Prognostics and Health Monitoring Methodologies and Approaches: A Review

Hasan A. Bjaili1* and Ali M. Rushdi1

1Department of Electrical and Computer Engineering, King Abdulaziz University, P.O. Box 80204
Jeddah 21589, Saudi Arabia.

Authors contributions

This work was carried out in collaboration between both authors. Both authors read and approved
the final manuscript.

Article Information

DOI: 10.9734/JERR/2020/v18i417215

Editor(s):
(1) Dr. Harekrushna Sutar, Indira Gandhi Institute of Technology, India.

Reviewer(s):
(1) Bashar Jabbar Hamza, Al-Furat Al-Awsat Technical University, Iraq.
(2) Fidias E. Leon-Sarmiento, University of Pennsylvania, Colombia.
Complete Peer review History: http://www.sdiarticle4.com/review-history/62205

Received 27 August 2020
Accepted 03 November 2020
Published 18 November 2020

Short Research Article

ABSTRACT

Prognostics is a term that engineering borrowed from medicine to refer to the discipline concerned
with the Remaining Useful Life (RUL) of an engineering device. This paper surveys the RUL
prediction techniques and classifies them into four categories of model-based techniques,
knowledge-based techniques, experience-based techniques, and data-driven techniques. A
comparative review is given for the main features, prominent advantages, potential shortcomings
and main subcategories for each of these categories. The survey is supported by an extensive list
for up-to-date references.

Keywords: Prognostics; State-space modeling; algorithms; Remaining Useful Life (RUL); lithiumion
battery; reliability.

2010 Mathematics Subject Classification: 53C25, 83C05, 57N16.

*Corresponding author: E-mail: hasan@bjaili.com;
1 INTRODUCTION

Prognostics is a term introduced recently into the reliability engineering discipline. This term (borrowed from that of medical prognostics) is concerned with the Remaining Useful Life (RUL) [1] (also known as the Mean Residual Life (MRL)) of engineering devices. It focuses on predicting the time in which the device will no longer perform its intended function. It is the field of predicting the future reliability and performance of an engineering device by assessing the deviation or degradation extent of the device from its normal operation condition expectation [2].

In recent years, the research on prognostics, reliability and asset life prediction has been enhanced in the field of Engineering Asset Management (EAM). Modern systems such as aircraft engines are often built with enormous complexities. These systems are often bundled with rich electronics and intricate connections as well as interactions among their subsystems and components. For example, a typical car consists of about 30,000 parts, 10 million lines of software code and 2,000 functional components [3]. An engine manufacturer such as Rolls-Royce® has developed a service agreement called TotalCare™ in which engines are not sold, but instead they are leased with a TotalCare™ service agreement. The cost of lease under the TotalCare™ service agreement is actually paid by the hour. The lease requires a support fee per hour of operation and the Original Equipment Manufacturer (OEM) or Engine Manufacturer provides the service support, including tracking of the inherent uncertainty. This service is also called Power-By-The-Hour. The Power-By-The-Hour policy eliminates the requirement for the lessee to pay hourly maintenance reserves for the use of the engines in addition to a fixed monthly lease fee. In the Power-By-The-Hour service, the lessee is protected from the unexpected cost of premature engine failure or capital cost to refurbish the supplier’s owned engines that require replacement due to life-limited parts (LLPs), Airworthiness Directive (AD) compliance and performance degradation. The lessee is responsible for the cost to refurbish or repair engines that require removal from the aircraft and a shop visit to resolve discrepancies.

1.1 Life-Limited Parts (LLPs)

Life-limited Parts (LLPs) are parts that have a limited service life. When this service life has been exhausted, the life-limited part must be removed from the aircraft and replaced before the aircraft is permitted to fly again. The service life of the life-limited parts may be expressed in hours of operation, cycles of operation, or calendar time.

1.2 Airworthiness Directives (AD)

Airworthiness Directives (AD) are legally enforceable regulations issued by the Saudi Arabian’s General Authority for Civil Aviation (GACA), United States Federal Aviation Administration (FAA), European Aviation Safety Agency (EASA) or any other country’s civil aviation authorities. These regulations are used to correct an unsafe condition in a product which could be an aircraft, engine, propeller, or appliance.

The Power-By-The-Hour service has led to a shape change in the product Life Cycle Cost (LCC) profile for the Original Equipment Manufacturer (OEM). In order for the engine manufacturer or lessor to account and budget for unexpected failures and LCC and still make it appealing for the lessee or operator, the OEM had to invest in research concerning the Mean Residual Life (MRL), which is frequently referred to as the Remaining Useful Life (RUL). Therefore, scientists studied and applied many life prediction approaches and techniques. Failure of an engineering device can be traced to an implicit deterioration and degradation mechanism that acts and evolves over time. Thus, the understanding and identification of the different potential failure mechanisms present in engineering devices is essential for accurate RUL and life prediction [4, 5, 6, 7]. The RUL and life prediction techniques can be classified into four categories (see Fig. 1):

- Model-Based Techniques,
- Knowledge-Based Techniques,
- Experience-Based Techniques,
- Data-driven Techniques.

Each of the following four sections is devoted to an overview of one the aforementioned categories.
2 MODEL-BASED TECHNIQUES

The term "Model-based techniques" typically refers to approaches using models derived from first principles (e.g., physics-based). It simply depicts methods that use mathematical models of system behavior. In essence, the model can be derived from either one of the following two ways:

- knowledge (when available) from first principles, known physical laws, expert experience, dimensional analysis, method of similitude, prototyping, and the like.
- when a large amount of data is available for both nominal and degraded behavior.

Models are typically developed from a mixture of system knowledge and system data. Systems can be described by discrete state-space equations of the form:

\[
\begin{align*}
x(k+1) &= f(k, x(k), \phi(k), u(k), v(k)), \\
y(k) &= h(k, x(k), \phi(k), u(k), n(k)),
\end{align*}
\]

(2.1)

where \(x\): states, \(\phi\): parameters, \(u\): input, \(y\): output, \(v\): process noise, \(n\): sensor noise.

A model-based technique is used when an accurate and explicit mathematical model of the degradation process can be developed and constructed from first principles [8, 5]. It utilizes the outcome of a consistency check between the sensed measurement of the actual system and the outcomes of a mathematical model. The difference is called the residual. If the residual is large, then there is a malfunction, and when the residual is small, then there is a normal disturbance, noise or modeling error [9]. The model-based approach is also called Analytical-Based or Accelerated Degradation Modeling (ADM).

The pros of the model-based technique include the features that it:

- relies on the understanding of the physics of the system,
- experiences changes in the feature vector due to and commensurate with a change in the model parameters [10],
- requires less data than the data-driven approach,
- demands lower cost for implementation.

On the other hand, the cons of the model-based technique are manifested in the fact that it

- requires more knowledge on the fundamental theory relevant to the monitored system,
- demands many assumptions about the practical operating conditions of the system,
- suffers from the dependence of its robustness and accuracy on the experimental condition under which the models were developed [11],
- requires the estimation of various physical parameters of the system,
- might not produce desirable and practical results since the fault type differs from one component to another.

The model-based technique can be further subcategorized as [12]:

- Physics-of-Failure (PoF) Models,
- Statistical Models,
- Kalman/Particle Filtering, or
- Non-linear Dynamics.

2.1 Physics-of-Failure (PoF) Models

The physics-of-failure (POF) provides a focus for the life and reliability aspects of components. It addresses the root causes of failure such as fatigue, fracture, wear, and corrosion [13]. The technique of Physics-of-Failure (PoF) prognostics depends on the product life cycle loading and failure mechanism to get the Remaining Useful Life (RUL) and perform prognostics evaluation [14]. This technique allows assessing the product reliability under its actual usage conditions. It utilizes the product model as well as \textit{in situ} measuring sensor data integrated together to assess the product deviation and degradation form its expected normal operation or operating conditions and predict of the future reliability state. Fig. 2. outlines the Physics-of-failure (PoF) based prognostics and health monitoring (PHM) methodology. First, the design data, life cycle, expected life cycle conditions, Failure Modes, Mechanisms and Effects Analysis (FMMEA) [14, 15, 2, 16, 17, 18], and Physics-of-failure models.
are combined to be the inputs and are called the virtual life assessment (Virtual Reliability). The critical failure modes and mechanisms are prioritized based on the virtual life assessment. The built-in-test results, existing sensory data, inspection and maintenance records, and warranty data are also utilized to identify the abnormal conditions. The monitoring parameters and sensor location can be determined using the aforementioned information. It is impossible to perform data-driven prognostics for a product that has not been manufactured since there is no data available for training the algorithm. For new-product prognostics, we only need to change the material geometries or properties to model the product. This is because most new products are not completely different from previous products. Similar products can be referenced via Failure Modes, Mechanisms and Effects Analysis (FMMEA) [14]. FMMEA is a systematic approach to identify failure mechanisms and models for all potential failure modes and then prioritize them. It is based on understanding the relationships between [19]:

- Requirements and the physical characteristics of the product.
- The interactions of product materials with loads.
- The influence on product failure susceptibility with respect to the use conditions.

The methodology of PoF-based prognostics comes in handy when prognostics analysis is needed for a new product or for the prognosis of a legacy system. In the case of legacy systems, it is difficult to obtain training data and also very hard to assess the Remaining Useful Life (RUL) if the failure mechanisms and their effect on the collected parameters are not well understood. The PoF-based prognostics approach depends on the understanding of the legacy system structure and life cycle conditions as well as the failure modes and mechanisms. According to [20], the PoF technique can be summarized as:

- Establish a list of the probable fault structure for example chemical, electrical, physical, mechanical, structural, or thermal processes leading to failure.

---

**Fig. 1. Taxonomy of prognostics approaches and techniques**
2.2 Statistical Models

Statistical modeling (which is also called the probabilistic-based modeling or usage-based prognostics modeling) is typically developed based on known information about the degradation of the measure parameters [21]. The parameters measurements are collected over time to assess the current severity of the parameters distribution shifts. Changes in these parameters can be accelerated using the accelerated degradation techniques.

![Fig. 2. PoF-based PHM methodology. Graph reproduced from [14]](image)

**Accelerated Life Test (Accelerated Degradation)** The accelerated life tests consist of different test techniques to shorten the life of products or to hasten the degradation of their performance. The purpose of such testing is to obtain degradation data quickly. Such data is then modeled and analyzed. We do this because in real life, devices of practical utility take long time to deteriorate or degrade. The real life degradation process is usually very slow at normal wear-and-tear conditions, i.e., the Mean Time To Failure (MTTF) is comparatively high. Therefore, to obtain statistical data in a fast manner from a degradation test, we employ an accelerated life test by changing and speeding up the environment variables such as temperature, vibration amplitude, load, voltage, corrosive media and pressure [22, 23]. There are many types of acceleration testing such as:

- **High Usage Rate:** Also called compressed time testing, is an easy method to hasten the deterioration and shorten the life of many products by running them at a higher usage rate. We can achieve such testing by either running the component faster or by reducing its time off period. An example of running a component faster is when a rolling bearing is run at three or four times its intended or normal speed. **Reduced time off** accelerated testing concerns home appliances such as washers and dryers when they are run 24-hour a day while in normal life they are run one or two hours a day. Another example of reduced time off accelerated testing is when a toaster or a coffee maker are cycled hundreds of times a day while in normal life they are cycled three to five times a day.

- **Overstress Testing:** In overstress accelerated testing, the product runs in an overstressed environment to shorten
its life. This is done by changing the component overstressed environment variables such as temperature, thermal cycling, humidity, vibration, voltage and/or mechanical load.

- **Stress Loading** Stress loading is a type of accelerated testing methods that is done by applying a stress on the component until it fails while degradation statistical data is being gathered. Stress loading can be carried out using constant stress, step stress, or linearly increasing stress. The constant stress refers to the case in which constant stress or force is applied while the component is running. The step stress can be done by applying stress to the component while changing the stress setting at specified times. The stress is applied at low level and at a specified time. If the component does not fail, the stress is raised and held for a determined time and so on and so forth until the component is failed. Simple step stress uses only two stress levels. The method of linearly increasing stress means simply applying stress and increasing it linearly until the component fails.

There are many statistical (probabilistic-based) models. However, this paper will only review three of them, which are:

### 2.2.1 Proportional Hazards Model (Cox Model)

The proportional hazards model (also called Cox model) is a type of a regression method [24]. This model is widely used in biomedical applications to investigate the effects of several covariates on a survival distribution at the same time. Cox models are a statistical technique for exploring the relationship between the survival of the subject (component, part, human being, etc.) in test and several explanatory variables. The survival analysis is mainly concerned with studying the time between entry to the study and a subsequent event such as inoperability (cease of operation), breakage, damage, etc. [27].

Cox proportional hazard model or Cox model cater for estimating the hazard (or risk) of parts/equipment damage, given its prognostics variables. The model is briefly explained below:

Let \( x_1, \ldots, x_n \) denote the variables, and let \( h_0(t) \) denote the unknown life distribution hazard function at \( x_1 = x_2 = \cdots = x_n = 0 \). Thus, the hazard function for the distribution at the variables \( x_1, \ldots, x_n \) is:

\[
h(t; x_1, \ldots, x_n) = h_0(t) \cdot e^{(\gamma_1 x_1 + \cdots + \gamma_n x_n)}.
\] (2.2)

The base hazard function \( h_0(t) \) and the coefficients \( \gamma_1, \gamma_2, \ldots, \gamma_n \) are estimated from data. The corresponding reliability function is:

\[
R(t; x_1, \ldots, x_n) = e^{-\int_0^t h_0(\tau; x_1, x_2, \ldots, x_n) d\tau} = [R_0(t)]^{e^{(\gamma_1 x_1 + \cdots + \gamma_n x_n)}},
\] (2.3)

where \( R_0(t) \) is the reliability function at \( x_1 = x_2 = \cdots = x_n = 0 \) and in given by

\[
R_0(t, 0, 0, \ldots, 0) = e^{-\int_0^t h_0(\tau) d\tau}.
\] (2.4)

The Proportional Hazards Model (Cox Model) is presented in detail in [25, 28, 29].

### 2.2.2 Logistic Regression Model

The logistic regression model is also widely used in biomedical applications in which the predictor or dependent variable is a binary 0 or 1, on or off, operational or inoperational, dead or alive, approved or disapproved, etc. This model could also be called the qualitative response/discrete choice model. The logistic regression model is a probability distribution of an event occurring depending on the values of the independent variables, which can be categorical or numerical. The model estimates the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur. It seeks the effect of a series of variables on binary response variables and classifies observations by estimating the probability that an event observation is a particular category, as mentioned above, on or off, etc.

The logistic proportional model for the proportion \( p \) in a particular category, for example, inoperational as a function of \( n \) independent variables \( x_1, \ldots, x_n \) is

\[
\ln \left( \frac{1-p}{p} \right) = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_n x_n.
\] (2.5)
where $\gamma_0, \gamma_1, \ldots, \gamma_n$ are unknown coefficients to be estimated from the data.

### 2.2.3 Cumulative Damage Model

The **Cumulative Damage Model** is defined to be the permanent (irreversible) damage accumulation in a component under a cyclical usage pattern [30]. It is developed based on the concepts of bounds on residual fatigue life in two-stage cycling [31]. Basically, it has been used in three different areas [32]:

- **Mechanical Systems**: to predict time of failure
- **Health and Safety**: to determine the humans tolerance level when a person is exposed to toxic or latent injurious materials.
- **Structural Engineering**: to calculate the structure safety of structural parts when they are exposed to loads over time.

The cumulative damage can be modeled using the non-linear Miner’s rule [6]:

$$ D = \left( \frac{n}{c} \right)^{r} \quad (2.6) $$

Where $D$ is the damage, $n$ is the number of cycles experienced, $r$ is the non-linear damage exponent, and $c$ is the number of cycles to crack initiation.

### 2.3 Kalman/Particle Filtering

Filtering is usually used in many situations in engineering. For example, radio communication signals are usually corrupted with noise which makes the signal unusable. Therefore, a good filtering algorithm may come in handy to remove noise [33]. In this subsection, we will discuss only two types of filtering that could be used in prognostics to estimate the model parameters: **Kalman Filters and Particle Filters**. Filtering using Kalman/Particle filters has been used in [34, 35] to predict the state of a battery charge and the Remaining Useful Life (RUL).

#### 2.3.1 Kalman Filtering

The Kalman filter is a statistical estimator to the linear-quadratic problem, which is technically the problem of estimating the instantaneous state of a linear dynamic system affected by a white noise [36]. Alternatively, Kalman filtering can be viewed as a recursive algorithm which can estimate the true (instantaneous) state of a noisy system [37]. Kalman filtering can only be applied to linear systems. Non-linear system can be filtered using the **Extended Kalman Filter** which is an improved version of the original Kalman filter. A linear system is simply a process that can be described by the following state (difference) and output equations:

$$ \begin{align*}
\dot{x}_{k+1} &= \tilde{A} \tilde{x}_k + \tilde{B} \tilde{u}_k + \tilde{w}_k \\
\tilde{y}_k &= \tilde{C} \tilde{x}_k + \tilde{z}_k
\end{align*} \quad (2.7) $$

where $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ are matrices, $\tilde{u}$ is a known control input, $\tilde{y}$ is the measured output, $\tilde{z}$ is the measured noise, $\tilde{w}$ is the process noise, and $k$ is the discrete time index.

The $\tilde{x}$, $\tilde{u}$, $\tilde{w}$, $\tilde{y}$ and $\tilde{z}$ quantities are vectors, in general; therefore, they each of them might than one element. Kalman filters can expressed in many forms and formulations, one of them is [33]:

$$ K_k = A P_k C^T \left( C P_k C^T + S_k \right)^{-1} \quad (2.8) $$

$$ \dot{x}_{k+1} = (A \tilde{x}_k + B \tilde{u}_k) + K_k (y_{k+1} - C \tilde{x}_k) \quad (2.9) $$

$$ P_{k+1} = A P_k A^T + S_w - A P_k C^T S_k^{-1} C P_k A^T \quad (2.10) $$

#### 2.3.2 Particle Filtering

Particle filters are sequential Monte Carlo statistical simulation techniques which can provide a very accurate approximation of the sequential Bayesian estimator [38]. The basic concept of the particle filters as in [39] is described next. We use a set of weighted particles as follows:

$$ \left\{ w_{i}^{(k)}, x_{i}^{(k)} \right\}_{i=1}^{N} \quad (2.11) $$

where $x_{i}^{(k-1)}$ is the state of the particle $i$ while the weight of this particle is $w_{i}^{(k-1)}$. These are used to approximate the posterior density at time $k-1$. The weights are normalized, i.e., $\sum_{i=1}^{N} w_{i}^{(k-1)} = 1$.

This approximation, then, can be expressed as:

$$ p \left( x_{k-1} | x_{1:k-1} \right) \approx \sum_{i=1}^{N} w_{i}^{(k-1)} \delta \left( x_{k-1} - x_{i}^{(k-1)} \right) \quad (2.12) $$
Where $\delta_a(x)$ is a Dirac delta that is concentrated at $a$ and $x_{t : k - 1}$ is available measurements up to time $k - 1$. Prognostics using Particle filtering and particle filters are described in detail in [40, 41, 34, 42, 43].

2.4 Nonlinear Dynamics

Linear dynamical systems can be described as objects of any type or nature. The state of a dynamical system evolves over time in accordance with dynamical rules or as a result of an evolving deterministic operator [44]. Every dynamical system has a mathematical model associated with it. A dynamical system is considered as a black box whose internal parts are unknown as shown in Fig. 3.

In this simple method, it is assumed that $y(t)$, a time series quantity, is obtained by observing the underlaying autonomous dynamical system and is mixed with unmonitored and unwanted force or noise [45]. Inherently, there is another time series quantity which we also observe, that is the input $u(t)$.

There are many approaches to develop a model for the dynamical system. The deterministic approach, one of them, assumes that the input-output time series relationships can be modeled as:

$$\frac{dx}{dt} = f(x(t), u(t)),$$  \hspace{1cm} (2.13)

$$y(t) = h(x(t)).$$  \hspace{1cm} (2.14)

Many researchers used the nonlinear dynamics techniques to model components for prognostics analysis. Gašperin et al. [46] modeled a gear to predict its Remaining Useful Life (RUL). The propagation of the damage of the gear is inherently a stochastic process; therefore, its distribution of the Remaining Useful Life (RUL) must be estimated by the propagation of the distribution of the current system state. The gear damage was modeled by a nonlinear dynamical system with two hidden states and a single measured output. The gear condition was described with a dynamical process, which is affected by random tribological inputs. These inputs occur due to the impact of the rotating parts. The aforementioned condition can then be described by the following state-space model which is a random process:

$$x_{t+1} = f(x_t, w_t, \Theta)$$  \hspace{1cm} (2.15)

$$y_t = g(x_t, e_t, \Theta)$$  \hspace{1cm} (2.16)

Prognostics using nonlinear dynamical systems are described in detail in [47, 48].
3 KNOWLEDGE-BASED TECHNIQUES

When it is difficult to perform prognostics analysis on engineering devices using model-based techniques, the knowledge-based techniques may come in handy. Knowledge-based prognostic techniques predict the Remaining Useful Life (RUL) by assessing the similarity between a data bank of predefined failures and an observed current situation [11]. The predefined failure data are usually collected from subject-matter experts as well as via an interpretation of a set of rules [49]. Fuzzy logic and expert systems are the most common techniques used to predict the Remaining Useful Life (RUL).

3.1 Expert Systems

An expert system is mainly a computer program or a set of computer programs that emulates the performance and decision-making ability of a knowledgeable human being. A knowledge-based expert system consists of either one of the following:

- Rule-based expert system,
- Model-based expert system,
- Case-based expert system.

3.1.1 Rule-based Expert System

In a rule-based expert system, the knowledge is represented in an IF Condition(s) THEN action(s) such as, "If Exhaust Gas Temperature (EGT) is more than 650° THEN the inlet fans are malfunctioned." These IF-THEN statements form a knowledge database comprised of heuristic factual data gathered by many experts over many years as shown in Fig. 4. In order for any rule-based expert system to be useful, it ought to be complete as well as exact as much as possible [50]. For a complete rule-based expert system, all rules and information available should have been one-to-one mapped. To be exact, the aforementioned mapping should be logically consistent such that no chain of logical rules pulled from the knowledge base contradicts with the human operator A rule-based system is often called a production system [51].

3.1.2 Model-based Expert System

A model-based expert system is used on a good model that mimics the structure and behavior of an engineering device. Generally, a model-based expert system, not to be confused with Model Based Techniques previously discussed in section 2, requires deeper reasoning to predict the Remaining Useful Life (RUL). In a model-based expert system, there are three fundamental tasks for prognostics:

1. Hypothesis Generation,
2. Hypothesis Testing,
3. Hypothesis Discrimination.

It is assumed that there is only a single point of failure [52]. The explanation of the aforementioned tasks is described in detail in [52]. In this process, the source of the problem prediction and the use of the observation get separated to improve the confidence of each problem source. The modeled system's behavior is then compared with the observed behavior of the same system and if they differ from one other, then there is a discrepancy in the system provided the model is correct and accurate.

3.1.3 Case-based Expert System

In our normal life, we tend to look for the services of experienced/older doctors and professionals. This is because we think that the experienced doctors or professionals have been around for a while and treated/fixed similar issues in the past; therefore, they quickly can figure out the remedy. A case-based expert system is a technique that depends on past experience (cases) to find a solution for the existing/current problem. Similarly, to the experienced doctors or professionals, a case-based method takes information on a current behavior of a problem and searches for the most similar past issue/problem (case). One of the advantages and strengths of the case-based expert systems is that they become better and better overtime when more cases are added and analyzed. The knowledge-based expert systems are all explained in detail in the Artificial Intelligence (AI) Literature in [51, 53, 54, 55, 56, 57, 52].
3.2 Fuzzy Logic

Fuzzy logic was presented in 1965 by Zadeh [58]. Basically, it was an extension to the classical Boolean logic in order to relax the tough constraint of the Excluded Middle (of being either absolutely true or absolutely false). Zadeh suggested that why not have something like 0.75 true [59]? The aforementioned 0.75 true statement can be interpreted as “not really true”. Fuzzy logic theory alleviates the harsh constraint of the classical logic set theory of true and false by allowing partial set association based on a variable degree of truth. It provides a mechanism for the linguistic phrases such as low, medium, high, often, few, etc. in a nonlinear mapping of the vector of the input data into a scalar single output [60].

A fuzzy prognostics system comprises of: a knowledge base, a fuzzy rule database and a program (algorithm) as shown in Fig. 5. The algorithm fuzzifies the crisp input, utilizing both databases to make inferences based on the fuzzified input, produces a fuzzy output, and finally defuzzifies it into a crisp output.

![Diagram of Rule-based architecture](image.png)
Similarly to all knowledge-based systems, the fuzzy logic system uses the IF-THEN rules to solve a problem, but unlike the other knowledge-based systems, fuzzy logic is intentionally made imprecise \[11\]. Fuzzy logic systems are regularly used to enhance other prognostics models \[61, 62, 63, 64\]. Majidian and Saidi \[65\] predicted the Remaining Useful Life (RUL) of a set of boiler tubes using fuzzy logic algorithms.

With its robust mathematical framework, fuzzy logic can deal with real-life imprecision and non-statistical uncertainty. Fuzzy logic and its applications are discussed in abundance in \[66, 59, 67, 68, 58, 65, 61, 62, 63, 64\].

4 EXPERIENCE-BASED TECHNIQUES

Probably the simplest technique of prognostics is based on the well-known probability distributions. Those are distribution functions based on identical events records of failures logged over a period of time. The experience-based technique is extensively used in reliability analysis. When studying probability, it is assumed that some probabilities and/or related quantities can be calculated based on this knowledge assumption. Conversely, in mathematical statistics, data is observed to compute related quantities and/or other probabilities or to predict the RUL \[70\].

The method of experience-based models depends mainly on mathematical statistics which is classified as parametric or non-parametric. This classification depends on our knowledge of the data distribution that belongs to a given family in which parameters such as the mean \((\mu)\) and the standard deviation \((\sigma)\) can easily be computed.

4.1 Parametric Distributions

A parametric distribution can concisely be described with just a few parameters without having to report the entire curve. A parametric model can be used to extrapolate—in time—to either the upper tail of a distribution or the lower part. It provides smooth estimates of the failure-time distribution which are perfect for prognostics analysis.

The following parametric distributions can be used for prognostics to predict the RUL:

- Location-scale and Log location-scale,
- Exponential,
- Normal and Lognormal,
- Smallest and Largest Extreme Value,
- Weibull,
- Logistic and Log-Logistic.
4.1.1 Location-Scale & Log Location-Scale Distribution

The Location-Scale is one of the parametrized probability distribution families. It is parametrized by a location and a scale. For any random variable X where x belongs to the location-scale family, its CDF is a function of \((x - a)/b\). The Location-Scale CDF is:

\[
F_x(x|a, b) = F \left( \frac{x - a}{b} \right) - \infty < a < \infty, b > 0
\]  

(4.1)

If the distribution of X is absolutely continuous the \((a, b)\) is the location-scale parameter with a pdf:

\[
f_x(x|a, b) = \frac{1}{b} f \left( \frac{x - a}{b} \right)
\]

(4.2)

Transforming the location-scale distribution using \(X = \log(Y)\) in which \(Y = e^x\) produces a log-location-scale distribution. Based on this transformation, the location parameter \(\mu\) becomes \(\theta = e^\mu\) and the scale parameter \(\sigma\) becomes \(\lambda = 1/\sigma\) where \(\theta\) is called the scale parameter and \(\lambda\) is called the power parameter [71].

Mukhopadhyay and Roy [72] used the log-location-scale distribution to perform Bayesian accelerated life testing as well as reliability analysis. The model used is:

\[
f_{Y_j} = (y|\theta_j, \tau_j, z) = \tau_j f_j(\tau_j(y - \mu_j(\theta_j, z)))
\]

(4.3)

4.1.2 Exponential Distribution

The exponential distribution which is also called the negative exponential distribution is a probability distribution that describes a time between two events occurring. The pdf of the exponential distribution is:

\[
f(x) = \begin{cases} 
\lambda e^{-\lambda x}, & \text{if } x \geq 0 \\
0, & \text{if } x < 0
\end{cases}
\]

(4.4)

The exponential distribution as a one-parameter distribution can be used in prognostics by taking the hazard function to be constant, \(\lambda(t) = \lambda > 0\), all over the range of \(t\) [26]. The failure frequency is:

\[f(t) = \lambda e^{-\lambda t}\]

(4.5)

Hence, the CDF is

\[F(t) = 1 - e^{-\lambda t}\]

(4.6)

Therefore, the reliability function will be

\[R(t) = e^{-\lambda t}\]

(4.7)

The reliability function which is also called the survival function [73] has a mean \((\mu)\) and variance \((\sigma^2)\) as follows:

\[
\mu = \frac{1}{\lambda}
\]

(4.8)

\[
\sigma^2 = \frac{1}{\lambda^2}
\]

(4.9)

Gebraeel et al. [74] modeled the exponential degradation function as:

\[
S(t_i) = \phi + \beta e^{(\mu_i)} e^{(\epsilon(t_i) - \sigma_i^2)}
\]

(4.10)

where \(S(t_i)\) denotes the degradation signal as a continuous stochastic process, continuous with respect to time \(t\), \(t_i \geq 0\), \(i = 1, 2, \ldots, \phi\) is a constant deterministic parameter, \(\theta\) is a lognormal random variable and \(ln(\theta)\) has a mean \(\mu_0\) and variance \(\sigma_0^2\), \(\beta\) is a normal random variable with a mean \(\mu_1\) and variance \(\sigma_1^2\), \(\epsilon(t_i)\) is a random error part following a normal distribution with a zero mean and variance \(\sigma^2\).

The hazard rate will then be:

\[h(t) = \lambda\]

(4.11)

which is constant and independent of time. The only distribution that has a constant failure rate is the exponential distribution [75].

Based on that, the Mean Time Between Failures (MTBF) is \(1/\lambda\) and the failure rate is \(\lambda\) while \(1/\lambda\) is the 63rd percentile meaning that 63% of the population will have failed by then.

Failure experts claim that electronic components fail randomly, thus, exhibiting exponential distribution failure. The exponential distribution is no longer used heavily in reliability applications [76].

4.1.3 Normal and Log-Normal Distribution

Without any doubts, the most famous and most widely used probability distribution model is the Normal Distribution which is also known as the
Gaussian Distribution. A random variable $X$ with pdf:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ (4.12)$$

is a normal random variable with the mean $\mu$ ($-\infty < \mu < \infty$) and standard deviation $\sigma > 0$.

The normal distribution hazard function increases without limits. Therefore, it can be used to model products with wear-out failure such as the life of incandescent lamp filaments as well as the electrical insulation. Nelson [27] used the normal distribution Cumulative Distribution Function (CDF) to model the population fraction failing by age $y$.

$$F(y) = \Phi\left\{ \frac{y - \mu}{\sigma} \right\}, \ -\infty < y < \infty \ (4.13)$$

The mean $\mu$ and standard deviation $\sigma$ should be in the same measurement of $y$ whatever it is, hours, months or cycles.

Evaluating 4.13 at $\mu = 0$ and $\sigma = 1$ yields:

$$F(y) = \Phi\left( \frac{y}{\sigma} \right), \ -\infty < y < \infty \ (4.14)$$

Which is the standard normal distribution function. Additionally, Daniel Inman et al. [77] used the normal distribution to model the statistical strength and toughness of composite materials under dispersed type load sharing as:

$$G_n(\sigma) = \Phi\left( \frac{\sigma - \mu_n}{\gamma_n} \right), \ \sigma \geq 0 \ (4.15)$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \ (4.16)$$

with mean $\mu$ and standard deviation $\gamma_n$.

The Lognormal distribution is widely used to predict metal fatigue, solid state components and electrical insulation life. Engineering devices that follow the lognormal distribution can be modeled as:

$$F(t) = \Phi\left[ \frac{(\log(t) - \mu)}{\sigma} \right], \ t > 0 \ (4.17)$$

which is the population fraction by age $t$. Note that $\log(.)$ here is base $e$ usually written as $(\ln(.))$ and should not be confused with base 10. Nelson [27] showed the lognormal reliability function to be:

$$R(t) = 1 - \Phi\left[ \frac{(\log(t) - \mu)}{\sigma} \right] \ (4.18)$$

### 4.1.4 Smallest and Largest Extreme Value Distribution

The smallest and largest extreme value distribution is characterized by scale and location parameters. The smallest extreme value distribution is skewed to the left and is used to model the minimum value from a distribution of random observations. It is also used to model time to failure for a system that fails when its weakest component fails. Whereas, the largest extreme value distribution is skewed to the right and is used to model the maximum value from a distribution of random observations. It is also used to describe extreme phenomena such as extreme velocities.

The smallest extreme value distribution is used, also, to predict the life of a series system in which the systems fails if any component fails. Whereas, the largest extreme value is used to predict the life of a parallel system in which the system fails when all its components fails.

The probability density function (pdf) of the smallest extreme (the minimum) is:

$$f(x) = \frac{1}{b} e^{\left(\frac{x-a}{b}\right)} e^{-\left(\frac{x-a}{b}\right)}, \ -\infty < x < \infty \ (4.19)$$

where $a$ is the location and $b$ is the scale parameter.

The pdf of the largest extreme (the maximum) is [78, 75]:

$$f(x) = \frac{1}{b} e^{\left(-\frac{x-a}{b}\right)} e^{-\left(-\frac{x-a}{b}\right)}, \ -\infty \leq x \leq \infty \ (4.20)$$

where again $a$ is the location and $b$ is the scale parameter. Note that the two exponentials in equation 4.20 have extra negative sign when compared with the corresponding ones in equation 4.19.

### 4.1.5 Weibull Distribution

The Weibull distribution is widely used in reliability analysis. It is frequently used to model the time to failure. The pdf of the Weibull distribution is:

$$f(x) = \frac{\beta}{\gamma} \left(\frac{x}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x}{\gamma}\right)^{\beta}}, \ for \ x > 0 \ (4.21)$$
where $\delta$ is the scale parameter and is $> 0$ and $\beta$ is the shape parameter and is also $> 0$.

The pdf of a Weibull distribution as used in prognostics analysis is:
\[
f(t) = \frac{\beta}{\eta} \left( \frac{t - \gamma}{\eta} \right)^{\beta - 1} e^{-\left( \frac{t - \gamma}{\eta} \right)^{\beta}} \quad (4.22)
\]
where $\eta$ is the characteristic life, $\beta$ is the shape factor and the $\gamma$ is the location parameter. Thus the Hazard rate of a Weibull distribution is:
\[
h(t) = \frac{\beta}{\eta} \left( \frac{t - \gamma}{\eta} \right)^{\beta - 1}, \quad t \geq \gamma; \eta > 0 \quad (4.23)
\]

The Cumulative Distribution Function (CDF) of a Weibull distribution is:
\[
F(t) = 1 - e^{-\left( \frac{t - \gamma}{\eta} \right)^{\beta}}, \quad t \geq \gamma; \beta, \eta > 0 \quad (4.24)
\]
The reliability function can be expressed as follows:
\[
R(t) = 1 - F(t) = e^{\left( \frac{t - \gamma}{\eta} \right)^{\beta}}, \quad t \geq \gamma; \beta, \eta > 0 \quad (4.25)
\]

4.1.6 Logistic and Log-logistic Distribution

The logistics distribution is usually used for growth models. Its Probability Density Function (pdf) is:
\[
F(t) = \frac{\exp\left( \frac{t - \mu}{\sigma} \right)}{1 + \exp\left( \frac{t - \mu}{\sigma} \right)}, \quad -\infty < t < \infty, -\infty < \mu < \infty, \sigma > 0 \quad (4.26)
\]
where $\mu$ is the location parameter and $\sigma$ is the scale parameter. The logistic failure rate and the reliability functions are:
\[
\lambda(t) = \frac{\exp\left( \frac{t - \mu}{\sigma} \right)}{\sigma \left(1 + \exp\left( \frac{t - \mu}{\sigma} \right) \right)} \quad (4.27)
\]
\[
R(t) = \frac{1}{1 + \exp\left( \frac{t - \mu}{\sigma} \right)} \quad (4.28)
\]
The Probability Density Function (pdf) of the log-logistic distribution is:
\[
f(t) = \frac{\left( \frac{\beta}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1}}{\left[1 + \left( \frac{t}{\alpha} \right)^{\beta} \right]^2}, \quad \alpha > 0, \beta > 0 \quad (4.29)
\]
The survivor and hazard functions are shown below:
\[
S(t) = \frac{1}{1 + \left( \frac{t}{\alpha} \right)^{\beta}} \quad (4.30)
\]
\[
h(t) = \frac{\left( \frac{\beta}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1}}{1 + \left( \frac{t}{\alpha} \right)^{\beta}} \quad (4.31)
\]
The log-logistics distribution got its name due to the fact that it handles a random variable $Y$, given by:
\[
Y = \log T \quad (4.32)
\]
where $T$ is the logistic random variable. The pdf of $Y$ is given by:
\[
f_Y(y) = \frac{(b-1) \left( \frac{y-\mu}{b} \right)^{\beta-1}}{(1 + \left( \frac{y-\mu}{b} \right)^{\beta})^2} \quad (4.33)
\]
where $\mu = \log \alpha$, $b = \beta^{-1}$, $-\infty < \mu < \infty$, $b > 0$.

4.2 Non-parametric Distributions

When it is difficult to fit the data to a known or specific probability distribution family, a non-parametric distribution is used. A method using this type of distribution is also called a distribution-free method or technique [75]. Additionally, no assumption about the distribution of the underlying population is made other than that it is continuous [79].

5 DATA-DRIVEN APPROACHES

Data driven approaches depend on large sets of time series data. They use statistical or artificial intelligence techniques on relatively large sets of measured component degradation or performance data to predict component reliability and states. Many of the existing approaches of the data-driven prognostics analysis use neural networks to model the system and hence its RUL. Data driven methods can be grouped into four main categories: multivariate statistical, black-box, signal analysis and graphical methods.

5.1 Multivariate Statistical Methods

Multivariate statistical analysis focuses on analyzing several related variables
simultaneously. Each of those variables is equally important. The most important statistical methods are:

5.1.1 Principal Component Analysis

The technique of principal component analysis is one of the simplest of the multivariate methods. The main objective of this analysis is to take a bunch of variables such as \( x_1, x_2, x_3, \ldots, x_p \), and then find a combination of these to generate indices \( Z_1, Z_2, Z_3, \ldots, Z_p \) which are uncorrelated [55]. This lack of correlation indicates that they measure different dimensions of the data and in the order such that \( \text{Var}(Z_1) \geq \text{Var}(Z_2) \geq \text{Var}(Z_3) \geq \cdots \geq \text{Var}(Z_p) \) where \( \text{Var}(\cdot) \) denotes the variance. Principal component analysis works as follows:

Assume we have a \( p \) number of \( X \) variables as \( x_1, x_2, x_3, \ldots, x_p \), then their combination is:

\[
Z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p
\]

(5.1)

Then the second principal component:

\[
Z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p
\]

(5.2)

The above is chosen so that \( \text{Var}(Z_2) \) is as large as possible subject to the constraint that:

\[
a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1
\]

(5.3)

Since \( Z_1 \) and \( Z_2 \) have zero correlation for the data. Therefore, the \( i^{th} \) principal component is:

\[
Z_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p
\]

(5.4)

In particular, \( \text{Var}(Z_i) = \lambda_i \) and the constants \( a_{11}, a_{12}, \ldots, a_{ip} \), are the elements of the corresponding eigenvectors scaled so that:

\[
a_{11}^2 + a_{12}^2 + \cdots + a_{ip}^2 = 1
\]

(5.5)

Future information on the principal components analysis can be found in [80, 78]

5.1.2 Linear and Quadratics Discriminant Analysis

Discriminant Analysis (DA) is a statistical model which uses a set of independent variables to discriminate between two or more groups of a dependent variable [81]. The linear discriminant analysis (LDA) is, also known as Fisher’s linear discriminant, a method used in machine learning, statistics and pattern recognition. LDA is most commonly used as dimensionality reduction technique to find a linear combination of features that characterizes or separates two or more classes of objects or events. The Quadratic Discriminant Analysis (QDA) is a generalization of the LDA.

The LDA and QDA are the simplest discrimination analyses due to their linearity. A linear decision boundary is easy to understand and visualize. Assume that we draw a sample from a multivariate normal distribution \( N(\mu, \Sigma) \) that has a mean \( \mu \) and covariance matrix \( \Sigma \). The multivariate density can then be:

\[
P(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)}
\]

(5.6)

A sample of data is drawn from two classes, each described by a multivariate normal density

\[
P(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)}
\]

(5.7)

LDA can be applied as follow: Assume that each of \( f_k(x) \), \( k = 1, \cdots, K \) follows a multivariate normal distribution \( (\mu_k, \Sigma) \) with a mean \( \mu_k \) and common covariance matrix \( \Sigma \) [24]. Then:

\[
\log \frac{P(G = k|x)}{P(G = j|x)} = \log \frac{\pi_k}{\pi_j} + \frac{1}{2} (\mu_k - \mu_j)^T \Sigma^{-1} (\mu_k - \mu_j) + x^T \Sigma^{-1} (\mu_k - \mu_j)
\]

(5.8)

If 5.8 is greater than zero then \( k \) will be grouped in a vector; otherwise \( j \) will be grouped instead. For classes \( k = 1, 2 \) the posterior probability \( P(k|x) \) of observing an instance of class \( k \) at point \( x \) can be found using Bayes rule as follows:

\[
P(k|x) = \frac{\pi_k P(k|x)}{P(x)}
\]

(5.9)

Note that the unconditional probability \( P(x) \) does not depend on \( k \). Taking the natural logarithm of the posterior odds:

\[
\log \frac{P(k = 1|x)}{P(k = 2|x)} = \log \frac{\pi_1}{\pi_2} - \frac{1}{2} \log \frac{|\Sigma_1|}{|\Sigma_2|} + x^T \left( \Sigma_1^{-1} \mu_1 - \Sigma_2^{-1} \mu_2 \right) - \frac{1}{2} \left( \Sigma_1^{-1} + \Sigma_2^{-1} \right) x
\]

(5.10)
We can obtain the hyperplane that separates the two classes by equating this log-ratio to zero [82]. The above is the quadratic function of \( x \). Hence, it is called the Quadratic Discriminant Analysis (QDA). If the covariance matrix \( \Sigma_1 = \Sigma_2 \), then the quadratic terms disappear making it Linear Discriminant Analysis (LDA).

5.1.3 Partial Least Squares

Partial Least Squares (PLS) is a regression approach that splits the predictors or independent variables to a smaller set of uncorrelated components then performs least square regression on them rather than on the whole original data [83, 84].

The core functionality of PLS is dimension reduction which works under the assumption of a basis latent decomposition of the predictor matrix in which \( X \in \mathbb{R}^{n \times p} \) and response matrix in which \( Y \in \mathbb{R}^{n \times q} \) as follows:

\[
X = TP^T + E \quad (5.11)
\]

\[
Y = UQ^T + F \quad (5.12)
\]

where, \( X \in \mathbb{R}^{n \times m} \) is a matrix that is capable of producing \( k \) linear combinations known as scores, \( Y \in \mathbb{R}^{n \times q} \) matrix of responses, \( T \in \mathbb{R}^{n \times l} \) is called X-scores \( U \in \mathbb{R}^{n \times l} \) is called the Y-scores, \( P \in \mathbb{R}^{m \times l} \) and \( Q \in \mathbb{R}^{p \times l} \) are matrices of coefficients known as loadings and \( E \in \mathbb{R}^{n \times p} \) and \( F \in \mathbb{R}^{n \times q} \) are random errors matrices.

5.1.4 Canonical Variate Analysis

Canonical Variate Analysis (CVA) is a technique that focuses on finding the relationship between groups of variables in a data set. Assumes that we have two data sets, namely \( X \) and \( Y \) such that \( X = \{x_1, x_2, x_3 \ldots x_n\} \) and \( Y = \{y_1, y_2, y_3 \ldots y_n\} \). The CVA main function is to find whether \( X \) and \( Y \) are related, i.e., can \( Y \) be represented by \( X \)? Basically, CVA finds the linear combinations of \( X \) and \( Y \) that are highly correlated.

6 CONCLUSION

The concept of the Remaining Useful Life (RUL) studied herein is also known as the Mean Residual Life (MRL) for an engineering device. This is the corner stone of the emerging field of engineering prognostics, which mimics the older field of medical prognostics, by allowing an engineering device to play a role similar to that of the human body. The Remaining Useful Life is simply a scientific prediction of the time in which an engineering device will no longer perform its intended function. Prognostics is, therefore, the field of predicting the future reliability and performance of an engineering device by assessing the deviation or degradation extent of the device from its normal operation condition expectation.

This paper surveys the prediction techniques for the Remaining Useful Life (RUL) of an engineering device under continuous operation. The paper classifies these techniques into four categories, namely model-based techniques, knowledge-based techniques, experience-based techniques, and data-driven techniques. A comparative exposition is given for the main features, prominent advantages, potential shortcomings and main subcategories for each of these technique categories. The survey is supported by an extensive list for up-to-date references. The present survey is intended to supplement many excellent reviews that address the prognostics of devices of contemporary vital importance [85, 86, 87, 88, 89, 90, 91, 92, 93, 94].

ACKNOWLEDGMENT

This article was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. The authors, therefore, acknowledge with thanks DSR technical and financial support.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES


