided that categories 1, 2 and 3 might adversely impact the wind turbine system and its components. But the status codes in Category 4 are not likely to seriously hinder the operations of a wind turbine. Statuses in categories 1, 2 and 3 are regarded as faults, and statuses in Category 4 are considered as statuses. The distribution of faults for all four turbines in each category is shown in Table 9.

As illustrated in Table 9, Category 4 statuses occur most frequently (87.33% on average for the four turbines). The most severe faults (Category 1) happen on average 1.50% of the time. The faults of Categories 2 and 3 occur more frequently than those of Category 1. The fault distribution of Turbine 4 is illustrated in Fig. 5.

There are 35 specific faults (11 in Category 1, 20 in Category 2, 4 in Category 3), and 31 different status occurrences. In total 233 (42 + 131 + 60) faults and 1096 statuses are captured during the three-month period.

4.3. Most frequent faults

The faults that occur relatively frequently (Fault frequency > 10) and their categories are listed in Table 10.

As illustrated in Table 10, only seven faults happen more than 10 times during the time period reflected in the data, including one fault from Category 1, five faults from Category 2, and one fault from Category 3. As the malfunction of the diverter (status code "296") occurs most frequently, it is selected for further analysis.

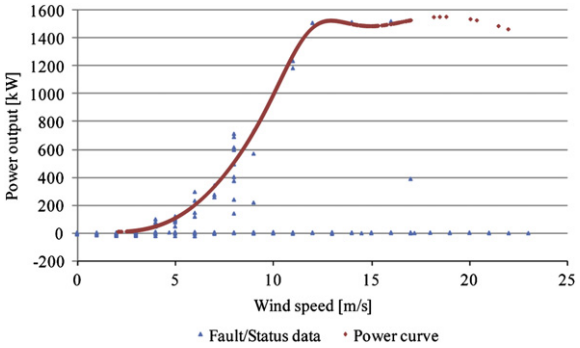

Fig. 3. Power curve of Turbine 4 and scattered points included in the status/fault file.

5. Proposed methodology for fault prediction

5.1. Three-level fault prediction

The proposed methodology for fault prediction of wind turbine systems involves three levels (see Fig. 6).

Level 1 : Predict status/fault
The goal of this level is to distinguish the status/fault data from the labeled SCADA data (to be discussed in Section 5.2). No differentiation is made between a status and a fault.

Level 2 : Predict category of status/fault
It is not enough to recognize whether a status/fault has occurred at a certain time. At this level, the category of a status or a fault is detected.

Level 3 : Predict specific fault
There are nearly 350 different status codes occurring with different frequencies for wind turbines. It is easier to detect statuses that are more frequent. At this level, fault "Malfunction of diverter" (shown in bold in Table 10) is predicted up to 60 min before it occurs.

For each level of fault prediction, the general process is divided into four steps: labeling SCADA data, data sampling, model extraction, and computational results analysis. The process of fault prediction is outlined in Fig. 7 and discussed in the next sections.

5.2. Labeling SCADA data with status/fault code and category

The data of Table 2 is generated whenever a status/fault occurs. The SCADA and the status/fault data is integrated by assigning status/fault codes and their categories to SCADA data according to (1)

$$\begin{aligned} &\text{If } T_{SCADA}(t-n) < T_{\text{fault}}\left(t_{\text{fault}}\right) < T_{SCADA}(t-n+1) \text{ Then} \\ &\quad \text{Status\_Code}(t-n) = \text{Status\_Code}\left(t_{\text{fault}}\right) \\ &\quad \text{Category}(t-n) = \text{Category}\left(t_{\text{fault}}\right) \end{aligned} \tag{1}$$

Table 8
Detailed information about status codes 181–185.

Status code	Status text	Category	Frequency
181	Idling position	4	214
182	Start-up	4	220
183	Load operation	4	102
184	Shut down	4	130
185	Manual operation of pitch	4	37

Table 9
Distribution of faults and statuses by category.

Turbine	Category 1		Category 2		Category 3		Category 4		Overall
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage	
1	40	1.68%	417	17.50%	44	1.85%	1882	78.98%	2383
2	14	0.53%	42	1.60%	36	1.37%	2527	96.49%	2619
3	11	1.35%	40	4.90%	29	3.55%	737	90.21%	817
4	42	3.16%	131	9.86%	60	4.51%	1096	82.47%	1329
Sum/Average	107	1.50%	630	8.81%	169	2.36%	6242	87.33%	7148

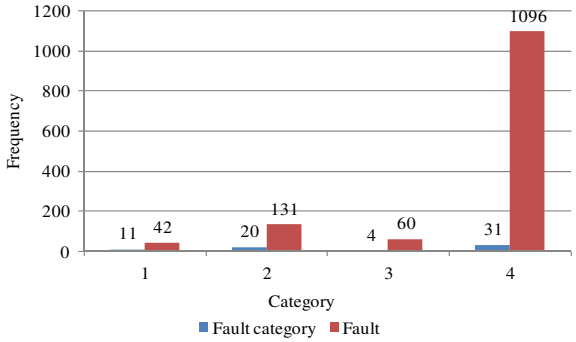

Fig. 5. Fault distribution for Turbine 4.

where Status_Code and Category are as shown in Table 2; n is the number of time stamps in advance of the status/fault. An attempt will be made to detect a status/fault $n \times 5$ min in advance. In this paper, n is assumed as 12; i.e., up to 60 min ahead of reporting the status/fault.

In the process of matching the status/fault data (see Table 2) with the SCADA data (unlabeled data), some status/fault data is deliberately ignored. The reason is that the status/fault data is recorded whenever the status or the fault happened, while the turbine operational data is reported at 5-min intervals. During the 5-min interval, the status code with the most severe category is considered. Table 11 shows a typical status code file.

As illustrated in Table 11, a number of status codes are reported at the same time, i.e., 11:23:27 PM. Of those, only the most severe category status code, e.g., status code "292" "Malfunction of cabinet heaters" is merged with the SCADA data. This way the 5-min record of the SCADA file corresponding to the time stamp 11:23:27 PM is assigned the Category 3 label.

After the turbine operations data have been labeled with the status/fault according to (1), 637 status/fault instances remain for Level 1 and Level 2 predictions. In other words, almost 50% of the status/fault information is lost. The fault "malfunction of diverter" with the status code 296 is used in this experiment. Of 55 status occurrences, 50 status/fault instances were used, and 5 instances were lost.

5.3. Data sampling

An ideal training data set should be balanced with status/fault and normal operations data. The selection of the status/fault data

Table 10
Most frequent faults of Turbine 4.

No	Status code	Status text	Category	Fault frequency
1	31	Timeout of yaw counter	2	12
2	45	Hydraulic pump time too high	2	20
3	52	Gearbox oil pressure too low	2	15
4	63	Safety chain	1	14
5	141	Rotor CCU collective faults	2	18
6	142	Line CCU collective faults	2	18
7	**296**	**Malfunction of diverter**	**3**	**55**

Fig. 6. Levels for fault prediction.

was discussed in Section 5.2. To construct a training data set reflecting normal turbine operations, direct use of the labeled SCADA data is not acceptable, as the number of records (8466) would vastly exceed the number of instances of the status/fault data and thus cause a prediction bias. Data sampling is an effective technique to deal with this issue.

A data sample is randomly selected from the normal instances of the labeled SCADA data of Turbine 4. To create a balanced data set, the size of the data sample depends on the size of the fault data at a particular level. Thus, for Level 1 and Level 2, 650 normal instances are selected, and for Level 3, 118 instances. At each level, the fault and normal instances are combined into one file. The combined data set for Level 1 and Level 2 predictions contains 650 normal instances and 637 fault instances. For Level 3 predictions, the combined data set contains 118 normal instances and 50 fault instances.

5.4. Test strategy

At each level of fault prediction, a training data set is created by randomly selecting 2/3 of the instances of the combined data set for each time stamp from t to $t - 12$. Specifically, 13 training data sets are provided for prediction from current time t to the proceeding 60 min, i.e., $t - 12$.

Two types of test data sets are provided. The first test data set uses 1/3 of the instances of the combined data for each time stamp. Normal instances and fault instances are sampled separately to avoid unbalanced fault distribution in the training and test data sets. For example, 16 faults of Category 1 are provided in the combined data sets; 10 are used for training and the other 6 are used for testing. The second test data set is created by randomly selecting 10% of the data from the labeled SCADA data. In the first test data set, the percentage of faults versus normal instances is much higher than that in the labeled SCADA data set. The second test data represents the distribution of the labeled SCADA data. The number of instances sampled for each data set (at each of the three levels) is illustrated in Table 12.

As illustrated in Table 12, the number of fault instances in the second test data set is limited. There are 61 status/fault instances

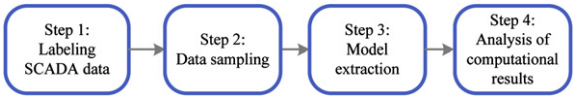

Fig. 7. Process of fault prediction.

Table 11
Ignored status code data while matching it with the SCADA data.

Date	Time	Status code	Status text	Category	Wind speed	Power output	Generator torque
4/3/2009	11:23:27 PM	95	PC restart	4	6	−7	22
4/3/2009	11:23:27 PM	156	Repair	4	6	−7	22
4/3/2009	11:23:27 PM	292	Malfunction of cabinet heaters	3	6	−7	22
4/3/2009	11:23:27 PM	293	Malfunction of temp switch cabinet	3	6	−7	22
4/3/2009	11:23:27 PM	296	Malfunction of diverter	3	6	−7	22

Table 12
Training and test data sets.

Level	Training data set					Test data set 1					Test data set 2				
Level 1	Normal	Status/Fault				Normal	Status/Fault				Normal	Status/Fault			
	433	425				217	212				2007	61			
Level 2	Normal	C1	C2	C3	C4	Normal	C1	C2	C3	C4	Normal	C1	C2	C3	C4
	433	10	22	16	375	217	4	12	9	187	2007	1	2	2	56
Level 3	No Fault 296	Fault 296				No Fault 296	Fault 296				No Fault 296	Fault 296			
	80	35				38	15								

for Level 1 predictions. Only 5 faults (1 for Category 1, 2 for Category 2, and 2 for Category 3) are provided among the sampled data for Level 2 predictions. For Level 3 predictions, the randomly sampled test data has a very low probability of including the fault "Malfunction of diverter".

6. Model extraction

The model's extraction process is illustrated in Fig. 8.

As illustrated in Fig. 8, input variables are wind speed $(t-n)$ and power output $(t-n)$. The target outputs are (1) fault-no fault at t_{fault}; (2) category of the fault at t_{fault}; and (3) fault "296" at t_{fault}. To compare the prediction results, three metrics defined in (2)–(4) are used.

$$\text{Accuarcy} = \frac{\text{Number of correctly predicted fault instances} + \text{Number of correctly predicted normal instances}}{\text{Number of fault instances} + \text{Number of normal instances}} \times 100\% \quad (2)$$

$$\text{Sensitivity} = \frac{\text{Number of correctly predicted fault instances}}{\text{Number of fault instances}} \times 100\% \quad (3)$$

$$\text{Specification} = \frac{\text{Number of correctly predicted normal instances}}{\text{Number of normal instances}} \times 100\% \quad (4)$$

Accuracy provides the percentage of correctly made predictions. Sensitivity expresses the percentage of correctly predicted faults, and specificity expresses the percentage of correctly predicted normal instances.

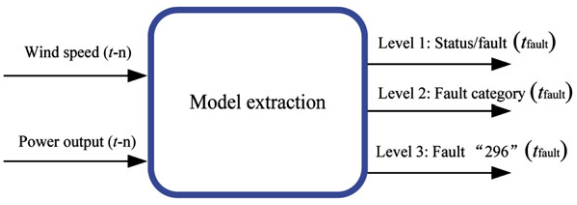

Fig. 8. The model extraction process.

6.1. Model extraction at Level 1

Four data-mining algorithms have been applied to extract the models, the Neural Network (NN), the Neural Network Ensemble (NN Ensemble), the Boosting Tree Algorithm (BTA), and the Support Vector Machine (SVM). The prediction results for the test data set 1 at current time t are shown in Table 13.

As illustrated in Table 13, the NN-ensemble makes the best quality predictions, and therefore it is recommended for building Level 1 models. To construct the NN-ensemble, 30 NNs are built and the best five are selected.

6.2. Model extraction at Level 2

Several data-mining algorithms have been applied to extract the models, including the Neural Network (NN), the Standard Classification and Regression Tree (CART), the Boosting Tree Algorithm (BTA), and the Support Vector Machine (SVM). The prediction accuracy (%) results for test data set 1 at current time t are compared in Table 14.

As illustrated in Table 14, CART exhibits the strongest potential and is selected for further predictions of fault categories.

6.3. Model extraction at Level 3

Several algorithms have been applied to extract the data-mining models, including the Neural Network (NN), Neural Network

Table 13
Performance of four algorithms predicting status/fault at time t.

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
NN	74.71	81.00	68.67
NN Ensemble	**74.56**	**83.67**	**65.81**
BTA	71.27	84.66	59.50
SVM	69.64	59.97	78.92

Table 14
Performance of four algorithms for fault category predictions.

Algorithm	Prediction accuracy for normal	Prediction accuracy for category 1	Prediction accuracy for category 2	Prediction accuracy for category 3	Prediction accuracy for category 4
NN	76.66	0.00	0.00	12.00	74.91
BTA	41.00	22.22	83.33	0.00	72.15
CART	**96.08**	**62.50**	**52.94**	**56.00**	**95.20**
SVM	80.88	0.00	0.00	0.00	69.28

Table 15
Performance of four algorithms in prediction of a specific fault.

Algorithm	Accuracy (%)	Sensitivity (%)	Specification (%)
BTA	**69.81**	**86.67**	**63.16**
NN	72.00	66.67	70.45
NN Ensemble	68.00	82.88	66.67
SVM	70.59	47.06	82.35

Ensemble (NN Ensemble), Boosting Tree Algorithm (BTA), and Support Vector Machine (SVM). The prediction results for test data set 1 at current time t are compared in Table 15.

As illustrated in Table 15, the BTA algorithm has been selected for prediction of the fault "Malfunction of diverter". The learning rate used by this algorithm was 0.1.

7. Computational results analysis

7.1. Computational results for Level 1

7.1.1. Performance of test data set 1

In this section, the models extracted in Section 6.1 have been applied to test data 1 (as illustrated in Table 12) for Level 1 predictions. The models are extracted from current time t to time stamp $t-12$ (13 prediction models). The prediction results for six models (one per time stamp) are illustrated in Table 16.

As illustrated in Table 16, the prediction accuracy is in the interval of [63%, 77%]. The sensitivity is relatively high, implying that most faults and statuses have been correctly identified. The accuracy and sensitivity at the time stamp $t-12$ are lower than at other periods.

7.1.2. Performance of test data set 2

In this section, the models extracted in Section 6.1 have been applied to test data set 2 (as illustrated in Table 12) for Level 1 prediction. The prediction results obtained at time stamps are illustrated in Table 17.

The results in Table 17 show that prediction accuracy at the time stamps t to $t-12$ is in the range of [65%, 78%]. Most statuses/faults have been correctly predicted. The percentage of correctly predicted statuses/faults is in the interval of [39%, 94%], and correctly predicted normal instances are in the range of [63%, 79%].

Table 16
Test 1 results for status/fault at six time stamps.

Time stamp	Accuracy (%)	Sensitivity (%)	Specification (%)
t	74.56	83.67	65.81
$t-1$	74.42	75.98	72.93
$t-3$	75.19	85.87	65.02
$t-6$	75.10	86.34	64.43
$t-9$	76.03	88.38	64.40
$t-12$	63.77	51.18	75.63

Table 17
Test 2 results for status/fault prediction at six time stamps.

Time stamp	Accuracy (%)	Sensitivity (%)	Specification (%)
t	68.63	78.69	63.88
$t-1$	66.17	93.18	65.58
$t-3$	65.88	83.61	65.37
$t-6$	66.70	65.57	66.77
$t-9$	65.39	73.77	65.37
$t-12$	77.42	39.34	78.57

Table 18
Test 1 results for prediction of status/fault category at six time stamps.

Time	Accuracy	Normal instances	Category 1	Category 2	Category 3	Category 4
t	93.39	96.08	62.50	52.94	56.00	95.20
$t-1$	93.39	95.79	68.75	47.06	52.00	95.91
$t-3$	94.31	92.27	50.00	70.59	56.00	94.31
$t-6$	92.21	94.79	56.25	67.65	68.00	92.70
$t-9$	91.86	94.83	68.75	52.94	48.00	93.24
$t-12$	90.72	94.24	56.25	38.24	40.00	92.88

7.2. Computational results for Level 2 prediction

7.2.1. Performance of test data set 1

In this section, the models extracted in Section 6.2 have been applied to test data 1 (as illustrated in Table 12) for Level 2 predictions. The models are extracted from the current time t to time stamp $t-12$ (13 prediction models). The prediction accuracy (%) results produced at six time stamps are illustrated in Table 18.

As illustrated in Table 18, the prediction accuracy for normal and status instances is high. However, the percentage of correctly predicted fault instances is in the range of [40%, 71%].

7.2.2. Performance of test data set 2

In this section, the models extracted in Section 6.2 have been applied to test data 2 at Level 2. The accuracy results (%) for six time stamps are illustrated in Table 19.

Despite the fact that the number of status/fault categories is small, the results presented in Table 19 are quite impressive. The variability in accuracy seen there is due to the small number of status/fault categories. For example, if one of the two status/fault categories is predicted in error, then the accuracy decreases from 100% to 50%. The prediction accuracy for normal instances is still high. However, the accuracy for status/fault category prediction is lower compared to test data set 1.

7.3. Computational results for Level 3 predictions

In this section, the models extracted in Section 6.2 have been applied to test data set 1 (as illustrated in Table 12) for Level 3 predictions. The models are extracted from current time t to time stamp $t-12$ (13 prediction models). The prediction results for six time stamps are shown in Table 20.

Table 19
Test 2 results for prediction of status/fault category at six time stamps.

Time	Accuracy	Normal instances	Category 1	Category 2	Category 3	Category 4
t	99.27	99.75	100.00	100.00	100.00	82.14
$t-1$	99.17	98.00	100.00	50.00	100.00	84.62
$t-3$	99.13	99.36	50.00		50.00	77.78
$t-6$	99.08	99.90	100.00	50.00	50.00	73.21
$t-9$	98.87	99.75	50		100	75.93
$t-12$	98.77	99.51	100	100	50	76.79

Table 20
Test 1 results for prediction of a specific fault at six time stamps.

Time stamp	Accuracy (%)	Sensitivity (%)	Specification (%)
t	69.81	86.67	63.16
$t-1$	64.15	66.67	63.16
$t-3$	67.92	73.33	65.79
$t-6$	67.92	73.33	65.79
$t-9$	66.04	33.33	78.95
$t-12$	49.06	24.53	34.21

As illustrated in Table 20, the prediction accuracy of the fault "Malfunction of diverter" is in the interval of [49%, 70%]. The percentage of correctly predicted faults is in the interval of [24%, 87%], and the correctly predicted instances without the fault "Malfunction of diverter" is in the interval of [34%, 79%].

8. Conclusion

A methodology to predict turbine faults using information provided by SCADA systems and fault files was presented. The methodology involves three levels: (1) the existence of a status/fault was identified; (2) the category (severity) of the fault was predicted; and (3) a specific fault was predicted. The computational results reported in the paper demonstrated that, in most cases, faults can be predicted with a reasonable accuracy 60 min before they occur. The prediction accuracy of the fault category is somewhat lower yet acceptable. Due to the data limitations, identifying a specific fault, though valuable, is less accurate.

The research reported in this paper was performed with industrial data collected at operating wind turbines. The major difficulty was with the low frequency data. The description of faults was not clear, and the number of fault occurrences was far from sufficient. A better prediction performance would have been achieved with higher quality data.

The limitations surrounding this research are as follows:

1) The volume of fault data was limited, and therefore many faults did not appear in the data or occurred only sporadically. Such rare faults are difficult to detect by any modeling approach.
2) The 5-min interval for collecting the vast majority of data was too long. Such a long interval led to a significant loss of the history of the fault emergence.
3) In this paper, every status code was considered independently. The relationship between faults has not been considered largely due to the low frequency data.

Acknowledgement

The research reported in the paper has been supported by funding from the Iowa Energy Center, Grant 07-01.

References

[1] McMillan D, Ault G. Condition monitoring benefit for onshore wind turbines sensitivity to operational parameters. Renewable Power Generation 2009;2(1):60–72.
[2] "20% Wind energy by 2030: increasing wind energy's contribution to U.S. electricity supply". United States Department of Energy; July 2008. Report No. DOE/GO-102008–2567.
[3] Strategic research agenda: Market deployment strategy from 2008 to 2030. European Wind Energy Technology Platform. Available online: http://www.windplatform.eu/fileadmin/ewetp_docs/Bibliography/SRA_MDS_July_2008.pdf; July 2008.
[4] Hatch C. Improved wind turbine condition monitoring using acceleration enveloping. Orbit; 2004. pp. 58–61.
[5] Walford C. Wind turbine reliability: Understanding and minimizing wind turbine operation and maintenance costs. Sandia National Laboratories; March 2006. Report No. SAND2006-1100.
[6] Hyers R, McGowan J, Sullivan K, Manwell J, Syrett B. Condition monitoring and prognosis of utility scale wind turbines. Energy Materials 2006;1(3):187–203.
[7] Hameed Z, Hong Y, Cho Y, Ahn S, Song CK. Condition monitoring and fault detection of wind turbines and related algorithms: A review. Renewable and Sustainable Energy Reviews 2009;13(1):1–39.
[8] Amirat Y, Benbouzid M, Bensaker B, Wamkeue R. "Condition monitoring and fault diagnosis in wind energy conversion systems: a review". In: Proc. 2007 IEEE International Electric Machines and Drives Conference, vol. 2; May 2007. p. 1434–9.
[9] Tavner P, Bussel GW, Spinato F. "Machine and converter reliabilities in wind turbines,". In: Proc. third IET international conference on power electronics, machines and drives; 2006. p. 127–30.
[10] Wilkinson M, Spinato F, Tavner P. "Condition monitoring of generators and other subassemblies in wind turbine drive trains". In: Proc. 2007 IEEE International symposium on diagnostics for electric machines, power electronics and drives; Sep 2007. p. 388–92.
[11] Amirat Y, Benbouzid M, Al-Ahmar E, Bensaker B, Turri S. A brief status on condition monitoring and fault diagnosis in wind energy conversation systems. Renewable and Sustainable Energy Reviews 2009;13(9):2629–36.
[12] Lu B, Li Y, Wu X, Yang Z. "A review of recent advances in wind turbine condition monitoring and fault diagnosis". In: Proc. IEEE conference on power electronics and machines in wind applications. 2009. p. 1–7.
[13] Becker E, Posta P. Keeping the blades turning: condition monitoring of wind turbine gears. Refocus 2006;7(2):26–32.
[14] Editorial. Managing the wind: reducing kilowatt–hour costs with condition monitoring. Refocus 2005;6(3):48–51.
[15] Caselitz P, Giebhardt J. Rotor condition monitoring for improved operational safety of offshore wind energy converters. Transactions ASME, Journal of Solar Energy Engineering 2005;127(2):253–61.
[16] Leany V, Sharpe D, Infield D. Condition monitoring techniques for optimization of wind farm performance. International Journal of COMADEM 1992;2(1):5–13.
[17] Sanz-Bobi M, Garcia M, Del P. SIMAP: Intelligent system for predictive maintenance application to the health condition monitoring of a wind turbine gearbox. Computers in Industry 2006;57(6):552–68.
[18] Rodriguez L, Garcia E, Morant F, Correcher A, Quiles E. "Application of latent nestling method using colored Petri nets for the fault diagnosis in the wind turbine subsets." In: Proc. 2008 IEEE Int. conf. emerging technologies and factory automation. p. 767–73.
[19] Echavarria E, Tomiyama T, van Bussel G. Fault diagnosis approach based on a model-based reasoner and a functional designer for a wind turbine: an approach towards self-maintenance. Journal of Physics Conference Series 2007;75:012078.
[20] Echavarria E, Tomiyama T, Huberts H, van Bussel G. "Fault diagnosis system for an offshore wind turbine using qualitative physics," In: Proc. EWEC 2008, Brussels, Belgium; 2008.
[21] Zaher A, McArthur S. "A multi-agent fault detection system for wind turbine defect recognition and diagnosis." In: Proc.2007 IEEE Lausanne POWERTECH. p. 22–7.
[22] Whelan M, Janoyan K, Tong Q. "Integrated monitoring of wind plant systems," In: Proc. SPIE smart sensor phenomena, technology, networks, and systems, 6933; 2008. p. 69330F.
[23] Kusiak A, Zheng H-Y, Song Z. Short-Term prediction of wind farm power: a data-mining approach. IEEE Transactions on Energy Conversion 2009;24(1):125–36.
[24] Kusiak A, Li W. Virtual models for prediction of wind turbine parameters. IEEE Transactions on Energy Conversion 2010;25(1):245–52.
[25] Kusiak A, Li W, Song Z. Dynamic control of wind turbines. Renewable Energy 2010;35(2):456–63.
[26] Tarek H, Ehab F, Magdy M. One day ahead prediction of wind speed and direction. IEEE Transactions on Energy Conversion 2009;23(1):191–201.
[27] Mohandes M, Halawani T, Rehman S, Hussain A. Support vector machines for wind speed prediction. Renewable Energy 2004;29(6):939–47.
[28] Kusiak A, Zheng H-Y, Song Z. On-line monitoring of power curves. Renewable Energy 2009;34(6):1487–93.