An Integrated Approach to Bearing Fault Diagnostics and Prognostics

Xiaodong Zhang, Roger Xu, Chiman Kwan, Steven Y. Liang, Qiulin Xie, and Leonard Haynes

Abstract—**This paper presents an integrated fault diagnostic and prognostic approach for bearing health monitoring and condition-based maintenance. The proposed scheme consists of three main components including principal component analysis (PCA), hidden Markov model (HMM), and an adaptive stochastic fault prediction model. The principal signal features extracted by PCA are utilized by HMM to generate a component health/degradation index, which is the input to an adaptive prognostics component for on-line remaining useful life prediction. The effectiveness of the scheme is shown by simulation studies using experimental vibration data obtained from a bearing health monitoring testbed.**

I. Introduction

The ability to accurately predict the remaining useful life of electromechanical systems is critical for affordable system operation and can also be used to enhance system safety. The theme of condition-based maintenance (CBM) is that maintenance is performed based on an assessment or prediction of the component health in stead of its service time, which achieves objectives of cost reduction and safety enhancement. If one can predict the degradation of a component before it actually fails, then it will provide ample time for maintenance engineers to schedule a repair, and to acquire replacement components before the components actually fail.

Bearings are of paramount importance to almost all forms of rotating machinery, and are among the most common machine elements. The failures of bearing without warning will result in catastrophic consequences in many situations, such as in helicopters, transportation vehicles, etc. Most of the current maintenance procedure includes periodic visual inspections and replacement of the components at fixed time intervals. The application of CBM will clearly result in a much lower operational cost and higher availability, which can be enhanced further by damage mitigation techniques. The key to successful CBM consists of early detection of faults of small magnitude, accurate assessment of current fault or defect size, and accurate prediction of fault progression [1].

Prior papers by the authors mainly deal with the problem of fault detection and isolation (FDI), and qualitative assessment of component health [2, 3]. In this paper, we present a unified fault diagnostic and prognostic (FDP) approach to anomaly detection, fault diagnosis, degradation status assessment, and remaining useful life prediction. The proposed FDP architecture consists of three diagnostics and prognostics tools including principal component analysis (PCA), hidden Markov models (HMM), and an adaptive prognostic model [4]. More specifically, the principal features extracted by PCA are utilized by HMM to generate a health/degradation index representing the current system health status, which is the input to an adaptive prognostics component for on-line remaining useful life prediction.

The paper is organized as follows. In Section II, the integrated FDP architecture is presented. The simulation results shown in Section III illustrate the effectiveness of the proposed method by using some real experimental vibration data obtained from a bearing testbed at Georgia Institute of Technology. Finally, some concluding remarks are given in Section IV.

II. The Integrated Diagnostic and Prognostic Architecture

A block diagram of the proposed FDP architecture is shown in Figure 1 (see page 6). Basically, it combines three diagnostics and prognostics tools, including principal component analysis (PCA), hidden Markov models (HMM), and adaptive prognostics for remaining useful life prediction. In addition, the preprocessing component makes the data suitable for PCA computation by performing some simple manipulations such as frame blocking, frequency spectral analysis, and noise filtering. In the proposed architecture, these three components operate in a unified fashion and provide complementary fault information to maintenance engineers. The principal features extracted by PCA are used by HMM to perform on-line assessment of

Manuscript received September 24, 2003. This work was supported by the USA Naval Sea Systems Command Indian Head under Grant N00174-03-C-0049.

X. Zhang, R. Xu, C. Kwan, and L. Haynes are with Intelligent Automation, Inc, 15400 Calhoun Dr., Suite 400, Rockville, MD 20855, USA. (phone: 301-294-5269; fax: 301-294-5201; e-mail: xzhang@i-a-i.com).

S. Y. Liang and Q. Xie are with the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

the component health/degradation status. Then based on the degradation index $D(t)$ generated by HMM, the adaptive prognostics component performs remaining useful life prediction. The degradation index prediction error $e(t) = D(t) - \hat{D}(t)$ is used to continuously update the fault propagation model to improve prediction accuracy.

More specifically, the main functions of each of these three components are summarized as follows:

1) PCA extracts principal signal features and achieves data reduction. The residual models generated by PCA are also useful for anomaly detection (for example, new and unanticipated faults for which no training data is available *a priori*).
2) HMM provides robust FDI decisions owing to its powerful stochastic modeling capability. The state sequence estimation generated by HMM gives qualitative information about the current component health/degradation status. In addition, the component health/degradation index generated by HMM provides a quantitative assessment of the current component health status, which can be used for fault prediciton.
3) The adaptive prognostics algorithm predicts the remaining useful life of the component by using the degradation index provided by HMM. This is very important for CBM avoid catastrophic failures or unnecessary replacement of components.

In the following sections, we will briefly describe the three main components shown in Figure 1.

A. Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical approach to analyze data sets with significant redundant information. This technique has been applied to many applications including signal processing, data transmission and storage, and pattern recognition, etc [5]. Its main function is to retain the most important characteristics of its input by using a small amount of vectors, i.e., the principal components. In recent years, PCA has also been applied to health monitoring applications [6].

Consider the preprocessed sensor data to be a $(q \times k)$ matrix X, where each row represents an observation. Let us denote p_i as the loading vectors [6], where $i \le k$. The first principal component is defined as the linear combination $t_1 = Xp_1$ that has the maximum variance subject to $|p_1| = 1$. The second principal component is the linear combination defined by $t_2 = Xp_2$ that has the next greatest variance subject to $|p_2| = 1$ and subject to the condition that it is orthogonal to the first principal component t_1. Up to k principal components can be similarly defined. Suppose the first N principal components capture an adequate approximation of the matrix X. Then we define the last $k - N$ component as minor components. Based on the above discussion, PCA decomposes the X matrix as

$$X = \sum_{n=1}^{N} t_n p_n^T + \sum_{m=N+1}^{k} t_m p_m^T \quad , \tag{1}$$

where t_n and t_m are the principal and minor components, respectively, p_n and p_m are the corresponding loading vectors. Denote $T \triangleq (t_1 \cdots t_N)$, $\tilde{T} \triangleq (t_{N+1} \cdots t_k)$, $P \triangleq (p_1 \cdots p_N)$, and $\tilde{P} \triangleq (p_{N+1} \cdots p_k)$. Then we can rewrite (1) as follows:

$$X = TP^\top + \tilde{T}\tilde{P}^\top .$$

Using the orthogonality of matrices P and $\tilde{P}$, we have

$$X\tilde{P} = \tilde{T}\tilde{P}^\top \tilde{P} = \tilde{T} \approx 0_{q \times (k-N)} , \tag{2}$$

where the entries of $\tilde{T}$ are minor components that are usually very small, as defined above.

Equation (2) provides a residual model that can be used for anomaly detection. Moreover, the residual models can be continuously updated to include the latest fault information. The idea of *adaptive PCA* is best illustrated by a simple example shown in Figure 2(a) and Figure 2(b). Suppose that at the beginning we only have the data corresponding to the normal operating condition. Then a residual model $B = \tilde{P}^\top$ corresponding to the normal operating condition can be generated by using (2). As long as the system continues to work under normal operating conditions, the output of the residual model will remain around zero. Otherwise, we can conclude that an unknown fault 1 has occurred. Then we can generate a new residual model B_1 using the sensor data corresponding to that particular type of fault. This procedure can be repeated for any other new types of faults.

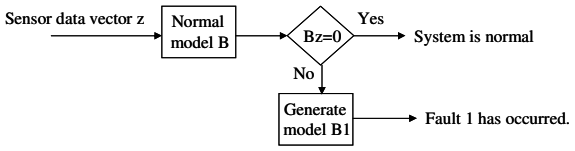


Figure 2 An illustrative example of adaptive PCA

B. HMM-based Diagnostics/Prognostics

The fault progression process of mechanical systems usually consists of a series of degraded states. This process can be ideally described by a mathematical model known as hidden Markov model (HMM). The word "hidden" means the HMM states are not directly observable. In other words, the HMM states can only be observed through a set of stochastic processes that produce the sequence of observations. A diagram depicting the state transitions of a bearing from normal to failure is shown in Figure 3.

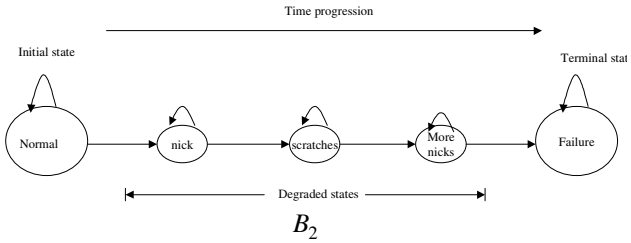


Figure 3 An HMM describing the failure mechanism of a bearing

A block diagram of the proposed HMM-based health monitoring scheme is shown in Figure 4 (see page 6). More specifically, a bank of trained continuous HMMs estimates on-line the health status of the system. Each HMM corresponds to a particular type of fault condition, such as normal condition (without faults), fault 1, fault 2, etc. Moreover, the states of each HMM represent different degradation status of a particular fault that grows from a very small size to a complete failure. A decision scheme integrates the fault information generated by each HMM. Two types of fault information are provided by the decision scheme: (1) fault detection and isolation by comparing the likelihood generated by each HMM; (2) health/degradation status estimation and degradation index generated after the isolation of a particular fault. More specifically, the HMM-based scheme is capable of providing the following three types of fault diagnostic/prognostic information:

1) Given some particular sensor data, we determine the type of fault that has occurred by choosing the HMM that has the largest likelihood (i.e., fault detection and isolation);
2) By monitoring the progression of the state sequence generated by that particular HMM, we can have a qualitative understanding of the past fault progression process and the current degradation status. The fault information obtained at this stage is qualitative because a range of fault/defect size can be classified as belonging to one HMM state.
3) Additionally, the system degradation index generated by HMM provides a quantitative estimation of the system health/degradation status, which can be used for component remaining useful life prediction.

In general, an HMM is described by $\lambda = (A, B, \pi)$, where the matrices A and B represent the state transition probability distribution and observation symbol probability distribution, respectively, and π is the initial state distribution. In this scheme, each of the hidden HMM states represents the degradation status of a particular fault, for instance small fault size, intermediate fault size, and significant fault size, etc. Over the past 30 years, various fast and efficient training algorithms have been presented in the speech recognition literature, for example the well-known Baum-Welch method [7]. However, such training algorithms are not directly applicable in fault diagnosis applications, since it is usually not possible to obtain adequate sequence of data describing a continuous fault propagation process over time from state to state, as discussed in the paper by Smith [8]. The HMM training algorithms used here follows the ideas of [8] and [9].

As described above, in addition to providing a qualitative estimation of the degradation, the proposed HMM also generates a degradation index by using the probability associated with the final state representing a complete failure. The basic idea is as follows. For all the sensor data, we can estimate its probability associated with the last state of the HMM. In a sense, this probability indicates the similarity between the current degradation status and a complete failure represented by the final state of the HMM. To get a better insight into this idea, let us consider the following scenario. When the component is under normal condition, its probability associated with the last state of HMM should be zero. After a fault occurs and starts to propagate, the probability of the input data associated with the last state of HMM will slowly increase, since the features are becoming more and more similar to a complete failure. At the end of the fault progression process, the probability of the sensor data associated with the final state of the HMM will become one, which means a complete failure. Therefore, the degradation index generated by HMM provides a measure of the current component health, which is very useful for fault prediction.

It is worth noting that by monitoring the progression of state sequences and degradation index generated by HMM, we can perform fault prognosis and diagnosis in a unified fashion. Conventional techniques using HMM only deal with the fault diagnosis problem [8, 10, 11].

C. Adaptive Prognostics for Remaining Useful Life Prediction

It is our belief that an accurate prognostic method should be intimately based on the diagnostic results. That is, prognostics should be done at regular intervals and should rely on on-line diagnostic results, which provide the latest information about the system health status. In other words, the prognostic algorithm needs to continuously update its model and offer a revised prediction based on the current health status of the system.

In this work, we employ an adaptive prognostic method presented in [4], which use a stochastic modeling method to characterize the uncertainty in material properties, process conditions, or environmental factors. More specifically, the fault progression model is given by

$$\dot{D} = \frac{dD}{dt} = C_0 D^n e^{Z(t)}, \qquad (3)$$

where $D(t)$ is the fault size, C_0 and n are constants depending on the material property of the system, and $e^{Z(t)}$ is called the lognormal random variable that characterize

the amount of uncertainty in material properties or environmental factors. The random variable $Z(t)$ is assumed to be a stationary, exponentially correlated, Gaussian-Markov Process with the following equation

$$\dot{Z} = -\xi Z(t) + w_z(t),$$

where $w_z(t)$ is a Gaussian noise with zero mean and with $\mathrm{E}\{w(t)w(\tau)\} = \sigma_w^2 \delta(t-\tau)$.

An adaptive algorithm has been derived to update the model parameters based on the latest diagnostic information $D(t)$. Therefore, the effects of maintenance activities and changing operating conditions are considered. Moreover, a fourth-order stochastic differential equation was developed for fault progression prediction, which takes into account uncertainties due to the complexity of defect progression and diagnostic model inaccuracies. Detailed description and analysis of the fault prediction algorithm can be found in [4].

III. Simulation Results

Simulation studies have been carried out to verify the effectiveness of the integrated diagnostic and prognostic scheme by using real experimental vibration data obtained from bearing testing equipment at Georgia Institute of Technology [4].

Defective bearing would cause mechanical systems to vibrate abnormally. The vibrations are typically monitored by accelerometers. In this work, the test bearing was Timken LM 50130 cup (out race) and LM 501349 cone (inner race). To accelerate the propagation process, a defect was artificially initiated on the outer raceway by an electrical discharge machine. The defect was placed in the loading zone. The testing was interrupted and disassembled to perform physical inspection of the defect at 0.378, 1.566, 3.51, 7.452 and 9.666 million cycles. The corresponding vibration data were recorded in each of these periods at a sampling rate of 30k. No defect propagation was observed before the 9.666 million cycles. In the last assembly after 9.666 million cycles, the testing continued until 13.986 million cycles. Excessive vibration was observed during that period. Since defect propagation was observed to happen only in the last running, the detailed analysis was performed only for that period.

Figure 5(a) and Figure 5(b) show the time segments of vibration signals from a normal bearing and a defective one, respectively. Their corresponding power spectrum densities (PSD) are shown in Figure 6(a) and Figure 6(b). It has been observed that the vibration level in the range of 3500-5000 Hz generated by the defective bearing is closely correlated to a defect propagation process. Therefore, a band-pass filter was applied to the raw vibration data.

In our simulation studies, we used two sets of real vibration data, which correspond to a normal bearing and a

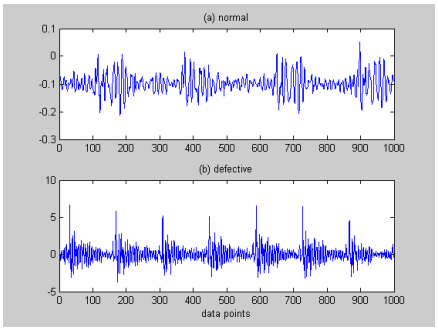


Figure 5: Raw vibration signals from a normal bearing and a defective bearing

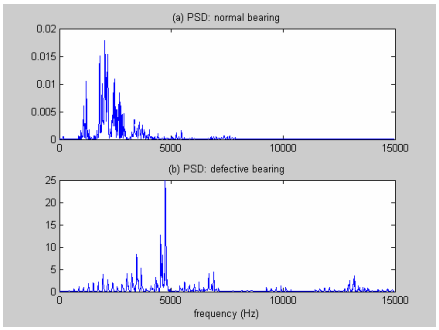


Figure 6: The corresponding power spectrum densities

significantly defective one, respectively. An additional set of data set were obtained via interpolation to represent an intermediate defect condition. Therefore, an HMM with three states was used to model these three different levels of defect, i.e., normal condition, intermediate defect, and significant defect that could lead to a catastrophic failure of the bearing.

As described in Section 2, the proposed fault prognostics scheme mainly consists of PCA for feature extraction, HMM for degradation state and index estimation, and an adaptive prognostic method for remaining useful life prediction. Below we show some simulation results.

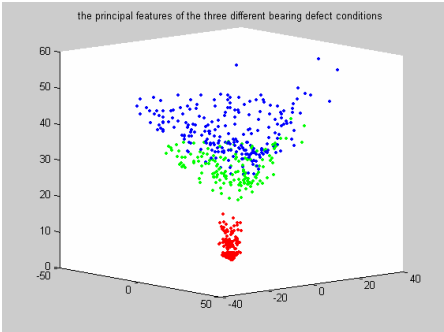


Figure 7 The principal features extracted by PCA

The principal features extracted by PCA are shown in Figure 7, where the clusters in red, green, and blue represent the normal condition, an intermediate defect, and a significant defect, respectively. Then the extracted

features are used to estimate the HMM model parameters in the training process. We conducted the following two types of tests to evaluate the performance of the trained HMM model. The first test is degradation state estimation, while the second test is degradation index estimation.

In our first test, three sets of new data corresponding respectively to the three different defect conditions (i.e., normal, intermediate, and significant) were used to constitute the test data set. The data length for each condition is the same. The degradation state estimation generated by the HMM is shown in Figure 8. As we can see, the HMM is able to detect the change of each defect condition immediately after it occurs.

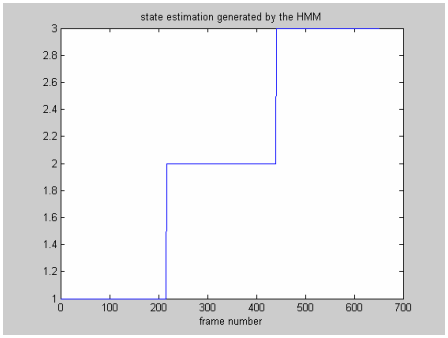


Figure 8 The degradation status estimation of bearing defect conditions generated by HMM

In our second test, interpolation was performed to the data corresponding to the three known different defect conditions to simulate a continuous defect propagation process. The data interpolation was based on the well-known Paris's formula [4]. The actual defect propagation curve obtained using the Paris's formula and its estimation generated by HMM are shown in Figure 9, respectively. The estimated degradation index gives a reasonable estimation of the true defect propagation curve. It is worth noting that the variations in the estimated curve is due to various modeling uncertainty in the fault progress process (see (3)).

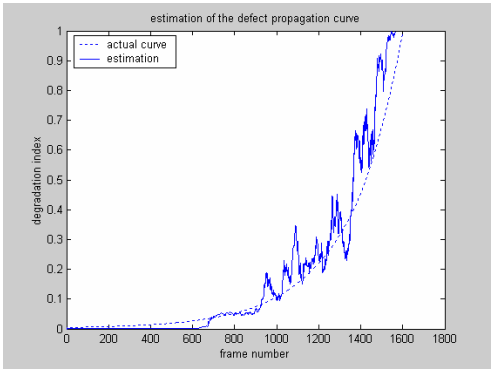
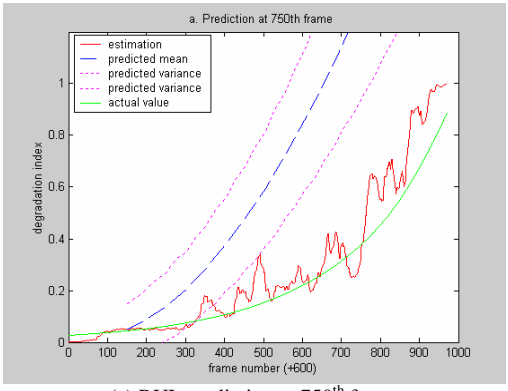
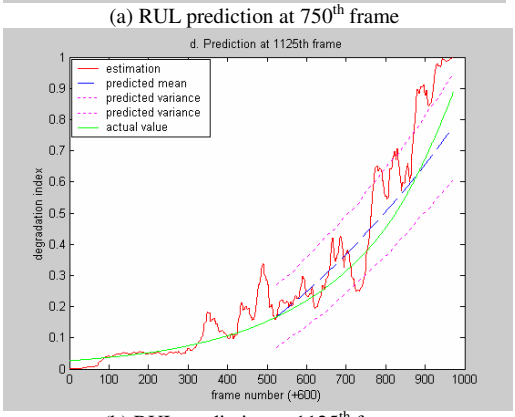


Figure 9 The actual degradation progression curve (dashed line) and its estimation (solid line) provided by HMM

Finally, the estimated degradation index was used to predict the remaining useful life of the bearing. As mentioned early, one of the key advantages of the proposed adaptive prognostics algorithm is its on-line learning capability. In other word, the defect progression model parameters are online updated using the latest diagnostic information, therefore improving the prediction accuracy [4]. It is worth noting since the first 600 frames shown in Figure 9 give little information about fault progression, so only the estimated degradation index starting from the 600th frame was used to predict the bearing's remaining useful life. Figure 10(a) and Figure 10(b) show the RUL prediction results obtained at 750th frame and 1125th frame, respectively. The actual degradation growth curve and its estimation provided by HMM are plotted in green and red solid lines, respectively. Moreover, the dashed blue line is the predicted mean, while the two dotted curves in magenta represent the predicted variances. By comparing Figure 10(a) and Figure 10(b), it is obvious that the fault prediction accuracy is significantly improved as the prediction model is continuously updated.



(a) RUL prediction at 750th frame



(b) RUL prediction at 1125th frame

Figure 10 Remaining useful life (RUL) prediction

It is worth noting that the component's *remaining useful life* can be easily determined by choosing a threshold representing a critical damage level that could lead to a failure of the bearing. Then the difference between the current time and the time instant when the predicted degradation index reaches this threshold defines the remaining useful life of the component.

IV. CONCLUSION

In this research work, three fault diagnostic and prognostic algorithms have been integrated for bearing health monitoring. The proposed unified framework is capable of performing anomaly detection, fault detection and isolation, health/degradation estimation, and remaining useful life prediction. Simulation results using some real bearing vibration data have shown the feasibility of the proposed method.

Future work includes the following: First, more data intensive validation and performance evaluation; Second, the development of open modular software architecture and the implementation of the fault diagnostics and prognostics algorithms.

REFERENCES

[1] L. Atlas, G. Bloor, T. Brotherton, L. Howard, L. Jaw, G. Kacprzynski, G. Karsai, R. Mackey, J. Mesick, R. Reuter, and M. Roemer, "An Evolvable Tri-Reasoner IVHM System", *Proceedings of the 2001 IEEE Aerospace Conference*, 11.0307, Big Sky, Montana, USA, March 13, 2001

[2] X. Zhang, M. M. Polycarpou, and T. Parisini, "A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems", *IEEE Transactions on Automatic Control*, vol. 47, no. 4, pp. 576-593, April 2002.

[3] C. Kwan, X. Zhang, R. Xu, and L. Haynes, "A novel approach to fault diagnostics and prognostics", *Proceedings of the 2003 IEEE International Conference on Robotics and Automation*, September 2003, pp. 604-609.

[4] Y. Li, T. R. Kurfess, and S. Y. Liang, "Stochastic prognostics for rolling element bearings", *Mechanical Systems and Signal processing*, 14(5), 2000, pp. 747-762.

[5] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Second Edition, Prentice Hall, 1999.

[6] T. Kourti, "Process analysis and abnormal situation detection: from theory to practice", *IEEE Control Systems Magazine*, 22(5), 2002, pp.10-25.

[7] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, no. 2, 1989.

[8] P. Smyth, "Markov monitoring with unknown states", *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 9, 1994, pp. 1600-1612.

[9] D. A., Reynolds, and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1., 1995.

[10] H. M. Ertunc, K. A. Loparo, and H. Ocak, "Two wear condition monitoring in drilling operations using hidden Markov models", *Machine Tools & Manufacture*, 41, pp. 1362-1384, 2001

[11] L. Wang, M. G. Mehrabi, and E. Kannatey-Asibu, Jr., "Hidden Markov model-based tool wear monitoring in turning", *Journal of Manufacturing Science and Engineering*, vol. 124, pp. 651-658, August 2002.

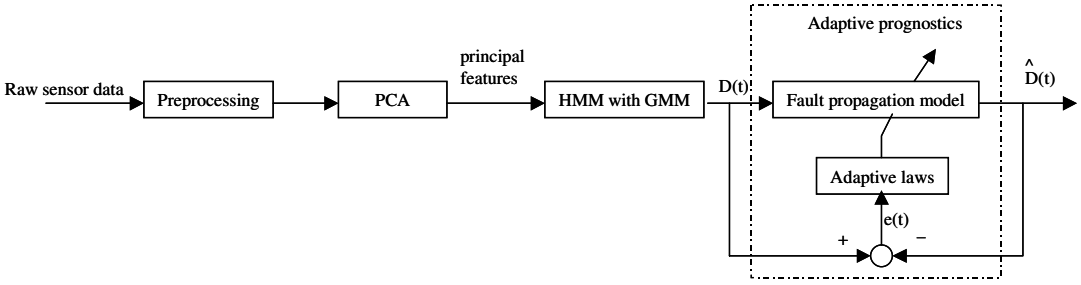


Figure 1 An integrated diagnostics and prognostics framework.

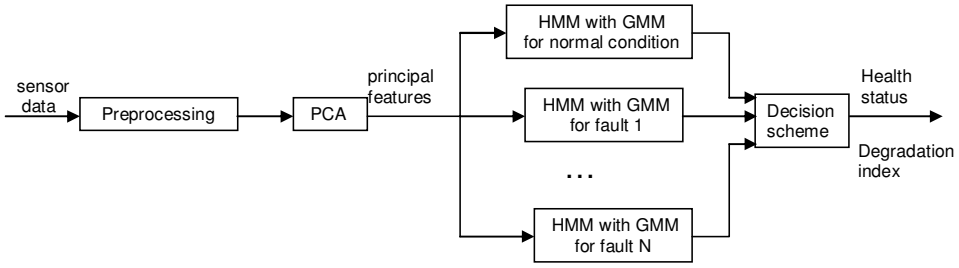


Figure 4 A block diagram of HMM-based fault diagnostic and prognostic scheme