Int J Adv Manuf Technol (2010) 50:297–313
DOI 10.1007/s00170-009-2482-0

ORIGINAL ARTICLE

Current status of machine prognostics in condition-based maintenance: a review

Ying Peng · Ming Dong · Ming Jian Zuo

Received: 9 March 2009 / Accepted: 10 December 2009 / Published online: 6 January 2010

Abstract Condition-based maintenance (CBM) is a decision-making strategy based on real-time diagnosis of impending failures and prognosis of future equipment health. It is a proactive process that requires the development of a predictive model that can trigger the alarm for corresponding maintenance. Prognostic methodologies for CBM have only recently been introduced into the technical literature and become such a focus in the field of maintenance research and development. There are many research and development on a variety of technologies and algorithms that can be regarded as the steps toward prognostic maintenance. They are needed in order to support decision making and manage operational reliability. In this paper, recent literature that focuses on the machine prognostics has been reviewed. Generally, prognostic models can be classified into four categories: physical model, knowledge-based model, data-driven model, and combination model. Various techniques and algorithms have been developed depending on what models they usually adopt. Based on the review of some typical approaches and new introduced methods, advantages and disadvantages of these methodologies are discussed. From the literature review, some increasing trends appeared in the research field of machine prognostics are summarized. Furthermore, the future research directions have been explored.

Keywords Condition-based maintenance · Prognostics

1 Introduction

The modern industry is increasingly demanded to work at high reliability, low environmental risks, and human safety while operating their processes at maximum yield. Technological development has resulted in increased complexity in both industrial machinery and production systems. It is difficult or almost impossible to identify and predict failure conditions in a timely manner. In industry systems, machines' breakdowns usually limit uptime in critical situations. The economical consequences from an unexpected 1-day stoppage in industry may become as high as up to 100,000 to 200,000 euros [1]. Operational reliability of industrial machinery and production systems has a significant influence on the profitability and competitiveness of industrial companies. This emphasizes the increasing importance of effective maintenance strategies of machinery, production processes, and systems in industry. It is critical to have a sound maintenance management system, which can control its maintenance costs at the lowest level and maintain its overall equipment effectiveness at the highest level [2].

Time-based maintenance (TBM) is a periodic preventive maintenance. In a TBM strategy, some primary preventive maintenance is carried out periodically, such as lubricating, refurbishing, calibrating, and inspecting equipment on a regularly scheduled basis. The goal of TBM is to slow down the deterioration processes leading to faults. TBM assumes that the estimated failure behavior of equipment,

Y. Peng · M. Dong (✉)
Antai College of Economics & Management,
Shanghai Jiao Tong University,
535 Fahua Zhen Road,
Shanghai 200052, People's Republic of China
e-mail: mdong@sjtu.edu.cn

M. J. Zuo
Department of Mechanical Engineering, University of Alberta,
Edmonton, AB T6G 2G8, Canada

e.g., mean time between failures is statistically or experientially known for equipment, and the system is degrading in normal usage [3]. TBM will also encounter some minor or major planned shutdowns of systems for predetermined overhaul or repair activities on still functioning equipment. This can prevent functional failures by replacing critical components at regular intervals just shorter than their expected residual useful lifetime. System overhaul and critical item replacement at fixed intervals are widely adopted in automated manufacturing and control systems. Although TBM can reduce the probability of system failures and the frequency of unplanned emergency repairs, it cannot eliminate the occurrences of random failures completely. Some TBM practices may be out of date and cannot satisfy the actual operating requirement of the modern industry. In TBM maintenance strategies, most decisions are made by experienced planners according to suggestions from equipment fabricant, historic breakdowns or failure data, operating experience, and judgment of maintenance staff and technicians. Under an uncertainty situation, it is very difficult to make a maintenance schedule properly in advance [4]. On the economy aspect, TBM tends to be too conservative that results in very high maintenance costs. In the literature on TBM strategies, the prognostic technology has been few referred.

Condition-based maintenance (CBM) is a decision-making strategy to enable real-time diagnosis of impending failures and prognosis of future equipment health, where the decision to perform maintenance is reached by observing the "condition" of the system and its components. The condition of a system is quantified by obtaining data from various sensors in the system periodically or even continually. CBM attempts to avoid unnecessary maintenance tasks by taking maintenance actions only when there is evidence of abnormal behaviors of a physical asset. It is a proactive process, which requires the development of a predictive model that can trigger alarm for corresponding maintenance. Generally, CBM can be treated as a method used to reduce the uncertainty of maintenance activities and is carried out according to the requirements indicated by the equipment condition. Figure 1 shows the relationship between the maintenance cost, residual useful life (RUL), and reliability of the system. When time to failure equals zero, the system will go into breakdown status. As the time to failure of the system approaches to zero, the reliability of the system decreases. CBM with capability of precisely predicting RUL and reliability of the system is desirable since useful information can be provided for the decision of an economical maintenance schedule.

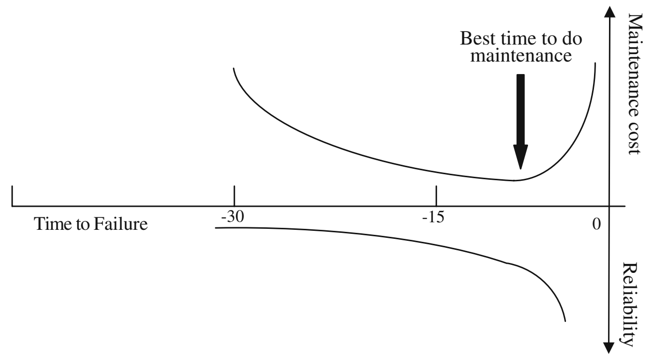


Fig. 1 The relationship between RUL, reliability, and maintenance cost

Diagnostics and prognostics are two important aspects in a CBM program. Diagnostics deals with fault detection, isolation, and identification when abnormity occurs. Prognostics deals with fault and degradation prediction before they occur. Prognostic algorithms for CBM have only recently been introduced into technical literature and become such a focus in the field of maintenance research and development. Soon a lot of prognostic algorithms were reported. However, many of them are more application specific rather than generic methods. For instance, prognostics has been an active area of research and development in aerospace, automotive, nuclear, process controls, and national defense fields. In many instances, these models could be loosely built based on the analysis of signals collected from system. In the future maintenance strategy, it may necessitate the development of prognosis models that can predict the future state of a system besides diagnosing the current state.

Prognostics addresses the use of automated methods to detect, diagnose, and analyze the degradation of physical system performance, calculating the remaining life in acceptable operating state before failure or unacceptable degradation of performance occurs. The major role of degradation analysis is to investigate evolution of physical characteristics, or performance measures, of a product leading up to its failure. Being able to perform precise and reliable prognostics is the key of CBM for an engineering system, and it is also critical for improving safety, planning missions, scheduling maintenance, reducing maintenance costs, and down time. Even though it is hard to do a prognostics with an acceptable precise, prognostic methodologies based on CBM have attracted a lot of attentions recently.

There are many studies and development on a variety of methods and technologies that can be regarded as the steps toward prognostic maintenance that are needed in order to support decision making and manage operational reliability (see Fig. 2). Some review works on prognostics have been done recently. Katipamula and Brambley [5] and Jardine et al. [6] made some simple review on the current status of prognostics, but their review emphases are mainly on diagnostics. Schwabacher and Goebel [7] made a survey of artificial intelligence (AI) for prognostics based on their

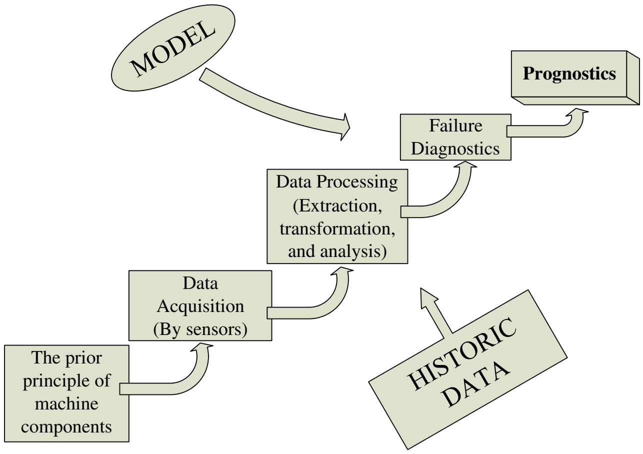


Fig. 2 Technology steps to form the basis of prognostic maintenance

new definition of AI and machine learning. Zhang and Li and Kothamasu et al. attempted to summarize recent research and development of fault prognostics in CBM [8, 9]. Based on the review of some literature, they proposed that future challenges need to be addressed in practical applications, such as development of fast and precise prognostic approaches, establishment of efficient validation approaches, and development of prognostic software toolkits. Goh et al. introduced some typical techniques used in prognostic maintenance and made a strengths, weaknesses, opportunities, and threats analysis of current prognostic research [10]. In the previous review literature, referred techniques and methods are limited to some traditional and typical ones such as artificial neural network (ANN), fuzzy logic (FL), and expert system (ES). In recent years, a large amount of papers, including theories and practical applications used in machine prognostics, appear in conference proceedings and academic journals. Many traditional methodologies used successfully in other areas or new methodologies have been introduced into the prognostic field. To improve the performance of prognostic models, some efforts that combine two or more techniques and methods together to build combination models have been done by many researchers. In this paper, prognostic models are divided into four categories: physical model, knowledge-based model, data-driven model, and combination model. Various techniques and algorithms have been reviewed by this category, depending on what models they usually adopt. Based on the review of some typical approaches and new introduced methods, advantages and disadvantages of these methodologies are discussed. The remaining part of the paper is organized as follows: Sections 2, 3, and 4 review and analyze physical model, knowledge-based model, data-driven model, and related methodologies, respectively. Section 5 reviews some existing combination/hybrid prognostic models that combine two or more techniques and methods together for machine prognostics. Finally, Section 6 concludes the paper by summarizing some increasing trends appeared in this area, pointing out some existing problems, exploring research directions and possible development trends needed for the future prognosis maintenance.

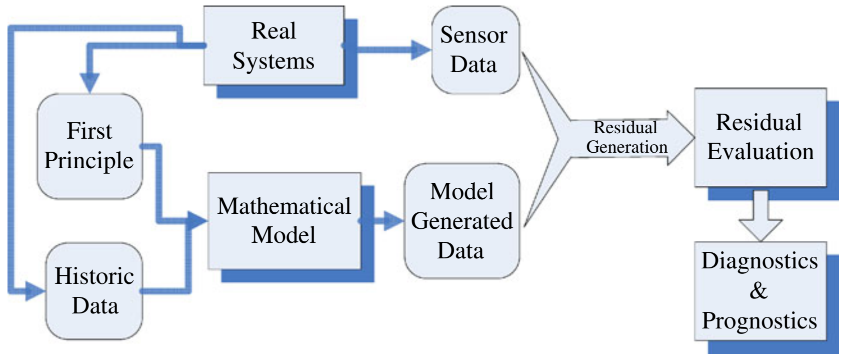
2 Physical model-based methodology

Physical model-based approaches usually employ mathematical models that are directly tied to physical processes that have direct or indirect effects on health of related components. Physical models are usually developed by domain experts, and the parameters in the model are validated by large sets of data. Physical model-based approaches used for prognostics require specific mechanistic knowledge and theories relevant to the monitored systems.

Physical model-based approaches use residuals as features, where the residuals are outcomes of consistency checks between sensed measurements of a real system and outputs of a mathematical model (see Fig. 3). The premise of such approaches is that the residuals are under a preset scale in the presence of normal disturbances, noise, and modeling errors and will exceed the scale in the presence of malfunctions. Statistical techniques are used to define thresholds for detecting the presence of faults. Physical model-based approaches are applicable in situations where accurate mathematical models can be constructed from first principles. These models must be configured specifically for the systems being monitored and should accurately simulate the responses of these systems for given command signals. In addition, the developed models can also be used to simulate component failures, which will not incur any cost associated with seeding faults on or in real hardware.

Physical models are useful in accounting for different operating conditions. With an intelligent monitoring system, most often, they work well under any load profile, including steady-state and transient performance and unanticipated conditions, loads, and operational regimes. Since they incorporate physical understanding of the system for monitoring, in many situations, the changes in feature vectors are closely related to model parameters. Therefore, a functional mapping between drifting parameters and selected prognosis features can be established. Moreover, if the understanding of system degradation improves, these models can be adapted to increase their accuracy and address subtle performance problems. The limitations are their higher costs and component speciality, which means that they cannot be applied to other types of components [11]. Furthermore, it is very difficult to build a good physical model.

Fig. 3 General flowchart of a model-based approach CBM system



Li et al. introduced a defect propagation model by mechanistic modeling approach for RUL estimation of bearings [12]. Engel et al. discussed some practical issues regarding accuracy, precision, and confidence of the RUL estimates [13]. Kacprzynski et al. utilized statistical physics-of-failure models to assess and predict gas turbine compressor performance degradation rates due to salt deposit ingestion [14]. Oppenheimer and Loparo applied a physical model to predict the machine condition, and in combination with fault strengths-to-life models based on crack growth law, it could be used to determine remaining machine life [15]. Luo et al. developed an integrated prognostic process based on data collected from model-based simulations under both nominal and degraded conditions [16]. Interacting multiple models is then used to track the hidden damages. Kacprzynski et al. proposed a helicopter gear prognosis method that fuses physics of failure modeling and relevant diagnostic information [17]. Byington et al. developed a specifically configured dynamic model for flight actuator fault detection and failure prediction [18]. This model-based prognostic approach applies physical modeling and advanced parametric identification techniques, along with fault detection and failure prediction algorithm to predict the time-to-failure of system. A different way of applying physical model-based approaches to prognosis is to derive the explicit relationship between the condition variables and the lifetimes (current lifetime and failure lifetime) through mechanistic modeling for machines considered as energy processors subject to vibration monitoring and for bearings with vibration monitoring [19, 20].

3 Knowledge-based methodology

It is usually a tough task to accurately build a mathematical model for a physical system with prior principles in real-world applications. So the uses of physical model-based methodologies are limited. For such a reason, knowledge-based methodology requiring no physical model appears to be promising. Two typical examples of knowledge-based approaches are ES and FL.

3.1 Expert systems

ES has been used since middle of 1960s. It is suitable for problems that usually solved by human specialists. ES can be considered as a computer system that is programmed to exhibit expert knowledge in solving a particular domain problem. The performance could be evaluated by combining the power of computers with the laws of reasoning. ES stores the so-called domain knowledge extracted by human experts into computers in the form of rules, simulates the way human experts do thinking and inference, and then uses these rules to generate solutions. The process of building ESs involves knowledge acquisition, knowledge representation, and the verification and validation of models.

Recently, ES become one of the major playing fields of artificial intelligence, and it has traditionally been used for fault diagnostics and prognostic applications. It is usually used for diagnosing, interpreting, and monitoring systems, planning for predictive repair activities and maintenance. Rule-based ESs are useful in encapsulating explicit knowledge from experts. Usually, rules are expressed in the form: IF condition, THEN consequence. The condition portion of a rule is usually some types of facts while the consequence portion can be outcomes that affect the outside world. Furthermore, the outcomes can also be used to test other conditions or rules, or even add a new fact to the knowledge base. These rules can be specific domain rules or heuristic rules (rules of thumb) and can be chained together using logical operators [21].

However, it is difficult to obtain domain knowledge and convert it to rules [22], and once built, an ES cannot handle new situations that not covered explicitly in its knowledge bases. Furthermore, when the number of rules increases dramatically, the "combinatorial explosion" that is involved in computation problems will be caused.

Lembessis et al. built an on-line ES for fault prognosis, which continuously monitors the condition/health of industrial equipment [23]. Butler proposed an ES-based framework for incipient failure detection and predictive maintenance (FDPM) [24]. The FDPM system is comprised of several ES-related components and databases by using the mathematical and neural network models. It can assess the integrity of a power distribution system component and predict the maintenance needs. Biagetti and Sciubba designed a prognostic and intelligent monitoring expert system that can generate real-time information on the existence of severity faults, forecast on the future time for both detected and likely faults, and give suggestions about how to control the problem [25].

3.2 Fuzzy logic

FL is a problem-solving methodology that provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. FL is widely used in various systems ranging from simple, small, embedded microcontrollers to large, networked PC or workstation-based data acquisition and control systems. It has the ability to model system behavior in continuum mathematics of fuzzy sets rather than with traditional discrete values, coupled with extensive simulation, thus offers a reasonable compromise between rigorous analytical modeling and purely qualitative simulation. By using linguistic variables, FL provides a very human-like and intuitive way of representing and reasoning with incomplete and inaccurate information.

FL is a superset of conventional Boolean logic with extensions to cater for imprecise information. Words and phrases such as “high”, “low”, “very high”, and “not very low” are used to describe continuous, overlapping states. It enables qualitative and imprecise reasoning statements to be incorporated with rule bases and produces simpler, more intuitive, and better behaved models. The linguistic description of a system is much more effective and less specific than the numerical or mathematical description. So the introduction of concepts issued from FL theory allows managing uncertainty in maintenance scheduling processes. Then, robustness of fault detection and prediction methods is improved, and the prognosis may be more reliable.

FL is based on the principle that every crisp value belongs to all relevant fuzzy sets to various extents, called the degrees of membership. They range from 0 (definitely not a member) to 1 (definitely a member) with values generated by a membership function. It contrasts with conventional Boolean logic, where membership of a set could only be false or true, i.e., 0 or 1. The graduation from zero to one enables the boundaries between sets to be smoothed out and overlapped. FL allows crisp values to belong to more than one fuzzy set, which means that in a fuzzy system, all rules can be used and every rule has some influences on the resulting outputs, whereas in a crisp system, only one rule might be fired and used.

The application of FL in fault prognostics is usually incorporated with other techniques such as ES, Kalman filter, and ANN. Alarm filtering and diagnostic system, an on-line fuzzy expert system developed by Choi et al., besides providing clean alarm pictures and system-wide failure information during abnormal states, can also carry out alarm prognosis to warn operators of process abnormalities [26]. A prognostic adaptive system based on fuzzy pattern recognition principles was designed by Frelicot [27]. The fault detection is achieved by a fuzzy classification rule including membership rejection (nondetection) and ambiguity rejection (multiple detections). A self-learning process based on fuzzy clustering is activated periodically in order to extract new faults from membership rejected patterns, and the classification rule is adapted. A multistep adaptive Kalman filter was introduced to predict membership vectors that are suitable for prognosis.

4 Data-driven methodology

Data-driven prognostic methodology is based upon statistical and learning techniques, most of which originated from the theory of pattern recognition. Data-driven methods can be classified into two categories: statistical approaches and AI approaches. Statistical approaches include multivariate statistical methods (e.g., static and dynamic principal components analysis (PCA), linear and quadratic discriminant, partial least square, canonical variety analysis, and learning vector quantization (LVQ)), state space models (e.g., Bayesian networks, hidden Markov models (HMM), and hidden semi-Markov models (HSMM)), and regressive models. Now most of the existing data-driven AI approaches for prognosis have employed ANN and its variants (e.g., polynomial neural networks (PNN), dynamic wavelet neural networks (DWNN), self-organizing feature maps (SOM), multilayer perceptron (MLP) neural network [28]). Data-driven models are usually developed from collected input/output data. These models can process a wide variety of data types and exploit the nuances in the data that cannot be discovered by rule-based systems.

Ray and Tangirala used a nonlinear stochastic model of fatigue crack dynamics for the real-time computation of time-dependent damage rate and accumulation in mechanical structures [29]. Given the condition monitoring history to date, Wang used the residual delay-time concept and stochastic filtering theory to derive the residual life distribution [30]. Yan et al. employed a logistic regression model to calculate the probability of failure for given condition variables and an autoregression moving average time series model to trend the condition variables for failure prediction [31]. Then, a predetermined level of failure probability was used to estimate RUL. Banjevic and Jardine discussed RUL estimation for a Markov failure time process that includes a joint model of proportional hazard model (PHM) and Markov property for covariate evolution as a special case [32]. Goebel et al. used two data-driven models to deal with the collected data [33]. They first employed a simple data-driven routine to establish a baseline for battery health prediction performance and uncertainty assessment. Then, they explored a Gaussian process regression method, which is a probabilistic technique for nonlinear regression that computes posterior degradation estimates by constraining the prior distribution to fit the available training data, to estimate the RUL. The variance was provided around the mean predictions to describe associated uncertainty in the predictions. Gebraeel et al. developed a Bayesian updating approach that uses real-time condition monitoring information and reliability characteristics of a device's population to periodically update stochastic parameters in exponential degradation models [34]. They used the degradation models to develop a closed-form residual-life distribution for a partially degraded bearing.

In the following sections, some widely used or new introduced data-driven methodologies, such as ANN, Bayesian-related methods, HMM and HSMM, hazard rate (HR) and proportional HR, and gray model, will be reviewed extensively. Some other data-driven methods will also be referred.

4.1 ANN-based methodology

ANN is a data processing system that consists of three types of layers: input layer, hidden layer, and output layer. Each layer has a number of simple, neuron-like processing elements called "nodes" or "neurons" that interact with each other by using numerically weighted connections [35]. ANN can be used to establish a complex regression function between a set of network inputs and outputs, which is achieved through a network training procedure. There are two main types of training methodologies: (1) supervised training where the network is trained using a specified sequence of inputs and outputs and (2) unsupervised training where primary function of the network is to classify network inputs.

There are several types of neural networks such as DWNN, PNN, and SOM neural network. ANN-based approaches are attractive in the fields of prognostic study for their potential to enhance process speed and reduce complexity. These approaches reduce complexity by providing generic and reusable software and hardware modules applicable in a wide variety of modeling and decision-aiding (e.g., classification) applications. ANN does not rely on priori principles or statistics models and can significantly simplify the model synthesized process. ANN can readily address modeling problems that are analytically difficult and for which conventional approaches are not practical, including complex physical processes having nonlinear, high-order, and time-varying dynamics and those for which analytic models do not yet exist. It can increase fault tolerance through adaptation and can be self-modifying over life-cycle maturation to compensate.

Due to the ability to fuse numeric information from multiple and possibly disparate channels instantaneously, appropriately synthesized ANN can be used to implement accurate and fast on-line pattern recognition. Thus, data extracted from these systems could be brought jointly for the purpose of diagnostics and prognostics.

There are two types of applications of ANN for prognostics. One is to be used as a nonlinear function approximator to predict system failure features and trends by estimating and classifying applications. The other is to be used with feedback connections to model dynamic processes of system degradation and make expectation of the RUL [36].

It is usually a tough problem for system designers to fit domain knowledge to ANN in practical applications. Besides, prognostic process itself is a "black box" for developers, which means it is very difficult or even impossible to have physical explanations of the networks' outputs, and as ANN grows in size, training can become a complicated issue. For example, how many hidden layers should be included, and what is the number of processing nodes that should be used for each layer are confused questions for model developers [11].

Zhang and Ganesan used a self-organizing neural network for multivariable trending of fault development to estimate RUL of a bearing system [37]. Yam et al. applied a recurrent neural network for predicting the machine condition trend [38]. Byington et al. applied a neural network approach to do RUL estimation of aircraft actuator components [39]. The problem of prognostic uncertainty representation and management is studied by Khawaja et al. [40]. They introduced a novel confidence prediction neural network constructed with a confidence distribution node based on the idea of estimating joint probability density function. Parzen windows centered at points in the training data are used to represent uncertainty, and a learning algorithm implemented as Q-learning routine improves online prognostic estimates. Yu et al. presented a neural network model to predict behavior of a boring process during its full life cycle [41]. The prognosis is achieved by fusing the predictions of three principal components extracted as features from the joint time–frequency distributions of spindle loads' energy observed during the boring process.

DWNN belongs to a new class of neural networks with unique capability in addressing identification and classification problems. Wavelets are a class of basic elements with oscillations of effectively finite duration that make them look like "little waves". The self-similar, multiple resolution nature of wavelets offers a natural framework for the analysis of physical signals and images. DWNN has recently been proposed to address the prediction/classification issues. It must be trained and validated before any on-line implementation and use. DWNN can be trained in a time-dependent way, using either a gradient-descent technique like the Levenberg–Marquardt algorithm or an evolutionary one such as the genetic algorithm. The data used to train the predictor must be recorded with time information, which is the basis for prognosis-oriented prediction task. The features are extracted in temporal series and are dynamic in the sense that DWNN processes them in a dynamic fashion. Then, the obtained features are fused into time-dependent feature vector that characterizes the process at designated time instants. A trained DWNN, along with the RUL calculation mechanism, can act as an on-line prognostic operator. A drawback of DWNN is that a substantially large database is required for feature extraction, training, validation, and optimization.

Wang and Vachtsevanos introduced a prognostic framework based upon the concept of DWNN [42]. DWNN incorporates temporal information and storage capacity into their functionality so they can carry out fault prognostic tasks for future. The prognostic architecture is based on a static "virtual sensor" that relates known measurements to fault data and a predictor which attempts to project the current state of the faulted component into the future, thus revealing time evolution of the failure mode and allowing estimation of the component's RUL. Both constructs are based on a DWNN model acting as the mapping tool. Gebraeel and Lawley developed a DWNN-based degradation model that utilizes condition-based sensory signals to compute and continuously update residual life distributions of partially degraded components [43]. Initial predicted failure times are estimated through "supervised" trained DWNN using real-time sensory signals. These estimates are used to derive a prior failure time distribution for the component that is being monitored and update the prior distributions by using a Bayesian approach. The novelty of this methodology lies in the ability to update a component's remaining life distribution using in situ condition-based sensory signals. The real-time sensory signals capture the latest degradation state of the component, and the resulting updated distributions are directly linked to the physical degradation state of the component.

The near universality of multinomial in representing physical processes and dynamic systems has been demonstrated by Kolmogorov, working in the 1940s, and Gabor, working in the 1950s. In principal, many kinds of building-block elements may be used in modeling by induction. Although algebraic elements are most often used, other elements such as logic and transcendental functions, including the popular sigmoid, are also worthy candidates. PNN utilizes composition of Kolmogorov–Gabor (KG) multinomial or algebraic sum of terms. The KG multinomial is

$$y = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \sum_i \sum_j \sum_k a_{ijk} x_i x_j x_k + \dots \quad (1)$$

where x_i represents input data to the nodal element, y represents output result of the nodal element, and a_i is the relationship parameter between input nodal elements. Compared to MLP neural networks that utilize just the first two items, PNNs take advantage of higher-order and cross-coupling nonlinearities.

PNN was used for a generic fault detection, isolation, and estimation scheme for analyses of normal and defective vibration signatures in helicopter transmissions by Parker et al. [44]. Data from nine seeded-fault test-rig experiments, each corresponding to one of six different fault/no-fault conditions, are used to train and evaluate polynomial neural networks for pattern classification tasks. Features are generated using amplitude spectra of the time series vibration signatures. The algorithm for synthesis of PNN, a neural network software package that utilizes a constrained and minimum-logistic-loss criterion for multiclass problems, is used to perform pattern recognition tasks. By employing a multiple-look postprocessing strategy, perfect vibration signature classification could be achieved.

SOM neural network-based method is usually used to deal with the problem of feature space's construction and system degradation detection. It is trained iteratively after normalizing the input data [45]. In each training step, one sample vector X from the input data set is chosen randomly, and the distances between it and all the weight vectors of the SOM, which are randomly initialized, are calculated using some distance measures, such as Euclidian distance. The best matching unit (BMU) is the map unit, whose weight vector is the closest to X. After the BMU is identified, the weight vectors of the BMU and its topological neighbors are updated to make them be closer to input vector X. The vectors are updated according to the following learning rule:

$$m_i(t+1) = m_i(t) + \alpha(t) h(n_{\mathrm{BMU}}, n_i, t)(X - m_i(t)) \quad (2)$$

where m_i is a multidimensional weight vector for map unit i, $h(n_{\mathrm{BMU}}, n_i, t)$ is the neighborhood function monotonically decreasing with respect to the distance (n_{BMU} is the distance between X and BMU, n_i is the distance between X and i), and the training time; $\alpha(t)$ is the learning rate, which is a decreasing function of t with $0<\alpha<1$. At the end of the learning process, the weight vectors are grouped into clusters depending on their distance in the input space.

SOM can be used to process highly deviating, nonlinear data while in contrast to some traditional methods such as PCA. As a network based on unsupervised learning, SOM does not require that target values corresponding to input vectors are known so it can be used to cluster data without knowing the class membership of the input data [46]. SOM is a powerful tool for discovering and visualizing general structures of the state space, and therefore, it is an efficient tool for visualizing the system behavior and an effective tool for condition monitoring and system degradation detection. Huang and Xi used minimum quantization error indicator derived from SOM, which is trained by six vibration features, to assess performance degradation for a ball bearing life prediction [47]. SOM sets up an appropriate degradation indicator from the incipient defect stage of a ball bearing to its final failure.

4.2 Bayesian network-related methods

Bayesian networks arise from a synthesis of probability and graph theory. A Bayesian network is a directed acyclic graph that is constituted with nodes and directed lines. Bayesian networks can be described as $B(G, P_r)$, where G is a directed acyclic graph with directed lines that present the probability transition relationship between nodes and P_r is the conditional probability that can be decomposed on the basis of recursive product [48].

Dynamic Bayesian network (DBN) is a directed graphical model of stochastic processes that enables users to monitor and update system as time proceeds, or even predict further states of the system. The purpose of a DBN is to model probability distributions over semi-infinite collections of random variables that progress according to some temporal models. For a DBN, the following elements need to be defined: a prior network $P(X_0)$, a transition network $P(X_t|X_{t-1})$, and an observation network $P(Y_t|X_t)$. The prior network represents the prior probabilities for all variables in the network at the initial time slice t=0. The transition network illustrates what the probabilities are for each variable at any time slice t=1, 2,…,n. The observation network specifies dependencies of observation nodes regarding to other nodes at time slice t. Each variable X in a DBN is associated with a time slice t and denoted by X_t. A key characteristic of DBN is the number of time slices needed to model a particular problem, which is called time span T. Another key characteristic is the number of variables associated with each time slice, which is called slice size n. A DBN is often assumed to satisfy the following conditions: (1) it has the same structure at any time slice t, (2) the variables with cross-slice edges have Markov property from slice t to slice t+1, that is, the value at slice t+1 only depends on slice t, and (3) the set of variables and probabilities definitions are the same for each time slice, with the exception of the prior network in the initial time slice having its own probability distributions. Therefore, given the initial and transition networks, a DBN of any length can be constructed as needed. Recently, DBN has received increased attentions as a tool for modeling complex stochastic processes since it generalizes the popular HMM and Kalman filter.

Sheppard and Kaufman [49] developed a diagnostic approach based on Bayesian networks that incorporates information on failure probability, instrument uncertainty, and the predictions for false indication. They also used a DBN, which is an extension of the Bayesian network, to perform prognosis by modeling changes over time. Przytula and Choi suggested the use of Bayesian belief net (BBN) since prognostics and calculation of remaining life can be accomplished within the framework of BBN [50]. In the prognostic model developed by Gebraeel and Lawley, Bayesian approach is employed to update the prior distributions for estimation of subsequent failure times [34]. Dong and Yang investigated a DBN-based prognosis method to predict RUL for drill-bits [51]. They provided some specific steps for building a DBN-based prognosis model and corresponding inference algorithms. A prognosis procedure based on particle filtering algorithms is used to predict RUL of the drill-bits of a vertical drilling machine.

4.3 HMM and HSMM

HMM, which is a stochastic process model, is also a powerful tool for RUL estimation. HMM characterizes doubly embedded stochastic process with an underlying hidden stochastic process that can be observed through some probabilistic behavior; this is where its name "hidden" comes from. HMM is a parametric model; its parameters can be estimated by the vast experimental data using statistical techniques. HMMs have some distinct characteristics that are not possessed by some traditional methods. They could not only reflect the randomicity of machine behaviors but also reveal their hidden states, changing processes. Furthermore, HMMs have a well-constructed theoretical basis and easy to realize in software. But HMMs have some inherent limitations. One is the assumption that successive system behavior observations are independent. The other is the Markov assumption itself that the probability in a given state at time t only depends on the state at time $t-1$ is clearly untenable in practical applications.

Bunks et al. and Zhang et al. first pointed out that HMM-based models could be applied in the area of prognostics in machining processes [52, 53]. The principle of HMM-based prognostics is as follows: Build and train N HMMs for all component health states. Between N trained HMMs, it is assumed that the state transition time of estimated vectors follows some multivariate distribution. Once the distribution is assessed, the conditional probability distribution of a distinct state transition given the previous state transition points can be estimated. The coordinates of the points of intersection of the log-likelihood trajectories for different HMMs along the life/usage axis represent the estimated "state transition time instants". That is, the probabilty distribution for state transition time is estimated from the estimation of "state transition time instants". The overall shapes of actual log-likelihood plots do not resemble the ideal plots on which the "state transition time instants" are estimated. This makes the estimations of "state transition time instants" more difficult.

Baruah and Chinnam presented a novel method to employ HMM for carrying out both diagnosis and prognosis [54]. The method applies HMM for modeling sensor signals and, in turn, identifies the health states as well as facilitates estimation of remaining useful life. Zhang et al. introduced an integrated fault diagnosis and prognosis approach for bearing health monitoring and CBM [53]. The proposed scheme consists of PCA, HMM, and an adaptive stochastic fault prediction model. Camci proposed an integrated diagnostics and prognostic architecture that employed support vector machine (SVM) and HMM [55].

HSMM is constructed by adding a temporal component into the well-defined HMM structure. In order to cope with the inaccurate durational modeling of HMM, some authors have proposed to model explicitly the state duration [56, 57]. A common idea is to replace the duration probability density function with some well-chosen probability functions closed to the durational distribution of real-life applications. HMM with such an explicitly added state durational probability function is called HSMM, since the transition properties are no longer governed by a Markov process. It is like a HMM except each state can emit a sequence of observations. HSMM models the observations during the stay in state i as a whole. Hidden semi-Markov chains possess both the flexibility of hidden Markov chains for approximating complex probability distributions and the flexibility of semi-Markov chains for representing temporal structures. Dong et al. proposed a HSMM for fault classification application of UH-60A Blackhawk main transmission planetary carriers [58]. Soon they further developed their HSMM models for both equipment diagnosis and prognosis [59, 60]. An integrated framework based on HSMM for multisensor equipment diagnosis and prognosis is presented in the later work. In this framework, they used states of HSMM to represent health status of a component. The trained HSMM can be used to diagnose the health status of a component. Through parameter estimation of the health-state duration probability distribution and the proposed backward recursive equations, RUL of the component can be predicted.

4.4 Hazard rate and proportional hazard rate

In the prognostics of RUL, many applications involve the use of HR, which is one of the useful indicators in lifetime analysis. The mathematical properties of HR functions reveal a variety of features in data. Let T denote the time to failure of a component under consideration, with lifetime distribution function $F(t)$ and reliability function $R(t)$, where $F(t)+R(t)=1$, $F(t)\in[0, 1]$. Assume that $F(0)=0$ and density function $f(t)=F'(t)$ exists, then HR function can be defined as

$$\lambda(t) = \lim_{\substack{N \to \infty \\ \Delta t \to 0}} \frac{\Delta n(t)}{[N-n(t)]\Delta t} = \frac{dn(t)}{[N-n(t)]dt}$$
$$= \frac{dn(t)/N}{1-F(t)} = f(t)/R(t) \tag{3}$$

where N is the total number of sample items, $n(t)$ is the number of items that fail before time t, and $\Delta n(t)$ is the number of items that fail in the time interval $(t, t+\Delta t)$ [61]. The RUL function $\mu(t)$ is the expected time remaining to failure, given that the system has survived to time t, then

$$\mu(t) = E[T-t|Tt] = (1/R(t)) \int_t^{\infty} R(x)dx \tag{4}$$

where $R(t)>0$. $\lambda(t)dt$ can be approximated as the conditional probability of failure in the time interval $(t, t+\Delta t)$ given survival to time "t" [62].

Wang [30] proposed a stochastic process, called gamma process, with HR as its mean for prediction of residual life. The condition information considered was the expert judgment based on vibration analysis. Victor et al. carried out a lifetime analysis for the generalized Birnbaum–Saunders distribution based mainly on the hazard function of their modeled system [63].

However, HR of a part or a system is influenced not only by the time it is operating but also by the covariates under which it operates. The PHM was introduced in order to estimate the effects of different covariates influencing the RUL of a system [64]. The model has been used rather extensively in biomedicine, and recently, interest in its application in reliability maintenance has increased. Since a frequently occurring problem in the analysis of reliability data is that not all parts of the data have been collected under similar conditions, PHM introduced the definition of covariate into the equation of HR. It is assumed that there exists an arbitrary and unspecified baseline HR $\lambda_0(t)$ that depends on time only. HR $\lambda(t)$ of a system is the product of $\lambda_0(t)$ and a positive functional term $\psi(z, \beta)$ that is basically independent of time. This product incorporates the effects of a number of covariates such as temperature, pressure, and changes in system. Thus, $\lambda(t, z)=\lambda_0(t)\psi(z, \beta)$, where z is a row vector consisting of the covariates associated with the system and β is a column vector consisting of the regression parameters that is unknown in model. β defines the effects of the covariates. The covariates may influence HR so that the observed HR is either greater (e.g., in the case of poor maintenance) or smaller (e.g., a new or improved component of a system) compared to the baseline HR. Proportional HR can provide a variety of data-driven risk models which effectively capture the effects of the covariates.

Lloyd et al. presented a PHM that establishes a framework suitable for performing reliability estimates and risk prognostics on a complex multicomponent system that could transfer at arbitrary time among a discrete set of nonstationary stochastic environments [65]. They made modifications to PHM that accommodates time-varying stochastic covariates and implements the model in a nonlinear network context. The baseline HR comes from a parameterized reliability model developed from the empirical reliability estimates. Liao et al. introduced a proportional hazard model for hard failures and multiple degradation features of an individual component [66]. The model can predict the mean RUL of a component based on on-line degradation information. Li et al. developed an algorithm to extract the frequent failure signatures [67]. By coding the failure signatures as time-dependent covariates and interactions, a Cox proportional hazard model was built to effectively handle the situation of a long event sequence and a large number of event types in the sequence based on the frequent failure signatures. The proposed model can help proactively diagnose machine faults with a sufficient lead time before the actual system failures to allow the scheduling of preventive maintenance.

4.5 Gray model (1, 1)

Gray system theory was originally proposed by Deng [68]. It is able to effectively deal with incomplete data for system analysis, modeling, prediction, decision making, and controlling. Gray system theories have been successfully applied in many fields such as management, economy, engineering, and finance. In a gray system, its information is neither totally clear as in a white system nor totally unknown as in a black system. Gray systems set each stochastic variable as a gray quantity that changes within a given range. They deal directly with the original data and search the intrinsic regularity of data [69]. A gray forecasting model can be applied to the circumstances with the minimum data and its computation is relatively simple in MATLAB environment.

Gray model GM (1, 1) is a time series forecasting model. It has five basic operations: (1) obtain the original data series: $X^{(0)}$, (2) do accumulated generating operation (AGO) to obtain data series: $X^{(1)}$, (3) estimate parameters, (4) predict future point $X^{(1)}$, and (5) apply inverse accumulated generating operation (IAGO) to predict values for original data series $X^{(0)}$. Gray forecasting model uses accumulated operations to construct differential equations. For an initial time sequence $X^{(0)}=\{x^{(0)}(1),\ x^{(0)}(2),\ldots,\ x^{(0)}(i)\ldots x^{(0)}(n)\}$, where $x^{(0)}(i)$ is the time series data at time i. AGO is used to transform an original set of data into a new set that highlights trends but has less noise and randomness. The equation used in generating the AGO series is as follows:

$$X^{(1)}(k) = \sum_{i=1}^{k} X^{(0)}(i) \tag{5}$$

where $X^{(1)}=\{x^{(1)}(1), x^{(1)}(2),\ldots, x^{(1)}(i)\ldots x^{(1)}(n)\}$. After $X^{(1)}$ is obtained, the gray differential equation with one variable is built as:

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = b \tag{6}$$

where coefficients a and b express the relationship between dX/dt (how fast current state changes) and X (current state). a and b can be determined by the least-square method. Then, the predicted data points for the AGO series are calculated. Let $\underline{X}^{(1)}$ represent the predicted AGO series. IAGO is used to transform the forecasted AGO data series back into the original data series with equation $\underline{X}^{(0)}(i+1) = \underline{X}^{(1)}(i+1) - \underline{X}^{(1)}(i)$, where $\underline{X}^{(0)}$ is the predicted original data series.

Ku and Huang explored the application of gray forecasting models for predicting and monitoring production processes [70]. When a mean shift occurs, the gray predictors are found to be superior to the sample mean, especially if the number of subgroups used to compute the gray predictors is small. It is shown that the gray predictor is very sensitive to the number of subgroups. Gu et al. is trying to use a gray prediction model in the failure prognostics for electronics. His work is still in exploring stage [71].

4.6 Some other methods used in date-driven methodology

PCA is usually used to extract the significant characteristics of structural failures out of the sensors. Its main function is to retain the most important characteristics of its inputs by using a small amount of data. The advantage of PCA procedure is that it significantly reduces the dimension of input data and consequently enhances training speed and recognition speed. Furthermore, PCA has the self-learning capability. The traditional PCA has been extended by many researchers. Some PCA-related methods, such as dynamic principal component analysis, multiway principal component analysis (MPCA), and kernel principal component analysis (KPCA) have been developed. PCA is basically a static method, which assumes that observations at current time instance and observations at past time instances are independent. As an extension of static PCA method, dynamic PCA reveals dynamic relationships among process variables. By applying PCA to dynamic data matrix constructed by He et al., dynamic relationships among variables are extracted from data rather than a static approximation [72]. KPCA has emerged in recent years as a promising method for tackling nonlinear systems. KPCA can efficiently compute principal components in high-dimensional feature spaces by means of integral operators and nonlinear kernel functions. The basic idea of KPCA is to first map the input space into a feature space via nonlinear mapping and then computes principal components in that feature space. In comparison to other nonlinear PCA techniques, KPCA requires only the solution of an eigenvalue problem and does not entail any nonlinear optimization [73]. Nomikos and MacGregor have extended the multivariate statistical process control method of PCA to batch processes, where the method is called MPCA [74]. MPCA allows the monitoring of a batch process to be achieved once a model has been developed from nominal or good batch operations.

PCA and PCA-related methods have been applied to many applications in signal processing, image processing, and pattern recognition. Recently, it has been introduced into the prognostic systems for the data preprocess work. Kwan and Zhang employed PCA for the data extraction work in their fault diagnostics and prognostic model [75]. Zhang et al. used PCA for the principal signal features extraction work that could help to generate a component's health/degradation index as the input of an on-line RUL prediction system [53]. Lee et al. designed a consecutively updated MPCA model for on-line batch monitoring [76]. The key to realize "real-time" is that whenever a batch successfully remains within the bounds of normal operation, the batch data are added to the historical database of normal data and a new MPCA model is developed based on the revised database.

LVQ is a supervised learning technique that uses class information to move the Voronoi vectors (vectors that can characterize a class) so as to improve the quality of the classifier decision regions. An input vector is picked randomly from the input space. If the class labels of input vector x and Voronoi vector w agree, then w will move in the direction of x. On the other hand, if the class labels of input vector x and Voronoi vector w disagree, w will move away from x. Usually, LVQ is employed to generate a sequence of codes to represent the fault signatures that used in a prognostic system. LVQ was used to generate a sequence of codes for representing fault signatures in the model proposed by Zhang et al. [53]. The code sequence generated by vector quantization captures the temporal characteristics of the fault. In their simulation, 50 Voronoi vectors were used to represent each fault class (Table 1).

5 Combination model

In the real-world prognostic processes, the trends of all characteristic parameters are diversified and difficult to be predicted by using a single prediction method. Thus, a combination prediction method is adopted for prognostics. Using a well-designed condition-based combination prediction method that combines two or more prognostic approaches together for data extraction, data analysis and modeling may have following advantages: (1) the demerit of individual theory will be offset and the merits of all prediction methods could be utilized, (2) complexity of the computation may be reduced, and (3) the prediction precision could be improved.

The application of ANN is usually incorporated with knowledge-based techniques such as ES and FL. Brotherton et al. combined a dynamically linked ellipsoidal basis function neural network with rule extractors and applied it to gas turbine engine prognostics [11]. Garga et al. introduced a hybrid reasoning method that integrates machinery data into a feed-forward neural network trained basing on a simply representation of explicit domain knowledge for gearbox's health prognosis [21]. The proposed neurofuzzy algorithm is a combination of ANN and fuzzy inference system. They combine the linguistic description of a typical fuzzy inference system with learning procedures inspired by neural networks. These algorithms are particularly adaptive, lucid, robust, and highly flexible.

Wang et al. compared the results of recurrent neural network (RNNs) and neural fuzzy (NF) inference systems for predicting the fault damage propagation trend [77]. By comparison, it is found that if an NF system is properly trained, it will perform better than RNNs in both forecasting accuracy and training efficiency. Chinnam and Baruah presented a NF approach to estimate RUL for the situation where failure data and specific failure definition model are not available, but domain experts with strong experiential knowledge are available [78]. Satish and Sarma combined

Table 1 Major methods used for prognostics

Category	Methods	Supportive techniques	Application domains	Advantages	Disadvantages	References
Physical model	First principle modeling	Residual evaluation, parameter drift analysis	Bearing, gear, rotor shaft, flight actuator, gas turbine compressor	Might be used for different operating conditions without the need for recollecting data	Component/system speciality, hard to build model	Li et al. [12] Engel et al. [13] Kacprzynsk et al. [14] Oppenheimer and Loparo [15]
	Parameter estimation					Luo et al. [16] Kacprzynski et al. [17] Byington et al. [18] Cempel et al. [19] Qiu et al. [20]
Knowledge-based model	Fuzzy logic	Fuzzy set, fuzzy classification rule	Nuclear power plant, manufacturing system	Able to deal with vague, imprecise, or missing information, model system in continuum mathematics	Weight decision is a hard work	Choi et al. [26] Frelicot [27]
	Expert system	Domain rules, heuristic rules, logical operators	Energy conversion processes, power distribution equipment, manufacturing system	Solving problem by mimicking how human expert make decisions	Hard to obtain domain knowledge, hard to convert domain knowledge to rules	Lembessis et al. [23] Butler [24] Biagetti and Sciubba [25]
Data-driven model	ANN	BP, DWNN, PNN, SOM	Bearing, gearbox, power plant, aircraft actuator, gear plate	Model analytically difficult system, able to implement accurate and fast on-line pattern recognition	Hard to fit domain knowledge to ANN, model retraining is needed if operating conditions change, black box system	Zhang and Ganesan [37] Yam et al. [38] Byington et al. [39] Khawaja et al. [40] Wang et al. [42] Gebraeel et al. [43] Parker et al. [44] Huang and Xi [47]
		Bayesian network-related method	Bearing, drill-bits	Have a well-constructed theoretical basis, easy to predict further states	A lot of historical state transition and fatal data are needed	Sheppard et al. [49] Przytula and Choi [50] Gebraeel et al. [34] Dong and

Table 1 (continued)

Category	Methods	Supportive techniques	Application domains	Advantages	Disadvantages	References
						Yang [51]
	State space model	HMM, HSMM	Pumps, aircraft system, drilling process	Reveal the hidden states change processes, easy to realize in software	The assumptions in HMM are not practical in real world, HSMM relax the assumptions but complicate the model	Baruah and Chinnam [54] Zhang et al. [53] Camci [55] Dong and He [59, 60]
	Hazard rate, proportional hazard rate	Statistical techniques such as maximum likelihood function	Pumps, turbine	A general model without making much specific assumptions	The application is restricted by the assumption of "good-as-new after repair"	Wang [30] Victor et al. [63] Lloyd et al. [65] Liao et al. [66] Li et al. [67]
	Gray model	GM (1, 1)	Production process, electric system	Deal with incomplete information, good at smoothing time series data	A new introduced method in prognostics, more research and demonstration is need	Ku and Huang [70] Gu et al. [71]

neural networks and FL to form a fuzzy back propagation network for identifying the present condition of a bearing and estimating the RUL of a motor [79]. Xue et al. developed a fuzzy mathematical model with radial basis function neural network to predict the potential faults of a coal-fired boiler [80]. Kothamasu and Huang presented a NF modeling approach based on an adaptive learning Mamdani fuzzy model for system diagnosis and prognosis [81]. A robust and lucid modeling system that can assume the role of a decision making aid in the process of CBM is developed. The comprehensibility of the system is emphasized so it can effectively serve as a decision aid for domain experts and be adaptive to some ordinary modifications and even continuous improvement by interacting with users. The comprehensibility of a NF system usually deteriorates once rules are tuned. Kullback–Leibler mean information is introduced into the system to solve the problem. It can evaluate and refine tuned rules to make the system easily interpretable.

Since GM (1, 1) is good at smoothing time series data and ANN has stronger ability in nonlinear time series prediction than others, Dong et al. designed a multiparameter condition prediction model based on the combination of GM (1, 1) and BP neural network to predict the conditions of equipments in a power plant [82, 83]. By fully using (extraction and analysis) the operating data, condition monitoring data, and operation statistic data, the conditions of equipments are predicted, and the results are more reasonable than singe characteristic parameter prediction.

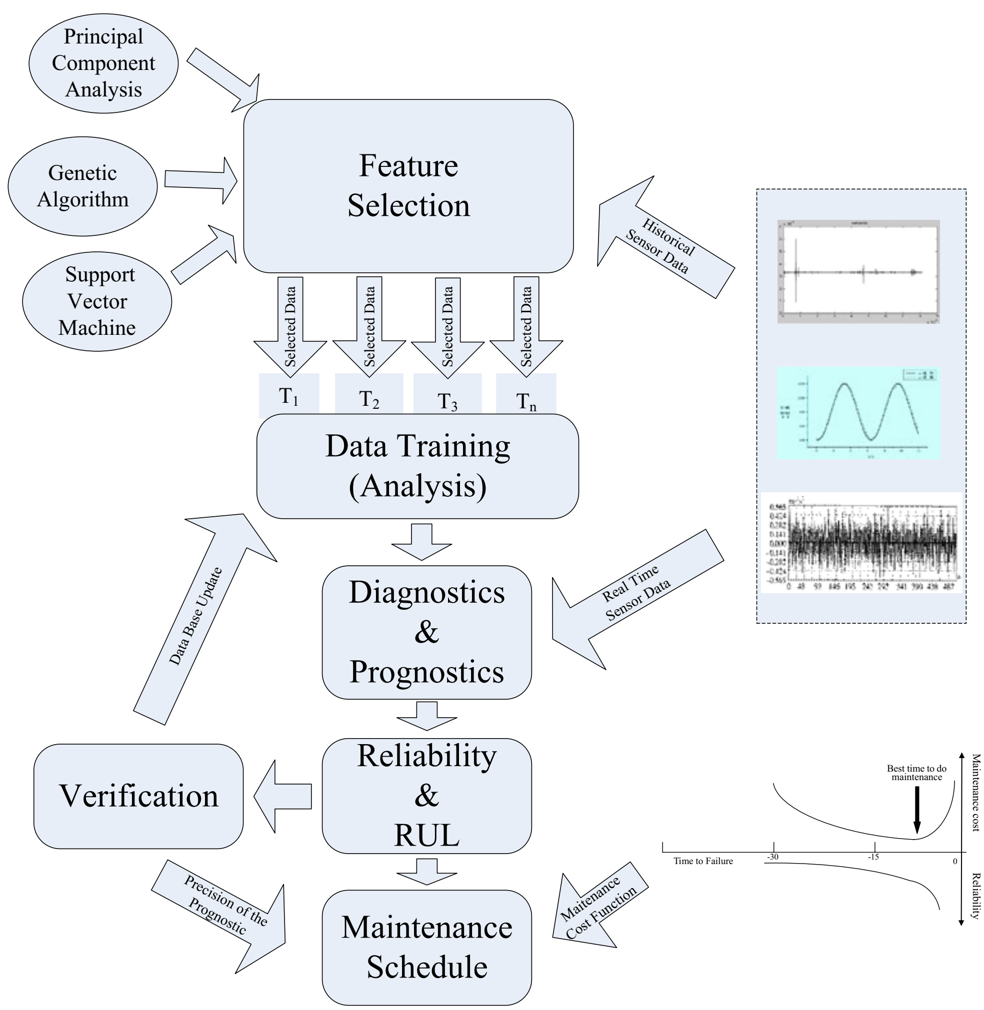
There are also some other combination models. Shetty et al. modeled a degrading system for aircraft auxiliary power units as a collection of prognostic states (health vectors) that evolve continuously over time [84]. The proposed multivariate state space model includes an age-dependent deterioration distribution, component interactions, as well as effects of discrete events arising from line maintenance actions and/or abrupt faults. Mathematically, the proposed model can be summarized as a continuously evolving dynamic model, driven by non-Gaussian input and switches according to the discrete events in the system. The system identification and recursive state estimation scheme for the developed non-Gaussian model is derived from a partially specified distribution framework. Mohanty et al. developed a hybrid prognosis model for real-time RUL estimation of metallic aircraft structural components [85]. The prognosis framework combines information from off-line physics-based, off-line data-driven, and on-line system identification-based predictive models. These components can be explicitly expressed by Gaussian processes, which are based on a data-driven approach within a Bayesian framework. The Gaussian process model projects the input space to an output space by probabilistically inferring an underlying nonlinear function for relating input and output. For the off-line prediction, the input space of the model is trained with parameters that affect fatigue crack growth. For the case of online prediction, the model input space is trained using features found from piezoelectric sensor signals rather than training input space with loading parameters, which are difficult to measure in a real flight-worthy structure. A new output space for corresponding unknown crack length or damage state can be predicted using the trained Gaussian process model.

6 Concluding remarks

This paper attempts to summarize the techniques and algorithms that appeared in recent research literature involved in machine prognostics for implementing CBM. Various techniques and algorithms have been categorized depending on what models are usually adopted. Due to the increasing maintenance cost, the tools to achieve precise prediction for maintenance strategy have received a great deal of attention recently. Many techniques and algorithms that used in other domain, such as pattern recognition and ailment prediction, have been introduced and modified for machine prognostics. From the literature review, some obviously increasing trends can be summarized as follows:

- More and more combination models are designed to deal with data extraction, data processing, and modeling for prognostics. From simple heuristic-based models to complex ANN models that impose AI knowledge, these methodologies have their own advantages and disadvantages. Since single-approach models have some difficulties in achieving satisfied results, it is a very challenging work to develop prognostic applications that can provide precise prediction. A well-designed combination model usually combines two or more theories and algorithms to model the system in order to eliminate the disadvantages of each individual theory and utilize the advantages of all combined methods. On the other hand, it is also a challenging work to choose appropriate methods and combine them together for modeling. Neutral networks have been combined with

Fig. 4 A systematic maintenance decision framework



many other data processing methods, such as ES, FL, and GM (1, 1). By the verification of practical data, these combination black-box models are demonstrated to be successful in the applications. HMM and HSMM are recently introduced methods in prognostics. Till now, the researches using HMM and HSMM for prognostics are in seedtime. Unlike black-box models, HMM and HSMM have a relatively explicit prediction process. Shao et al. developed a SVM–HMM method to identify nonstationary time series that occur when the plant proceeds to an abnormal state or a transient situation from a normal state [86]. Such HMM or HSMM combination models may be explored and designed for machine prognostics in future research.

- Many "new" theories and algorithms such as gray model, which have been successfully used in other domains before, have been introduced to deal with prognostics. These "new" theories and algorithms have their own advantages to deal with data processing and analysis and hence reduce the complexity of computation and increase the precision of the predictions.

Prognostics is still a relatively new research field in CBM. There exist some weaknesses in previous literature.

- Most of the research limits to a specific domain and lacks generality. Every method/algorithm in machine prognostics has its own application domain. A general methodology is needed for prognostics.
- It is a tough and high cost task to collect abundant data. Many algorithms and data training models require a large amount of historic data, including normal state data and failure (even fatal failure) data that need to destroy the components/systems artificially.
- Now for most of the "real-time" maintenance systems, only real-time input data collected from sensors are brought into trained models for diagnostics and prognostics. The prognostic model itself is usually not updated through online data. These static training data will decrease the precision of diagnostic and prognostic results. Methods that have fast computation speed are needed to achieve online updating of the prognostic model.
- A criterion of the judgment for prognosis accuracy and precision is lacking. Therefore, there is no consistent way to evaluate and compare the performances of proposed prognostic algorithms.
- Most researches are still resting at theoretical phase. Few of them have been applied to practical applications successfully.

The above weaknesses will lead to a long-learning time, long implementation time prognosis maintenance system with little practical value for real application. Generally, future researches on prognostics for implementing CBM may contribute to effectively measurement and management of the uncertainties in RUL and system degradation predictions. For example, when a probabilistic estimation of RUL is given, the error between predicted RUL and real RUL is expected to be within a given bound. At the same time, these new introduced prognostic methodologies need to be applied to more applications to demonstrate their effectiveness. Rather than domain specific models, holistic methodologies should be emphasized to adapt to the fast changing modern industry.

Another research direction is to join the prognostics with the maintenance scheduling. Elwany and Gebraeel built a sensor-driven prognostic model for an equipment replacement and spare parts inventory management system [87]. Accurate prediction of equipment failure time can help to develop more effective replacement and spare parts inventory policies. In summary, Fig. 4 proposes a systematic maintenance framework basing on the real-time diagnostics and prognostics.

References

1. Holmberg K, Komonen K, Oedewald P, Peltonen M, Reiman T, Rouhiainen V, Tervo J, Heino P (2004) Safety and reliability technology review. Res Rep BTUO43-031209. VTT Industrial Systems, Espoo
2. Tu PYL, Yam R, Tse P, Sun AOW (2001) An integrated maintenance management system for an advanced manufacturing company. Int J Adv Manuf Technol 17:692–703
3. Gertsbakh IB (1977) Models of preventive maintenance. North-Holland, Amsterdam
4. Tu Y (1995) Decision support system for equipment diagnosis and maintenance management: an artificial intelligent approach. Research Proposal, City University of Hong Kong, HK
5. Katipamula S, Brambley MR (2005) Methods for fault detection, diagnostics, and prognostics for building systems—a review, part I. HVAC&R Res 11(1):3–25
6. Jardine AKS, Daming L, Dragan B (2006) A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech Sys Signal Process 20:1483–1510
7. Schwabacher M, Goebel K (2005) A survey of artificial intelligence for prognostics. AAAI Fall Symposium—Tech Rep, pp 107–114
8. Zhang L, Li XS (2006) A review of fault prognostics in condition based maintenance. The Sixth International Symposium on Instrumentation and Control Technology: Signal Analysis, Measurement Theory, Photo-electronic Technology, and Artificial Intelligence, 13–15 Oct 2006, Beijing (China), vol 6357, art. no. 635752
9. Kothamasu R, Huang SH, VerDuin WH (2006) System health monitoring and prognostics—a review of current paradigms and practices. Int J Adv Manuf Technol 28:1012–1024
10. Goh KM, Tjahj B, Baines T, Subramaniam S (2006) A review of research in manufacturing prognostics. 2006 IEEE Int Conf on Ind, 16–18 Aug. 2006, Singapore (Singapore), vol 16–18, pp 417–422

11. Brotherton T, Jahns J, Jacobs J, Wroblewski D (2000) Prognosis of faults in gas turbine engines. Proceedings of the IEEE Aerospace Conf, 18–25 Mar. 2000, Big Sky, MT (USA), vol 6, pp 163–171
12. Li Y, Billington S, Zhang C, Kurfess T, Danyluk S, Liang S (1999) Adaptive prognostics for rolling element bearing condition. Mech Sys Signal Process 13:103–113
13. Engel SJ, Gilmartin BJ, Bongort K, Hess A (2000) Prognostics, the real issues involved with predicting life remaining. Proceedings of the IEEE Aerospace Conf, 18–25 Mar. 2000, Big Sky, MT (USA), vol 6, pp 457–469
14. Kacprzynski GJ, Gumina M, Roemer MJ, Caguiat DE (2001) A prognostic modeling approach for predicting recurring maintenance for shipboard propulsion systems. Proceedings of ASME Turbo Expo, 4–7 June 2001, New Orleans, LA (USA), P884310
15. Oppenheimer CH, Loparo KA (2002) Physically based diagnosis and prognosis of cracked rotor shafts. Proceedings of SPIE Compo and Sys Diagnostics, Prognostics, and Health Management II, vol 4733, pp 122–132
16. Luo J, Bixby A, Pattipati K, Liu Q, Kawamoto M, Chigusa S (2003) An interacting multiple model approach to model-based prognostics. Syst Man Cybern 1:189–194
17. Kacprzynski GJ, Sarlashkar A, Roemer MJ (2004) Predicting remaining life by fusing the physics of failure modeling with diagnostics. J Met 56:29–35
18. Byington CS, Watson M, Edwards D, Stoelting P (2004) A model-based approach to prognostics and health management for flight control actuators. Proceedings of the IEEE Aerospace Conf, 6–13 Mar 2004, Big Sky, MT (USA), vol 6, pp 3351–3362
19. Cempel C, Natke HG, Tabaszewski M (1997) A passive diagnostic experiment with ergodic properties. Mech Sys Signal Process 11:107–117
20. Qiu J, Zhang C, Seth BB, Liang SY (2002) Damage mechanics approach for bearing lifetime prognostics. Mech Sys Signal Process 16:817–829
21. Garga AK, McClintic KT, Campbell RL, Yang CC, Lebold MS, Hay TA, Byington CS (2001) Hybrid reasoning for prognostic learning in CBM systems. Proceedings of the IEEE Aerospace Conf, 10–17 Mar 2001, Big Sky, MT (USA), vol 6, pp 2957–2969
22. David JM, Krivine JP (1987) Three artificial intelligence issues in fault diagnosis: declarative programming, expert systems, and model-based reasoning. Proceedings of the Sec Euro Workshop on Fault Diagnostics, Reliability and Related Knowledge Based Approaches, vol 6–8, pp 190–196
23. Lembessis E, Antonopoulos G, King RE, Halatsis C, Torres J (1989) 'CASSANDRA': an on-line expert system for fault prognosis. Proceedings of the 5th CIM Euro Conf, 1989, pp 371–377
24. Butler KL (1996) An expert system based framework for an incipient failure detection and predictive maintenance system. Proceedings of the Int Conf on Intelligent Sys Applications to Power Sys, ISAP, 28 Jan–2 Feb 1996, Orlando, FL (USA), pp 321–326
25. Biagetti T, Sciubba E (2004) Automatic diagnostics and prognostics of energy conversion processes via knowledge based systems. Energy 29(12–15):2553–2572
26. Choi SS, Kang KS, Kim HG, Chang SH (1995) Development of an on-line fuzzy expert system for integrated alarm processing in nuclear power plants. IEEE Trans Nucl Sci 42(4):1406–1418
27. Frelicot C (1996) A fuzzy-based prognostic adaptive system. RAIRO-APII-JESA. J Eur Sys Autom 30(2–3):281–299
28. Luo J, Namburu M, Pattipati K, Liu Q, Kawamoto M, Chigusa S (2003) Model-based prognostic techniques [maintenance applications]. Proceedings of AUTOTESTCON, IEEE Sys Readiness Tech Conf, 22–25 Sep 2003, Anaheim, CA (USA), vol 22–25, pp 330–340
29. Ray A, Tangirala S (1996) Stochastic modeling of fatigue crack dynamics for on-line failure prognostics. IEEE Trans Control Sys Technol 4:443–451
30. Wang W (2002) A model to predict the residual life of rolling element bearings given monitored condition information to date. IMA J Manag Math 13:3–16
31. Yan J, Koc M, Lee J (2004) A prognostic algorithm for machine performance assessment and its application. Prod Plan Control 15:796–801
32. Banjevic D, Jardine AKS (2006) Calculation of reliability function and remaining useful life for a Markov failure time process. IMA J Manag Math 17(2):115–130
33. Goebel K, Saha B, Saxena A (2005) Prognostics in battery health management. IEEE Instrum Meas Mag 11(4):33–40
34. Gebraeel NZ, Lawley MA, Li R, Ryan JK (2005) Residual-life distributions from component degradation signals: a Bayesian approach. IIE Trans 37:543–557
35. Tsoukalas LH, Uhrig RE (1997) Fuzzy and neural approaches in engineering. Wiley, New York
36. Connor JT, Matinem RD, Atlas LE (1994) Recurrent neutral networks and robust time series prediction. IEEE Trans Neural Netw 5:240–254
37. Zhang S, Ganesan R (1997) Multivariable trend analysis using neural networks for intelligent diagnostics of rotating machinery. Trans ASME J Eng Gas Turbine Power 119:378–384
38. Yam RCM, Tse PW, Li L, Tu P (2001) Intelligent predictive decision support system for condition-based maintenance. Int J Adv Manuf Technol 17:383–391
39. Byington CS, Watson M, Edwards D (2004) Data-driven neural network methodology to remaining life predictions for aircraft actuator components. Proceedings of the IEEE Aerospace Conf, 6–13 Mar 2004, Big Sky, MT (USA), vol 6, pp 3581–3589
40. Khawaja T, Vachtsevanos G, Wu B (2005) Reasoning about uncertainty in prognosis: a confidence prediction neural network approach. Annual Conf of the North American Fuzzy Inf Proc Society—NAFIPS, 26–28 Jun 2005, Detroit, MI (USA), vol 2005, pp 7–12
41. Yu G, Qiu H, Djurdjanovic D, Jay L (2006) Feature signature prediction of a boring process using neural network modeling with confidence bounds. Int J Adv Manuf Technol 30:614–621
42. Wang P, Vachtsevanos G (2001) Fault prognosis using dynamic wavelet neural networks. Artif Intell Eng Des Anal Manuf (AIEDAM) 15(4):349–365
43. Gebraeel NZ, Lawley MA (2008) A neural network degradation model for computing and updating residual life distributions. IEEE Trans Autom Sci Eng 5(1):387–401
44. Parker BE, Nigro TM, Carley MP, Barron RL, Ward DG, Poor HV, Rock D (1993) Helicopter gearbox diagnostics and prognostics using vibration signature analysis. Proceedings of the SPIE—The Int Society for Optical Eng, 13 April 1993, Orlando, FL (USA), vol 1965, pp 531–542
45. Kohonen T (1995) Self-organizing maps. Springer, Berlin
46. Jamsa-Jounela SL, Vermasvuori M, Enden P, Haavisto S (2003) A process monitoring system based on the Kohonen self-organizing maps. Control Eng Pract 11:83–92
47. Huang RQ, Xi LF (2007) Residual life predictions for ball bearing based on neural networks. Chinese Journal of Mechanical Engineering 43(10):137–143
48. Liu Y, Li SQ (2007) Decision support for maintenance management using Bayesian networks. Int Conf on Wireless Communications, Networking and Mobile Computing, art. no. 4341175, 5708–5711
49. Sheppard JW, Kaufman MA (2005) Bayesian diagnosis and prognosis using instrument uncertainty. Proceedings of AUTOTESTCON, 26–29 Sep 2005, Orlando, FL (USA), vol 2005, pp 417–423

50. Przytula KW, Choi A (2007) Reasoning framework for diagnosis and prognosis. Proceedings of 2007 IEEE Aerospace Conf, 3–10 Mar 2007, Big Sky, MT (USA), art. no. 4161649
51. Dong M, Yang ZB (2008) Dynamic Bayesian network based prognosis in machining processes. J Shanghai Jiaotong Univ (Sci) 13(3):318–322
52. Bunks C, McCarthy D, Al-Ani T (2000) Condition-based maintenance of machines using hidden Markov models. Mech Sys Signal Process 14(4):597–612
53. Zhang XD, Xu R, Chiman K, Liang SY, Xie QL, Haynes L (2005) An integrated approach to bearing fault diagnostics and prognostics. Proceedings of the 2005 American Control Conf, 8–10 Jun 2005, Portland, OR (USA), vol 4, pp 2750–2755
54. Baruah P, Chinnam RB (2003) HMMs for diagnostics and prognostics in machining processes. Int J Prod Res 43(6):1275–1293
55. Camci F (2005) Process monitoring, diagnostics and prognostics using support vector machines and hidden Markov models. Graduate School of Wanye State University, Detroit
56. Gu HY, Tseng CY, Lee LS (1991) Isolated-utterance speech recognition using hidden Markov models with bounded state duration. IEEE Trans Signal Process 39(8):1743–1751
57. Falaschi A (1992) Continuously variable transition probability HMM for speech recognition. In: Laface P, De Mori R (eds) Speech recognition and understanding. Springer, Berlin, pp 125–130
58. Dong M, He D, Banerjee P, Keller J (2006) Equipment health diagnosis and prognosis using hidden semi-Markov models. Int J Adv Manuf Technol 30:738–749
59. Dong M, He D (2007) Hidden semi-Markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. Eur J Oper Res 178(3):858–878
60. Dong M, He D (2007) A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology. Mech Sys Signal Process 21(5):2248–2266
61. Kumar D, Klefsjo B (1994) Proportional hazards model: a review. Reliab Eng Sys Saf 44:177–188
62. Bebbington M, Lai CD, Zitikis R (2008) Reduction in mean residual life in the presence of a constant competing risk. Appl Stoch Models Bus Ind 24(1):51–63
63. Victor L, Riquelme M, Balakrishnan N, Sanhueza A (2008) Lifetime analysis based on the generalized Birnbaum–Saunders distribution. Comput Stat Data Anal 52:2079–2097
64. Cox DR (1972) Regression models and life-tables. J R Stat Soc 134:187–220
65. Lloyd GM, Hasselman T, Paez T (2005) A proportional hazards neural network for performing reliability estimates and risk prognostics for mobile systems subject to stochastic covariates. Eng/Tech Management, Safety Eng and Risk Analysis, Tech and Soc, Eng Bus Management, Health and Safety 2005, pp 97–106
66. Liao H, Qiu H, Lee J, Lin D, Banjevic D, Jardine A (2005) A predictive tool for remaining useful life estimation of rotating machinery components. Proceedings of the ASME Int Design Eng Tech Conf and Computers and Inf in Eng Conf-DETC, 24–28 Sep 2005, Long Beach, CA (USA), vol 1A, pp 509–515
67. Li ZG, Zhou S, Choubey S, Sievenpiper C (2007) Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures. IIE Trans 39:303–315
68. Deng JL (1989) Introduction to grey system theory. J Grey Syst 1 (1):1–24
69. Huang YP, Huang CC, Hung CH (1994) Determination of the preferred fuzzy variables and applications to the prediction control by the grey modeling. The Sec National Conf on Fuzzy Theory and Application, Taipei (Taiwan), pp 406–409
70. Ku LL, Huang TC (2006) Sequential monitoring of manufacturing processes: an application of grey forecasting models. Int J Adv Manuf Technol 27(5–6):543–546
71. Gu J, Bilal M, Pecht M (2008) Grey prediction method used in failure prognostics for electronics. http://www.prognostics.umd.edu/PHM_Website_new/PHM_grey_prediction.pdf
72. He QH, He XY, Zhu JX (2008) Fault detection of excavator's hydraulic system based on dynamic principal component analysis. J Cent South Univ Technol 15:700–705
73. Lee JM, Yoo CK, Choi SW, Vanrolleghem PA (2004) Nonlinear process monitoring using kernel principal component analysis. Chem Eng Sci 59:223–234
74. Nomikos P, MacGregor JF (2004) Monitoring batch processes using multi-way principal component analysis. AIChE J 40:1361–1375
75. Kwan C, Zhang X, Xu R, Haynes L (2003) A novel approach to fault diagnostics and prognostics. Proceedings of the 2003 IEEE Int Conf on Robotics and Automation, 14–19 Sep 2003, Taipei (Taiwan), vol 1, pp 604–609
76. Lee JM, Yoo CK, Lee IB (2003) On-line batch process monitoring using a consecutively updated multiway principal component analysis model. Comput Chem Eng 27:1903–1912
77. Wang WQ, Golnaraghi MF, Ismail F (2004) Prognosis of machine health condition using neuro-fuzzy systems. Mech Sys Signal Process 18:813–831
78. Chinnam RB, Baruah P (2004) A neuro-fuzzy approach for estimating mean residual life in condition-based maintenance systems. Int J Mater Prod Technol 20:166–179
79. Satish B, Sarma NDR (2005) A fuzzy BP approach for diagnosis and prognosis of bearing faults in induction motors. IEEE Power Eng Soc General Meeting, 12–16 Jun 2005, San Francisco, CA (USA), vol 3, pp 2291–2294
80. Xue GX, Xiao LC, Bie MH, Lu SW (2005) Fault prediction of boilers with fuzzy mathematics and RBF neural network. Proceedings of Int Conf on Communications, Circuits and Sys, 27–30 May 2005, Hong Kong (China), vol 2, pp 1012–1016
81. Kothamasu R, Huang SH (2007) Adaptive Mamdani fuzzy model for condition-based maintenance. Fuzzy Sets Syst 158:2715–2733
82. Chen MY (1990) Uncertainty analysis and grey modeling. Proceedings of the 1st International Symposium on Uncertainty Modeling Analysis, pp 469–473
83. Dong YL, Gu YJ, Yang K, Zhang WK (2004) A combining condition prediction model and its application in power plant. Proceedings of the Int Conf on Machine Learning and Cybernetics, 26–29 Aug 2004, Shanghai (China), vol 6, pp 3474–3478
84. Shetty P, Mylaraswamy D, Ekambaram T (2008) A hybrid prognostic model formulation and health estimation of auxiliary power units. J Eng Gas Turbine Power 130(2):021601
85. Mohanty S, Chattopadhyay A, Peralta P, Das S, Willhauck C (2008) Fatigue life prediction using multivariate Gaussian process. Collection of Tech Papers AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conf, 7–10 Apr 2008, Schaumburg, IL (USA), art. no. 2008-1837
86. Shao Q, Shao C, Feng CJ (2008) Identification of non-stationary time series based on SVM-HMM method. Proceedings of 2008 IEEE Int Conf on Service Operations and Logistics, and Informatics, 12–15 Oct 2008, Beijing, China, vol 1, pp 293–298
87. Elwany AH, Gebraeel NZ (2008) Sensor-driven prognostic models for equipment replacement and spare parts inventory. IIE Trans 40(7):629–639