

Introduction to Data-Driven Methodologies for Prognostics and Health Management

Jay Lee, Chao Jin, Zongchang Liu and Hossein Davari Ardakani

Abstract This book chapter gives an overview of prognostics and health management (PHM) methodologies followed by a case study in the development of PHM solutions for wind turbines. Research topics in PHM are identified and commonly used methods are briefly introduced. The case study in wind turbine prognostics has shown in detail how to develop a PHM system for an industrial asset. With the advancement of sensing technologies and computational capability, more and more industrial applications are emerging. Current gaps and future directions in PHM are discussed at the end.

Keywords Prognostics and health management • Wind energy • Data-driven • Prognostics

1 Overview of Prognostics and Health Management (PHM)

1.1 Definition and the Value of Prognostics and Health Management

Prognostics and health management (PHM) is an engineering discipline that aims at minimizing maintenance cost by the assessment, prognosis, diagnosis, and health management of engineered systems. With an increasing prevalence of smart sensing and with more powerful computing, PHM has been gaining popularity across a growing spectrum of industry such as aerospace, smart manufacturing, transportation, and energy at breakneck speed. Regardless of application, one common expectation of PHM is its capability to translate raw data into actionable information to facilitate maintenance decision making. This practice in industry is often

J. Lee (✉) • C. Jin • Z. Liu • H. Davari Ardakani
NSF IUCRC for Intelligent Maintenance Systems (IMS), University of Cincinnati,
Cincinnati, OH, USA
e-mail: jay.lee@uc.edu

referred to as Predictive Maintenance, which, as estimated by Accenture [1], could possibly save up to 12% over scheduled repairs, reduce overall maintenance costs by up to 30% and eliminate asset failures up to 70%. For example, a study performed by National Science Foundation (NSF) indicates that Center for Intelligent Maintenance Systems (IMS), which is a leading research center in the field of PHM, has created more than \$855 M of economic impact to the industry with a benefit cost ratio of 238:1 [2] through the development and deployment of PHM technologies to achieve near-zero unplanned downtime and a more optimized maintenance practice.

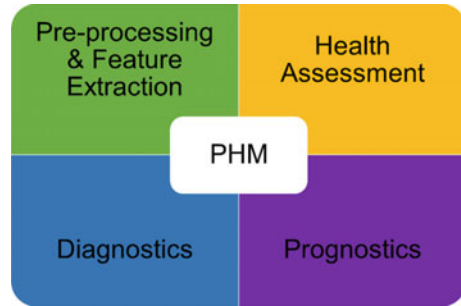
However, the value of PHM does not stop at maintenance alone. By performing smart analytics to asset usage data, users would be able to gain knowledge about how to achieve optimized performance of the asset. For instance, the State of Health (SoH) and Remaining Useful Life (RUL) of batteries on electric vehicles are highly dependent on the driving behavior. By analyzing the relationship between driving behavior and battery condition, a customized solution could be provided for the improvement of user's driving behavior and thus prolong battery life. Also, by relating process data with product quality metrics, predictive error compensation can be realized for increased product quality assurance. Additionally, asset usage and failure analysis could be fed back to the designers and manufacturers of the asset to nurture customer co-creation for an improved product design.

1.2 Research in Data-Driven Prognostics and Health Management

Besides huge economic potential in various industrial applications, PHM also holds great research value in the fields of signal processing, machine learning and data mining. Research in data-driven PHM requires an interdisciplinary background of computer science, signal processing, statistics, and necessary domain knowledge. In general, there are four major tasks for data analysis and modeling in PHM: pre-processing and feature extraction, health assessment, diagnostics, and prognostics, as indicated in Fig. 1. Prior to doing such tasks, it is critical to perform an overall analysis of the system to find out its critical components and the associated failure modes. Once the critical components of an asset have been determined, a data acquisition system needs to be devised to collect a sufficient set of measurements from the system for further analysis. Below is a description of the four data analysis steps for data-driven modeling of engineering systems:

- The task of pre-processing and feature extraction includes data quality evaluation, data cleaning, regime identification, and segmentation. Even though pre-processing does not directly offer immediate actionable information, it is a critical step and requires both domain knowledge and data processing skills to maintain the valuable parts of the data while removing its unwanted components.

Fig. 1 Research tasks of data-driven PHM analytics



- The task of health assessment consists of estimating and quantifying the health condition of an asset by analyzing the collected data. If there is data of failed condition, then a Confidence Value (CV) could be generated to indicate the probability of asset failure. However, if asset failure data are not available, health assessment could be transformed into either a degradation monitoring problem for gradual faults or a fault detection problem for abrupt faults.
- In PHM, diagnostics refers to the classification of different failure modes by extracting the fault signatures from the data. For rotating machinery, for example, this process consists of enhancing the signal-to-noise ratio of the vibration signals and extracting the cyclo-stationary components which can represent defects in certain components of the machine. A collection of various features can be used along with a clustering or classification algorithm for developing a data-driven model for machine fault diagnosis.
- The task of prognostics refers to the prediction of asset health condition. If a short-term prediction is desired, time-series modeling is often utilized to predict when the machine would go out of threshold. If a long-term prediction is preferred, then the problem becomes a remaining useful life (RUL) prediction with many existing machine learning and statistics tools available. A confidence range will need to be defined for such predictions as the performance of the machine is also highly dependent on the usage pattern and proper maintenance actions that will be taken.

Besides the aforementioned research topics, feature selection and dimension reduction is of vital importance to achieve better PHM results. Health management approaches such as maintenance scheduling and operation management are also within the scope of PHM discipline, but this introduction will only focus on the analytics aspect of PHM.

1.3 Methodology

In this section, a systematic approach for designing and implementing PHM for industrial applications is provided, as described in Fig. 2. The process is separated

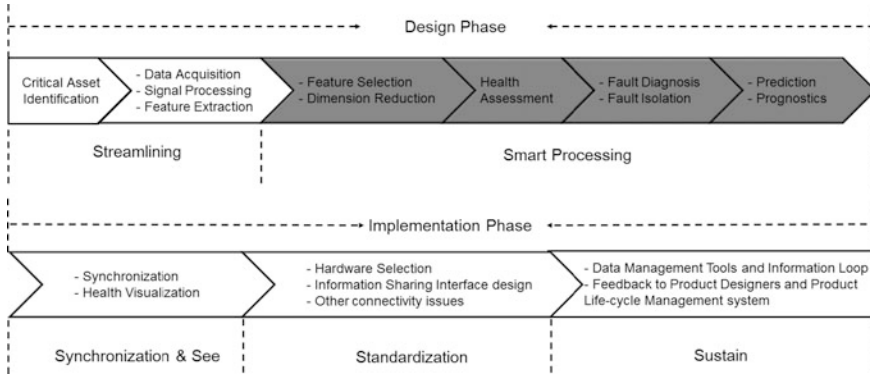


Fig. 2 General procedures for implementing PHM solutions

into five major steps following the “5S” methodology proposed in [3]. Within the 5S methodology, smart processing fulfills the major tasks of a PHM system and comprises the major intelligence of a system.

Smart processing focuses on utilizing data-to-information conversion tools to convert asset raw data into actionable information such as health indicators, maintenance recommendations, and performance predictions. All of this information is crucial for users to fully understand the current situation of the monitored asset and to make optimized decisions. Available data-to-information tools for smart processing includes: physics-based models, statistical models and machine learning/data mining algorithms. Among all three options, machine learning/data mining has its origins in computer practices and holds many advantages in industrial applications [4–6]: (1) Less domain knowledge requirements [7]: without building an exact mathematical model for the physical system, machine learning can also extract useful information by observing the input-output data pairs like the physical models do. This feature makes machine learning useful for complex engineering systems and industrial processes where prior knowledge is inadequate to build satisfactory physical models. (2) Scalable for a variety of applications [8]. (3) Easy implementation: compared with physics-based models, machine learning is more suitable to handle large-scale datasets since it requires less computation resources.

1.3.1 Algorithms

PHM algorithms refer to the data-driven models that serve as the computational core to transforming features into meaningful information. In Fig. 3, an example of this process for health assessment of induction motors is described. The process will be similar for diagnostics and prognostics, but the final visualization tool will be different depending on the purpose of users.

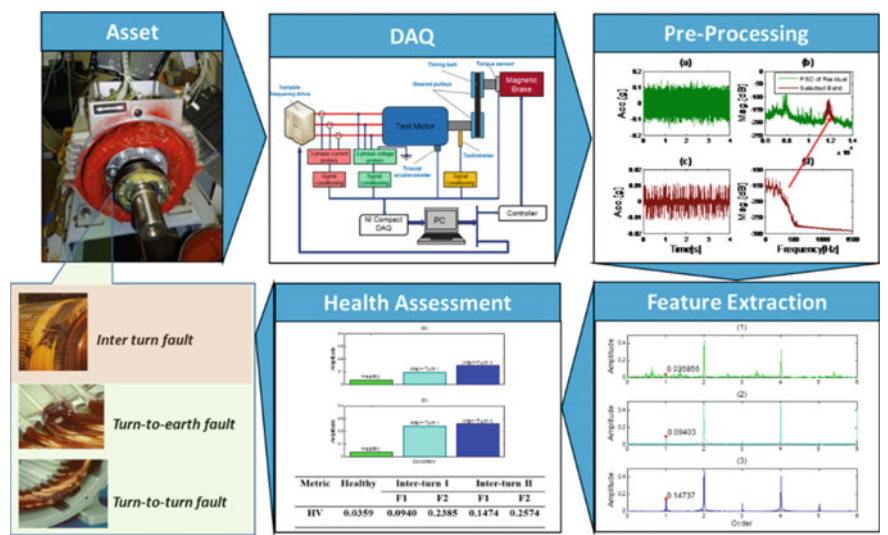


Fig. 3 Example of health assessment process [9]

Table 1 Summary of data pre-processing methods

	Method	Pros	Cons
1	Data quality inspection	Requires prior knowledge of signal type but effective in particular for vibration signal validation	Thresholds are needed for determining whether to include the signal in the analysis
2	Regime identification	Regime identification is important for developing baseline data sets in each operating condition	More sophisticated methods are needed for identifying the operating regime if the system changes operating conditions quickly
	Abnormality removal	Outliers, constant values, and missing values can dramatically increase the false alarm rate	Inclusion of domain knowledge is helpful for outlier detection

1.3.2 Data Pre-processing Algorithms

A summary of data preprocessing tools along with the merits and disadvantages of each method are provided in Table 1. In many applications, one or more of these methods is needed for ensuring that the data is suitable for additional processing and algorithm development. Data quality inspection is particular important to ensure that there are no sensor or data acquisition errors. However, developing an effective algorithm requires some prior knowledge of the signal characteristics or the distribution of the signal.

For more complicated working regimes, it might be necessary to have more advanced techniques for identifying the operating conditions. The regime

information allows one to develop baseline data sets in each operating condition and have a fair comparison and a local health model in each operating condition. Outlier removal from the measured data is also crucial in some of the applications as the presence of outliers can significantly affect the output of the analysis. Various methods exist for removing outlier instances from the signal or extracted features. However, these methods are purely based on the data distribution or characteristics, engineering experience and domain knowledge could be used to further improve the outlier removal algorithm.

1.3.3 Feature Extraction Algorithms

Numerous methods and algorithms are available for extracting characteristics or features from the measured signals, and an overview of some of the available techniques is provided in Table 2. For high frequency type signals such as vibration or current, there are well-established signal processing and feature extraction methods for extracting information from the time and frequency domain representation of the signal [10]. For rolling element bearings, mechanical shafts and gear wheels, there are several specific processing methods for extracting degradation features for these components [11]. Although it is advantageous to use the component specific feature extraction methods for high frequency vibration and current signals, they require a higher sampling rate, more computation, and more costly data acquisition systems.

For applications in which the monitored set of signals consists of trace signals such as temperature, pressure, or other controller signals, a different set of feature extraction algorithms would be recommended. Residual-based processing algorithms such as auto-associative neural networks, or principal component based

Table 2 Summary of feature extraction algorithms

	Method	Pros	Cons
1	Frequency based feature extraction methods	Frequency domain and envelope processing allows for component specific fault features to be extracted	Requires a higher sampling rate and more costly data acquisition
2	Residual based	More suited for low frequency signals and signals with a potential correlation	Residual processing algorithms can involve training a neural network which requires more computation
3	Statistics for each process segment or time slice	Ideal for process signals and provides a simple way of capturing the key aspects of the measured signal	Requires context information for identifying the various time slices of a process signal
4	Time statistics	Requires the least amount of domain knowledge and easiest to implement	Provides less specific information than other methods

methods are example methods that can be used to process trace signals [12]. For these types of algorithms, a baseline is established based on the normal operation of the machine. This baseline is used for comparing the predicted sensor values and processing the residuals as a sign of the drift in machine performance. The extraction of various statistical parameters is a straightforward but effective approach for characterizing the system condition from the available controller signals. In many instances, more insight can be gained by extracting statistics during different time slices, and a time slice could represent a different motion or action that is being performed by the monitored system [13]. If the context information regarding the process signals is not available, then extracting time statistics without any segmentation is a suitable alternative.

1.3.4 Health Assessment and Anomaly Detection Algorithms

A listing of the more commonly used algorithms for assessing machine health is provided in Table 3. The simplest approach for health assessment is to extract a health metric based on the weighted summation of the feature values. This health metric is simple to calculate and statistical thresholds for degradation detection can then be derived based on the distribution of the health value [14].

Various distances from normal health metrics can be used for determining the health condition of the monitored system or component. Mahlanobis distance and principal component analysis based Hotelling’s T^2 statistics are distance metrics that incorporate the covariance relationship among the variables; however, Euclidean and other distance metrics are also commonly used [15]. Distance based

Table 3 Summary of health assessment algorithms

	Method	Pros	Cons
1	Weighed combination of features	Simple to implement, easier for setting thresholds based on the health value distribution	Does not account for the correlation relationship in the features
2	Distance from normal	Requires only baseline data sets for training the algorithm, distance methods can also account for the variable covariance relationship	Does not account for whether the features are lower or higher than expected
3	Statistical hypothesis Testing	Simple to implement and can be used to test whether the system is in a normal condition	Data might not fit assumed distribution for the hypothesis testing
4	Regression methods	Provides a mapping between the features and an output defect level or health value.	Requires an output value that is related to the health condition of the system and multiple data sets for training
5	One-class classifiers	Support vector data description algorithms can provide a boundary for detecting anomalies	Requires experience on selecting the appropriate parameters and kernel function

methods do not account for whether the features are higher or lower than normal. This has potential drawbacks in that a system can have a lower vibration than normal and this would still trigger a higher distance based health value and an anomalous condition.

A simple but effective approach for anomaly detection is the use of statistical hypothesis testing. Sequential probability ratio test, rank permutation test, and a T-test, are all examples of some of the more commonly used hypothesis test for anomaly detection [16]. Other anomaly detection based methods include a one-class classifier, such as the support vector data description (SVDD) algorithm [17]. Regarding SVDD, one disadvantage is the lack of guidance in the literature on which kernel functions or settings to use for a given application. Regression or neural network based methods are particular effective if sufficient data is available for developing the regression models. A neural network or regression model can be used to provide a mapping between the feature values and a health value or defect size [18]. The ability for the model to generalize usually requires multiple training data sets.

1.3.5 Health Diagnostic Algorithms

For root-cause analysis and diagnosis, there are many different methods and algorithms to perform this task, and a sample of some of the more commonly used methods are listed in Table 4. Incorporating engineering knowledge and experience into the diagnostic algorithm makes the use of fuzzy membership functions and rules an attractive technique [19]. However, it becomes more challenging to use fuzzy based diagnostic algorithm for new applications in which there is not sufficient experience on the failure modes and their signatures.

The use of a classification algorithm is a popular alternative if there is data from multiple health states including a baseline condition and several of the different failure modes that can occur. The use of neural networks, support vector machines, and Naïve Bayes algorithm are some of the more common classification algorithms used for machine condition monitoring [20]. By learning the relationship between the extracted features and the baseline and failure signatures, the classification

Table 4 Summary of health diagnostics algorithms

	Method	Pros	Cons
1	Fuzzy membership rules	Can include engineering knowledge and experience in the diagnostic algorithm	Requires experience for determining the rules and membership functions
2	Machine learning classifier algorithm	Can learn the relationship between the feature values and the output health label	Requires data sets from each fault class for training the algorithm
3	Bayesian belief network	Models the cause and effect relationship between the feature values and various health states	Determining the BBN structure requires experience or learning the network structure from data

method can accurately diagnose and label the health condition from the monitored system. Another method for diagnostics includes the use of a Bayesian Belief Network (BBN) which provides a network representing the casual relationship between the measured variable and the different failure modes or system conditions that can occur [21].

1.3.6 Prognostics Algorithms

A sample of the more commonly used remaining useful life prediction algorithms is presented in Table 5, along with the advantages and disadvantages of each method. Curve fitting based methods are relatively simple to apply as they do not require a substantial amount of training data or a detailed physical model that describes the fault progression. Neural network or regression based methods can directly relate the feature values with the remaining useful life of the monitored component or system [22]. These methods require substantial training data for learning this relationship and obtaining multiple run-to-failure data sets is not feasible in many applications.

A similarity-based prognostic algorithm is a unique method that matches the previous degradation patterns to the current degradation pattern of the monitored system [23]. The similarity-based prognostic algorithm can be quite accurate; however, it requires several runs to failure data sets in order to obtain a library for performing the degradation trajectory matching. In contrast to the previously described data-driven prognostic algorithms, the incorporation of the physics of failure with a stochastic filtering algorithm is an effective approach. Whether the fault propagation dynamic equations are linear or non-linear, and also whether the measurement noise is Gaussian or not, this group of methods can be used [24]. Applying model-based prediction algorithms using stochastic filtering does have some potential challenges. Only for a subset of applications does one have established models for describing the failure mechanism.

Table 5 Summary of prognostics algorithms

	Method	Pros	Cons
1	Curve fitting methods	Simple to implement, does not require substantial training data sets	Results are dependent on selecting an appropriate curve fitting model form
2	Neural network methods	Provides a mapping between the feature pattern and the remaining useful life	Requires several run-to-failure data sets for learning this relationship
3	Stochastic filtering methods	Incorporates the failure physics and can handle uncertainties in the modeling and sensor data	Requires a physical model to describe the failure mechanism
4	Similarity based prediction method	Accurate and can account for different degradation patterns or initial degradation conditions	Requires several run-to-failure data sets

In the following section, a case study is provided which further elaborates on how a comprehensive PHM system can be applied to a real-world application and how the industry can benefit from such a system.

2 Case Study in Wind Turbine Monitoring System

2.1 Project Background

Wind power industry has been growing exponentially world-wide since 2000. By the year of 2012, there were more than 200,000 wind turbines operating, with a total nameplate capacity of 282,482 MW [25]. The US Department of Energy (DoE) claims that it is technically feasible to meet its goal of 20% of the total energy requirements by 2030, but this will involve extensive research in all aspects such as structural design, manufacturing, operation and maintenance, and construction [26]. In spite of the inspiring facts about wind power industry, there are also some hidden risks and concerns for all the players. In addition to the initial investment for turbine construction, operation and maintenance is estimated to take 20–25% of the total cost for on-shore turbines and 18% for offshore turbines over the lifetime, and can increase to 30–35% share of cost by the end of life [27].

Prognostics and health management (PHM) can play an important rule in promoting the development of wind energy and reducing the operation costs by ensuring wind turbines are more reliable and productive. According to *Wind System Magazine*, 70% of total wind turbine maintenance costs are from unscheduled breakdown. And for a 100 MW scale wind farm, only 1% of availability increase can worth between \$300–500 K of revenue per year. This view is also shared by the European Wind Energy Association (EWEA) to suggest that condition monitoring is a critical and integral system to the operation and maintenance [28].

By moving to condition based maintenance for wind turbine applications, there are numerous savings and benefits that can be provided as a service to the customer. The potential benefits include lower maintenance cost, a reduced risk of unplanned downtime, and higher asset utilization and uptime [3]. Different industries, such as semiconductor manufacturing, automotive, and machine tool among others, have benefitted from condition monitoring system integrated with advanced PHM tools. However, a similar level of advancement has not been developed in the wind turbine industry due to the very strict system integration requirements and low public acceptance.

The existing monitoring systems for wind turbines mainly fall into two categories: Supervised Control and Data Acquisition (SCADA) system and Condition Monitoring System (CMS). SCADA system has a variety of sensors to collect data from critical components and external environment. The data is used as the inputs for the control systems of pitch angle, yaw, and braking to name a few. CMS mainly have accelerometers and AE sensors mounted on the critical parts of

drivetrain and generator. Vibration level and related features are used to assess the health conditions of the components nearby. However, neither of the two systems are more than data collection system, and very limited and indirect information of operation risks is given to the operator to make optimal maintenance decisions.

2.2 *Benefits to Users*

The users that can directly benefit from the condition monitoring system are wind turbine OEMs and operators. For OEMs, the benefits include increasing the competitiveness and reliability of their wind turbines, reducing the maintenance and repair costs in warranty period, and with a minimum increase in prices. For operators, they can benefit from increasing availability of the turbines, reducing safety hazards, and reducing operation costs while increasing the revenue.

The benefits to turbine OEM and operators are summarized from the following aspects:

Increase Reliability of wind turbines (OEM, Operator): The condition monitoring system can benefit OEMs to increase the reliability of their wind turbines to increase the competitiveness of their product. This requires reducing the unplanned breakdown by detecting any incipient faults and repairing in the early stages of failure.

Reduce Warranty Costs (OEM): The warranty for wind turbines is usually 5-10 years with coverage of defects or faults in material, wear and tear, gradual deterioration, inherent vice, and latent defects. The price of an extended warranty ranges from \$30,000 per turbines per year for a 1.5 MW turbine to \$150,000 per turbine per year for a 3 MW turbine (Mike McMullen, *Wind Power Magazine*). However, the cost for warranty on the OEM's side is highly dependent on the failure rate of the components. Repair and replacement of critical components are usually very expensive. Gearbox replacement can cost \$250–350 K, generator replacement is \$90–120 K, and blade replacement can be \$120–200 K [29]. Hence, it is very important to monitor the critical components to avoid severe damage and to reduce warranty costs.

Increase Availability of Wind Turbines (Operators): For operators, it is ideal that the turbine can run 24/365 to maximize the revenue. It can be seen from simple match that for a 100 MW scale wind farm, only 1% of availability increase can be worth between \$300–500 K of revenue per year. The goal of increasing availability of wind turbines can be achieved from two aspects: decrease the failure rate of critical component, and decrease the maintenance and repair response time.

Reduce the Redundancy of Maintenance (OEMs and Operators): Whatever information or suggestions are given by the system should be accurate. False alarms and fault misdetection should be controlled to a very low limit. This requires the DAQ system and analytical modules to have a high level of accuracy and optimal settings for threshold.

2.3 Method Development

2.3.1 Identify Critical Subsystems/Components

According to the studies performed by Faulstich and Hahn [29], the component failure frequency and the average downtime per failure are listed in Fig. 4. The components that are suitable for predictive maintenance are those with a low failure rate but with a very high downtime per failure. Hence the components that will be included in the fault localization and diagnosis system are: the drivetrain, the yaw system, the pitch system the rotor blade, and other selected critical electrical components. The total breakdown of those components causes more than 80% of total unplanned downtime of the wind turbine.

2.3.2 Data Acquisition/Signal Selection

Most wind turbines are instrumented with supervised control and data acquisition system (SCADA), while CMS is installed based on customers’ (operator) requirement. As shown in Fig. 5, the SCADA system includes variables related to the operating condition of the key components. The sampling frequency is usually very low, and the output data is the statistical values such as mean, peak values and standard deviation. The sampling frequency of the SCADA system varies for different wind turbine settings, but is able to be customized according to data analysis requirement. For data storage, since the data volume is not very large, and SCADA

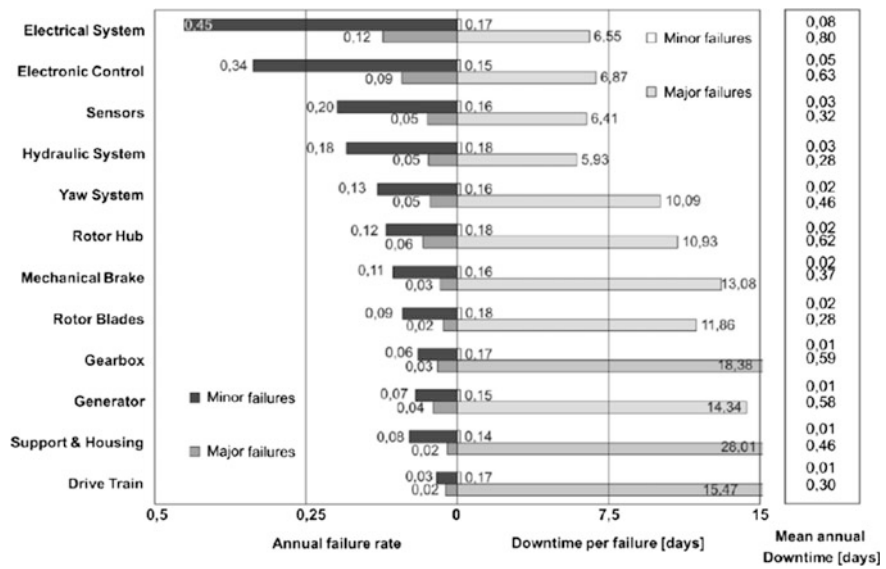


Fig. 4 Wind turbine failure rate and caused downtime [29]

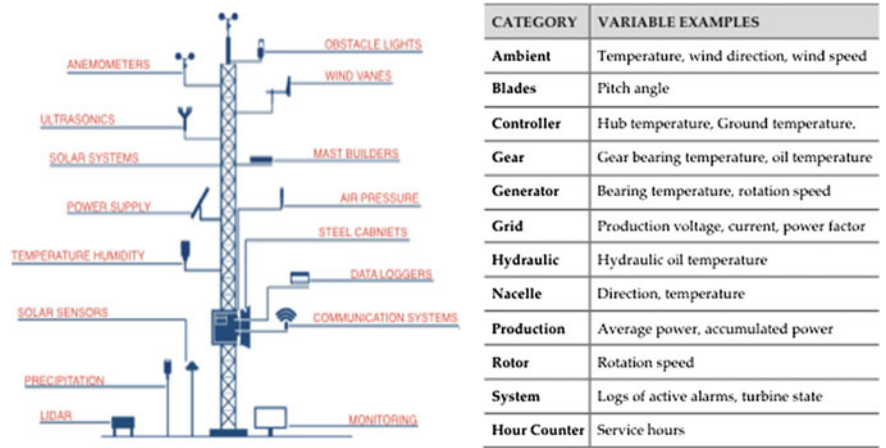


Fig. 5 List of SCADA variables [27]

Table 6 Design and operations summary of DAQ system

DAQ system	Sampling frequency	Collection interval	Data length	Data storage
SCADA	Range from 1/600–1/30	Continuous collection	1 sample per collection	From beginning of life
CMS	And from 1 K to 10 s KHz	Periodically or event based	From seconds to minutes	From beginning of life

data is important to provide reference information for maintenance actions, the historical data are always stored through lifetime.

The CMS data shall be available for fault localization and diagnosis for critical components. For data format, the sampling frequency determines the range of frequency spectrum while the data length determines the resolution of frequency spectrum in FFT analysis. The sampling frequency should be high enough to capture gear mesh frequencies and bearing characteristic frequencies for high-speed components in drivetrain, and the resolution of the frequency spectrum should be reasonably high. Since CMS collects data in very high frequency, and usually the monitored components are very stable and reliable, there is no need for continuous collection to reduce data storage and processing burden. Hence the data acquisition is usually triggered based on routine or event-based manners. Table 6 shows a summary for the DAQ system of wind turbine.

2.3.3 Multi-regime Modeling for Turbine Global Health Assessment

The global health of wind turbines shall reflect their generation efficiency, namely the efficiency to convert wind energy to electrical energy. Data from SCADA

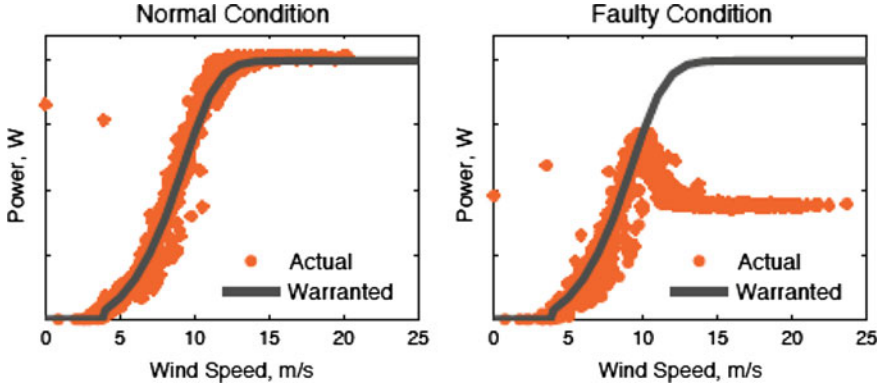


Fig. 6 Wind turbine power curve under normal and faulty conditions [30]

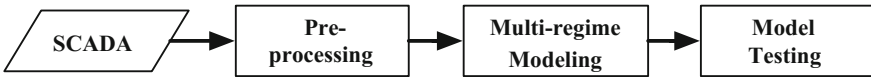


Fig. 7 Flow of wind turbine prognostic modeling

system shall be used to estimate the global health value. Figure 6 shows the power curve of wind turbines under normal and faulty conditions, and it is desired to quantify this kind of change and relate it to the loss of power efficiency.

SCADA parameters, including output power, wind speed, wind direction and pitch angle, is used to model turbine performance. Historical data is fed through a pre-processing module first to remove any outliers and undesired operating regimes for performance analysis. A multi-regime method is used to model the baseline behavior based on data from a selected training duration. Data from subsequent duration is then modeled and tested against the trained model, using distance metrics as a comparison method to quantify the deviation of testing data. Continuous testing can generate frequent evaluation of turbine performance and provide insight of turbine degradation over time with considerable time granularity, which could lead to valuable prediction (Fig. 7).

2.3.4 The Proposed Approach to Assessing Turbine Performance

SCADA variables are first fed into a pre-processing module to go through the following analysis.

1. Outlier filtering based on iterative Grubbs' test.
2. Rule-based non-operational regime filtering, where observations are filtered when wind speed is below cut-in and power output is zero.

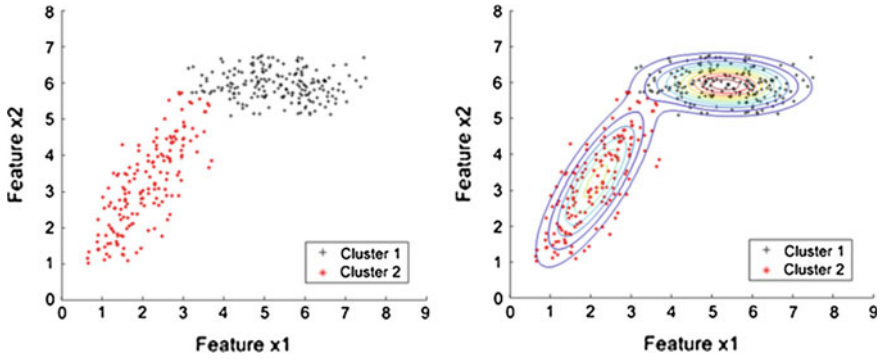


Fig. 8 Gaussian mixture model (GMM) based multi-regime clustering

3. Curtailment event filtering, where observations are filtered when a stalling event due to wind gust is determined to have occurred based on wind speed deviation and pitch angle deviation.
4. Data standardization, which equalizes variable contribution in the multivariate dataset (Fig. 8)

Gaussian Mixture Model (GMM), a probabilistic clustering model, is used to model the training data.

$$H(x) = \sum_{i=1}^n p_i h(x; \theta_i). \quad (1)$$

It partitions data into a mixture of Gaussian components with membership assignments where the component parameters and membership weights are estimated using techniques such as Expectation Maximization. The number of cluster, n , is decided based on the “goodness-of-fit” of the model when n is chosen as different numbers. A scoring method such as Bayesian Information Criterion (BIC) is used to evaluate model accuracy.

A testing model can be estimated with same GMM method for new data. An L2 distance between the two mixture models can be calculated, based on computing distance between all possible pairings of Gaussian components of the training and testing model. The normalized L2 distance is considered as a confidence value (CV) of turbine power performance.

$$\|H(x) \cdot G(x)\|_{L2} = \sum_{i=1}^n \sum_{j=1}^m p_i q_j \|h(x; \theta_i) \cdot h(x; \phi_j)\|_{L2}, \quad (2)$$

$$CV = \frac{\|H(x) \cdot G(x)\|_{L2}}{\|H(x)\|_{L2} \|G(x)\|_{L2}}. \quad (3)$$

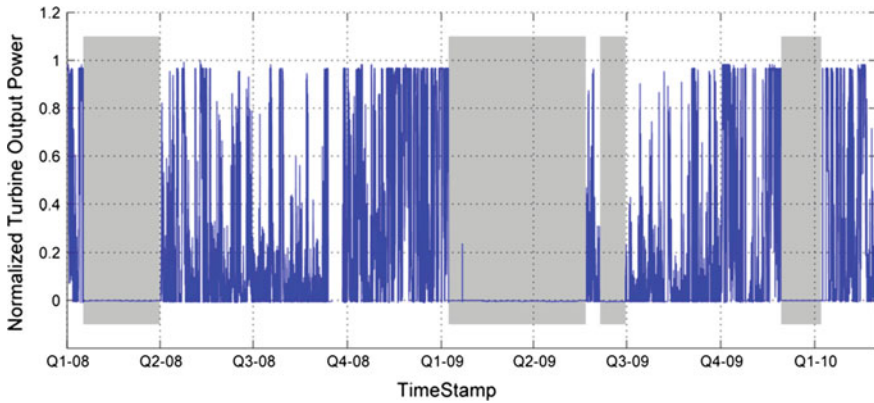


Fig. 9 Active power and failure events

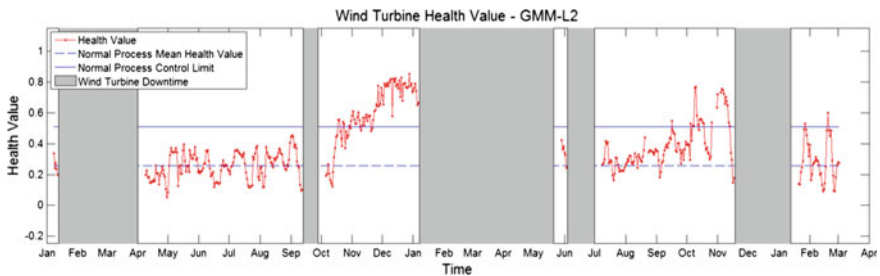


Fig. 10 Health risk increases between downtime

A SCADA dataset acquired from an onshore large-scale turbine is used to validate the proposed methodology for estimating the global health estimator (GHE). The duration of the data is 26 months, during which different parameters are extracted from the SCADA module every 10 min. The actual power output is shown in Fig. 6 where three major downtime events are highlighted in grey shadowed areas: (1) Q1-08 – Q2-08, (2) Q1-09 – Q3-09 and (3) Q4-09 – Q1-10 (Figs. 9 and 10).

2.3.5 Vibration-Based Condition Monitoring for Drivetrain System

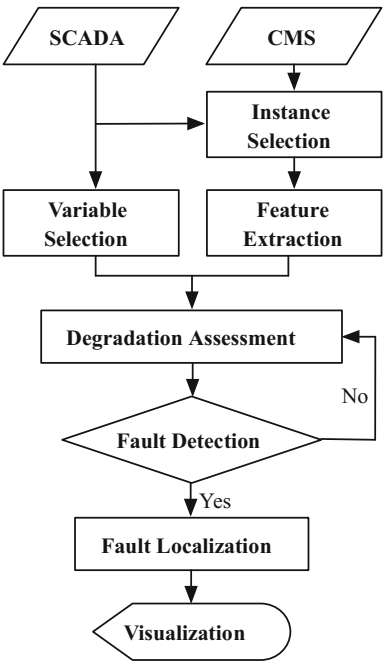
Prognostics techniques are mainly developed for key drive train components that cause costly and sometimes catastrophic failures, including rotor, gearbox and generator. In many applications, CMS and SCADA systems are used separately for condition monitoring purposes, mainly due to the issue of data availability shared by different shareholders. A degradation assessment framework is proposed to

integrate both the CMS data and SCADA variables for the evaluation of drivetrain degradation for the scenario when both data resources are available.

The framework of prognostics and diagnosis for drivetrain gearbox based on CMS vibration signals is shown in Fig. 11. In the framework, vibration signals first go through an automatic data quality check process to make sure the DAQ system is problem free. The data quality check process includes a series of check criteria as shown in Table 2. Afterwards, instance selection of vibration signals is performed based on the working regime variables from SCADA data. Vibration signals under certain conditions such as no energy generation will be removed. The selected vibration signals are then applied to the signal processing and feature extraction process to extract gearbox health related features. A collection of features will be extracted from the processed signals from frequency spectrum, TVDFT order tracing, spectral kurtosis filtering, Cepstrum analysis, and envelop analysis. Those features will be used as input for degradation assessment algorithms such as Self-Organizing Map Minimize Quantitation Error (SOM-MQE) and Principle Component Analysis techniques, so that the deviation of current condition from baseline model is quantified by distance metric.

In the training process, the input features are projected to output units by the weight vector in the mapping layer. The output units will compete with each other to determine the cluster of units with the best similarity, and the mapping layer will adjust the weight vector for the winning units. Hence when the training process is

Fig. 11 Framework for drivetrain components health assessment [31]



finished, the units in output layer will gather together to form several clusters, and each cluster corresponds to the same class of input data. If the training data are all under healthy condition, the units will gather to form one (single-regime) or several (multi-regime) clusters that represent the feature characteristics under healthy condition. In the testing process, the input feature vector (x_i) is projected into the output layer with the same weight vector, and its Euclidean distances to the units (w_j) are calculated based on Eq. 4. The unit that has the smallest distance to the projected vector is called the ‘Best Matching Unit (BMU)’, and the corresponding distance is called ‘Minimum Quantitation Error (MQE)’. The MQE value is calculated by (Table 7):

Table 7 Data quality check criteria [32]

Check method	Data processing	Check value	Threshold
Mean check	Mean value of vibration signal	Mean value	Smaller than $1e-5$ (should be decently small)
RMS check	RMS value of vibration signal	RMS value	$1e-5-0.05$ (minimum energy rule and dynamic range rule)
Parseval’s theorem-based Energy conservation rule	Time domain RMS and frequency domain RMS level should be close (conservation of energy for FFT)	$RMS(x(t)) - RMS(X(f))$	Smaller than 0.1%
Statistical distribution rule	Fit normal distribution of vibration signal	Hellinger-like distance and Komogorov distance of empirical and fitted distribution	<0.12 for K-distance <0.1 for H-distance
N-point rule	N neighbor points with the same value	N-point	Depends on sampling frequency. (<1 for our case due to very high sampling frequency)
U-point check	Number of unique points in the vibration signal	Portion of unique points to the length of dataset	$>99.99\%$ for our case
Positive and negative point check	Portion of positive and negative points to the length of dataset	$Max[P(+), P(-)]$	$<52\%$ for our case (the value should be close to 50%)
Derivative check	Derivative of vibration signal	RMS value of derivative signal; number of derivative value that exceeds threshold	0.015 for RMS derivative;

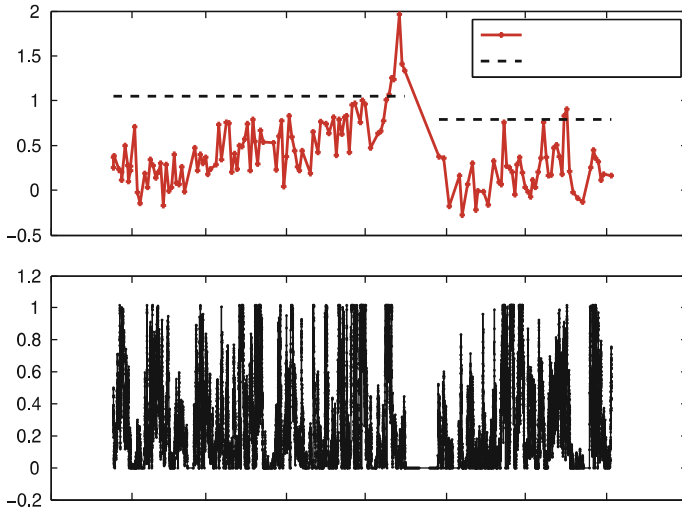
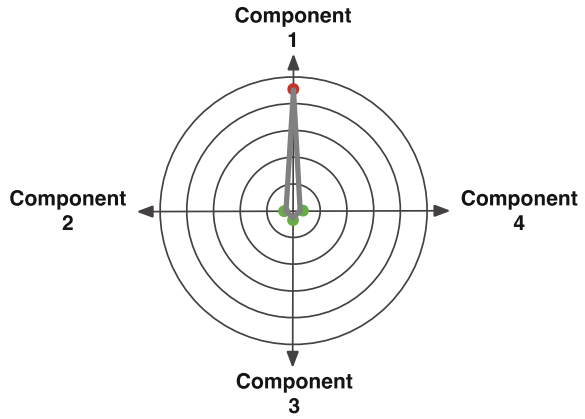


Fig. 12 Fault prognostic with SOM-MQE approach for offshore wind turbine drivetrain (*Note* Legend and axes labels were intentionally removed to keep the confidentiality of the data.)

Fig. 13 Radar chart for fault localization



$$\|x_i - w_c\| = \min_i \{\|x_i - w_j\|\}, \quad (4)$$

$$MQE = \|x - w_{BMU}\|. \quad (5)$$

This method is validated based on the historical data collected from an offshore wind turbine. The data span was 15 months, and had a severe damage at its rotor bearing that has caused downtime of 2 weeks. Figure 12 shows the change of MQE value before and after the breakdown. An incipient fault was detected five days before the severe damage.

As observed from the SOM-MQE result, there is a short duration in the middle of the history when MQE value noticeably exceeded the MQE threshold. The MQE excess occurred about five days before an operation pause due to a failure. The result shows that SOM-MQE is capable of detecting drivetrain anomaly at an early stage. A radar chart is created to view component criticality simultaneously. In this chart, each axis represents the contribution of each component to MQE abnormality. The closer the data point is to the center, the smaller the contribution is (Fig. 13).

3 Industrial Implementation and Gaps

3.1 Available Software and Platforms

As the industry recognizes the potential and business values in different sectors, a lot of companies have developed their PHM solutions/software/platform to achieve predictive modeling for industrial users.

- GE has announced Predix™ as a cloud-based service platform to enable industrial-scale analytics for management of asset performance and optimization of operations [33].
- National Instruments introduced Big Analog Data™ three-tier architecture solution [34], as well as LabVIEW Watchdog Agent™ Toolkit to support smart analytics solutions throughout different big data applications [35, 36].
- Many startup companies emerge recent years providing scalable PHM solutions, such as:
 - Predictrics (<http://www.predictrics.com/>) provides vertical solutions in various industrial applications from component level to fleet systems.
 - Uptake (<http://www.uptake.com/>) has been strategically working with Caterpillar and aims at developing a general PHM software.
 - Sparkcognition (<http://www.sparkcognition.com/>) has products concerning both cyber security and machine prognostics.
 - Trendminer (<https://www.trendminer.com/>) provides predictive analytics solutions to majorly process industry.

Besides, equipment makers themselves are also developing customized PHM systems for their own machines. For example, Prizm™ by Applied Materials for semiconductor manufacturing equipment [37], or RigWatch® by Canrig for their oil and gas applications [38].

3.2 *Gaps and Future Directions*

3.2.1 Preprocessing

“Industrial Big Data” is usually more structured, more correlated, more orderly in time and more ready for analytics [6]. This is because “Industrial Big Data” is generated by automated equipment and processes, where the environment and operations are more controlled and human involvement is reduced to minimum. Nevertheless, the values in “Industrial Big Data” will not reveal themselves after connectivity is realized by “Industrial Internet”. Even though machines are more connected and networked, “Industrial Big Data” usually possess the characteristics of “3B” [6], namely:

- Below-Surface
 - General “Big Data” analytics often focuses on the mining of relationships and capturing the phenomena. Yet “Industrial Big Data” analytics is more interested in finding the physical root cause behind features extracted from the phenomena. This means effective “Industrial Big Data” analytics will require more domain know-how than general “Big Data” analytics.
- Broken
 - Compared to “Big Data” analytics, “Industrial Big Data” analytics favors the “completeness” of data over the “volume” of the data, which means that in order to construct an accurate data-driven analytical system, it is necessary to prepare data from different working conditions. Due to communication issues and multiple sources, data from the system might be discrete and un-synchronized. That is why pre-processing is an important procedure before actually analyzing the data to make sure that the data are complete, continuous and synchronized.
- Bad-Quality
 - The focus of “Big Data” analytics is mining and discovering, which means that the volume of the data might compensate the low-quality of the data. However, for “Industrial Big Data”, since variables usually possess clear physical meanings, data integrity is of vital importance to the development of the analytical system. Low-quality data or incorrect recordings will alter the relationship between different variables and will have a catastrophic impact on the estimation accuracy.

Therefore, preprocessing and how to ensure data quality would be an important issue in PHM. The evaluation of data quality does not have to be limited to the inspection of signal validity, but can also include trend detection to evaluate the predictability, cluster analysis to evaluate potential for fault diagnosis, etc. [39].

3.2.2 Fleet-Based PHM

A fleet refers to a set of assets/machines that share some common characteristics that can be used to group them together according to a specific purpose. e.g. air crafts, vessels, wind turbines, trains, etc. Modern manufacturing enterprise scale is becoming larger and individual asset-based PHM might not be able to sufficiently fit in the changing environment in future.

Fleet-based PHM will be more accurate than conventional individual asset-based PHM:

- Prediction: similarity-based prediction
- Fault detection: peer-to-peer comparison, multiple kernel learning
- Compensation of training data insufficiency
 - Peer comparison without long history of individual baseline data for training.

3.2.3 General PHM Platform

Today's PHM solutions are still very customized and confined to one application. Different applications would have different data acquisition and storage system, different domain knowledge-dependent features and different monitoring purposes. It is very difficult to create a platform that could cover all kinds of applications.

One way of expanding the scope of a PHM solution is to combine several mainstream component/machine level PHM solutions together, and have users choose tools from similar applications. An alternative approach is to build up a standard platform where analytical tools are available but not customized. For such platform, background knowledge about how to use these tools for their application is required.

References

1. A. Alter, P. Banerjee, P. E. Daugherty, W. Negm, *Driving Unconventional Growth through the Industrial Internet of Things*, 2014
2. D.O. Gray, D. Rivers, *Measuring the Economic Impacts of the NSF Industry/University Cooperative Research Centers Program: A Feasibility Study*, 2012
3. J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mech. Syst. Signal Process.* **42**(1), 314–334 (2014)
4. M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms* (Wiley, 2011)
5. I.H. Witten, E. Frank, *Data Mining: Practical Machine learning tools and Techniques* (Morgan Kaufmann, 2005)
6. K.P. Murphy, *Machine Learning: a Probabilistic Perspective* (MIT press, 2012)
7. M. Pecht, R. Jaai, A prognostics and health management roadmap for information and electronics-rich systems. *Microelectron. Reliab.* **50**(3), 317–323 (2010)

8. Z. Ge, Z. Song, *Multivariate Statistical Process Control: Process Monitoring Methods and Applications* (Springer Science & Business Media, 2012)
9. C. Jin, A.P. Ompusunggu, Z. Liu, H.D. Ardakani, F. Petre, J. Lee, Envelope analysis on vibration signals for stator winding fault early detection in 3-phase induction motors. *Int. J. Progn. Health Manag.* **6**, 12 (2015)
10. A.K.S. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **20**(7), 1483–1510 (2006)
11. R.B. Randall, *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications* (John Wiley & Sons, 2011)
12. J.W. Hines, R. Seibert, Technical review of on-line monitoring techniques for performance assessment. *State-of-the-Art* **1** (2006)
13. G.A. Cherry, *Methods for Improving the Reliability of Semiconductor Fault Detection and Diagnosis with Principal Component Analysis*, 2006
14. E. Bechhoefer, D. He, P. Dempsey, Gear health threshold setting based on a probability of false alarm, in *Proceedings of Annual Conference of the Prognostics and Health Management Society*, 2011
15. H. Oh, M.H. Azarian, M. Pecht, Estimation of fan bearing degradation using acoustic emission analysis and Mahalanobis distance, in *Proceedings of the Applied Systems Health Management Conference*, pp. 1–12, 2011
16. R. Ganesan, A. N. V. Rao, and T. K. Das, A Multiscale Bayesian SPRT Approach for Online Process Monitoring, in *IEEE Transactions of Semiconductor Manufacturing*, vol. 21.3, pp. 399–412, 2008
17. D. Tax, A. Ypma, R. Duin, Support vector data description applied to machine vibration analysis, in *Proceedings of 5th Annual Conference of the Advanced School for Computing and Imaging (Heijen, NL)*, pp. 398–405, 1999
18. D. He, E. Bechhoefer, Development and validation of bearing diagnostic and prognostic tools using HUMS condition indicators, in *Proceedings of 2008 IEEE Aerospace Conference*, pp. 1–8, 2008
19. D.J. Cleary, P.E. Cuddihy, A novel approach to aircraft engine anomaly detection and diagnostics, in *Proceedings of 2004 IEEE Aerospace Conference*, vol. 5, pp. 3468–3475, (2004)
20. W. Yan, F. Xue, Jet engine gas path fault diagnosis using dynamic fusion of multiple classifiers, in *Proceedings of 2008 IEEE International Joint Conference on Neural Networks*, pp. 1585–1591, 2008
21. L. Yang, J. Lee, Bayesian Belief Network-based approach for diagnostics and prognostics of semiconductor manufacturing systems. *Robot. Comput.-Integr. Manuf.* **28**(1), 66–74 (2012)
22. N. Gebraeel, M. Lawley, R. Liu, V. Parmeshwaran, Residual life predictions from vibration-based degradation signals: a neural network approach. *Ind. Electron. IEEE Trans.* **51**(3), 694–700 (2004)
23. T. Wang, J. Yu, D. Siegel, J. Lee, A similarity-based prognostics approach for remaining useful life estimation of engineered systems, in *Proceedings of International Conference on Prognostics and Health Management*, pp. 1–6, 2008
24. M.E. Orchard, *A Particle Filtering-Based Framework for On-Line Fault Diagnosis and Failure Prognosis* (Georgia Institute of Technology)
25. S. Sawyer, K. Rave, *Global Wind Report—Annual Market Update 2012*, (GWEC, Glob. Wind Energy Council, 2013)
26. U.S. Department of Energy, *Wind Power Today 2010*, 2010
27. S. Sheng, P.S. Veers, *Wind Turbine Drivetrain Condition Monitoring—An Overview* (National Renewable Energy Laboratory, 2011)
28. P. Gardner, A. Garrad, L.F. Hansen, A. Tindal, J.I. Cruz, L. Arribas, N. Fichaux, *Wind Energy—The Facts Part 1 Technology* (EWEA, Garrad Hassan Partners, UK CIEMAT, Spain, 2009)

29. S. Faulstich, B. Hahn, P.J. Tavner, Wind turbine downtime and its importance for offshore deployment. *Wind Energy* **14**(3), 327–337 (2011)
30. E.R. Lapira, *Fault Detection in a Network of Similar Machines Using Clustering Approach*, (University of Cincinnati, 2012)
31. D. Siegel, W. Zhao, E. Lapira, M. AbuAli, J. Lee, A comparative study on vibration-based condition monitoring algorithms for wind turbine drive trains. *Wind Energy* **17**(5), 695–714 (2014)
32. A. Jabłoński, T. Barszcz, M. Bielecka, Automatic validation of vibration signals in wind farm distributed monitoring systems. *Measurement* **44**(10), 1954–1967 (2011)
33. General Electric, Predix. <https://www.ge.com/digital/predix>
34. National Instruments, Big Analog Data™ Solutions. <http://www.ni.com/white-paper/14667/en/>
35. Center for Intelligent Maintenance Systems, Development of Smart Prognostics Agents (WATCHDOG AGENT®). <http://www.imscenter.net/front-page/Resources/WD.pdf>
36. National Instruments, Watchdog Agent™ Prognostics Toolkit for LabVIEW—IMS Center. <http://sine.ni.com/nips/cds/view/p/lang/en/nid/210191>
37. Applied Materials, Applied TechEdge™ Prizm™. <http://www.appliedmaterials.com/media/documents/techedge-prizm-overview>
38. CANRIG, RigWatch® Instrumentation and Equipment Condition Monitoring. <http://www.canrigdrillingtechnology.com/rigwatch.php>
39. Y. Chen, J. Lee, *Data Quality Assessment Methodology for Improved Prognostics Modeling* (University of Cincinnati, Cincinnati, OH, 2012)

Probabilistic Prognostics and Health Management of
Energy Systems

Ekwaro-Osire, S.; Goncalves, A.C.; Alemayehu, F.M.
(Eds.)

2017, X, 277 p. 121 illus., Hardcover

ISBN: 978-3-319-55851-6