Fault Prognosis Using Dynamic Wavelet Neural Networks

P. Wang and G. Vachtsevanos

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250
Email: gjv@ece.gatech.edu

Abstract

Large manufacturing companies are considering to deliver to their customer base “guaranteed uptime” instead of the conventional service contracts. Modern industry is concerned about extending the lifetime of its critical processes and maintaining them only when required. Significant aspects of these trends include the ability to diagnose impending failures, prognose the remaining useful lifetime of the process and schedule maintenance operations so that uptime is maximized. Prognosis is probably the most difficult of the three issues leading to condition-based maintenance. This paper attempts to address this challenging problem with intelligence-oriented techniques, specifically dynamic wavelet neural networks. Dynamic wavelet neural networks incorporate temporal information and storage capacity into their functionality so that they can predict into the future, carrying out fault prognostic tasks. An example is presented in which a trained dynamic wavelet neural network successfully prognoses a defective bearing with a crack in its inner race.

Introduction

The manufacturing and industrial sectors of our economy are increasingly called to produce at higher throughput and better quality while operating their processes at maximum yield. As manufacturing facilities become more complex and highly sophisticated, the quality of the production phase has become more crucial. The manufacture of such typical products as aircraft, automobiles, appliances, medical equipment, etc, involves a large number of complex processes most of which are characterized by highly nonlinear dynamics coupling a variety of physical phenomena in the temporal and spatial domains. It is not surprising, therefore, that these processes are not well understood and their operation is “tuned” by experience rather than through the application of scientific principles. Machine breakdowns are common limiting uptime in critical situations. Failure conditions are difficult and, in certain cases, almost impossible to identify and localize in a timely manner. Scheduled maintenance practices tend to reduce machine lifetime and increase down-time, resulting in loss of productivity. Recent advances in instrumentation, telecommunications and computing are making available to manufacturing companies new sensors and sensing strategies, plant-wide networking and information technologies that are assisting to improve substantially the production cycle. Machine diagnostics/prognostics for condition-based maintenance involves an integrated system architecture with a diagnostic module – the diagnostician – which assesses through on-line sensor measurements the current state of critical machine components, a prognostics module – the prognosticator – which takes into account input from the diagnostician and decides upon the need to maintain certain machine components on the basis of historical failure rate data and appropriate fault models, and a maintenance scheduler whose task is to schedule maintenance operations without affecting adversely the overall system functionalities of which the machine in question is only one of its constituent elements.

This paper addresses issues relating to the prognostic module – the Achilles heel of the Condition-Based-Maintenance (CBM) architecture. Fault diagnosis is a mature field with contributions ranging from model-based techniques to data-driven configurations that capitalize upon soft computing and other “intelligent” tools [1][2]. Condition-based maintenance scheduling is a complex task that involves finding the “optimum” time to perform maintenance within the window prescribed by the Prognosticator while meeting a host of constraints. This scheduling problem may be formulated as a multi-objective optimization problem where the main objective is to maximize process uptime while satisfying a set of constraints that relate to resource and maintenance personnel availability, production and scheduling requirements, redundant or relocatable machines, timing constraints, etc [3]-[5]. The word “prognosis” implies the foretelling of the probable course of a disease [6], a term widely used in medical practice. In the industrial and manufacturing arenas, prognosis is interpreted to answer the question: what is the remaining useful lifetime of a machine or a component once an impending failure condition is detected and identified? Stochastic Auto-Regressive Integrated Moving Average (ARIMA) models [7], fuzzy pattern recognition principles [8], knowledge-intensive expert systems [9], nonlinear stochastic models of fatigue crack dynamics [10], polynomial neural networks [11] and other techniques have been introduced over the past years to address the diagnostic/prognostic problem. This paper attempts to address this issue by introducing a novel combination of a “virtual” sensor as a mapping tool between known measurements and “difficult-to-access” quantities and a dynamic wavelet neural network as the “predictor”, i.e. the construct that projects into the future the temporal behavior of a faulted component.

The Prognosticator

The prognosticator performs the vital function of linking the diagnostic information with the maintenance scheduler. It is probably the least understood but most crucial component of the diagnostic/prognostic/CBM hierarchical architecture. Furthermore, it entails ambiguity and large-grain uncertainty since the historical evolution of a failure event – the growth of a structural fault, for example – is difficult if not impossible to model accurately, historical data is not readily available and the particular growth phenomenon may be strongly dependent on the system structure, operating conditions, environmental effects, etc. It is viewed as a dynamic predictor which receives fault data from the diagnostic module and determines the allowable time window during which machine maintenance must be performed if the integrity of the process is to be kept as high as possible. The term “dynamic predictor” implies also the functional requirement that the target output, i.e. remaining useful lifetime or time-to-failure, is dynamically updated as more information becomes available from the diagnostician. Thus, this scheme should reduce the uncertainty and improve the prediction accuracy as the accumulated evidence grows.

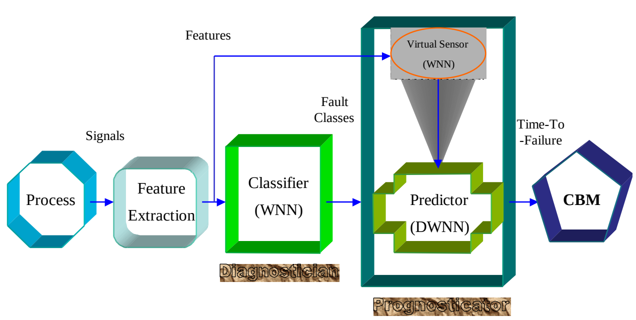

Figure 1 The overall architecture of the prognostic system

Figure 1 depicts the overall architecture of the prognostic system. The diagnostician monitors continuously critical sensor data and decides upon the existence of impending or incipient failure conditions. The detection and identification of an impending failure triggers the prognosticator. The latter reports to the CBM module the remaining useful lifetime of the failing machine or component. The CBM module schedules the maintenance so that uptime is maximized while certain constraints are satisfied. The schematic of Figure 1 focuses on the functionalities of the prognosticator. The diagnostician alerts the prognostic module and provides failure and other pertinent sensor data to it. The prognostic architecture is based on two constructs: 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 the time evolution of the failure mode and allowing the estimation of the component’s remaining useful lifetime. Both constructs rely upon a wavelet neural network model acting as the mapping tool. It is appropriate, therefore, to digress for a brief discussion of the Wavelet Neural Network (WNN).

The Wavelet Neural Networks

The Wavelet Neural Network belongs to a new class of neural networks with unique capabilities in addressing identification and classification problems. Wavelets are a class of basic elements with oscillations of effectively finite-duration that makes 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. On the other hand, artificial neural networks constitute a powerful class of nonlinear function approximants for model-free estimation. A common ground between these two technologies may be coherently exploited by introducing a WNN. Indeed, the implementation of a neural network is closely related to a truncated version of the wavelet series.

A MIMO WNN can be formulated as [12]:

$$y = [\psi_{A_1,b_1}(x)\ \psi_{A_2,b_2}(x)\ \cdots\cdots\ \psi_{A_M,b_M}(x)]C + [x\ 1]C_{lin} \qquad (1)$$

where x is the 1×n input row-vector; y is the 1xK output row-vector and K is the number of outputs; A_j is the nxn squashing matrix for the jth node; b_j is the 1xn translation vector for the jth node; C is the M×K matrix of output coefficients, where M is the number of wavelet nodes; C_{lin} is the (n+1)×K matrix of output coefficients for the linear direct link; and ψ is the wavelet function that can take the form:

$$\psi_{A,b}(x) = |A|^{1/4}\psi\left(\sqrt{(x-b)A(x-b)^T}\right) \qquad (2)$$

where x is the input row-vector; A the squashing matrix for the wavelet; b the translation vector; and T the transpose operator. Composed of localized basis functions, the WNNs are suitable for capturing the local nature of the data patterns and thus are efficient tools for both classification and approximation problems.

The WNN of (1) is a static model in the sense that it establishes a static relation between its inputs and outputs. All signals flow in a forward direction only with this configuration. Dynamic or recurrent neural networks, on the other hand, are required to model the time evolution of dynamic systems. Signals in such a network configuration can flow not only in the forward direction but also can propagate backwards, in a feedback sense, from the output to the input nodes. Dynamic Wavelet Neural Nets have recently been proposed to address the prediction/classification issues. A multi-resolution dynamic predictor that utilizes the discrete wavelet transform and recurrent neural networks forming nonlinear models for prediction was designed and employed for multi-step prediction of the intra-cranial pressure signal [13]. A recurrent wavelet neural network was developed for the blind equalization of nonlinear communication channels [14]; recurrent wavelet neural networks were also derived in [15] using the real-time Back-Propagation algorithm.

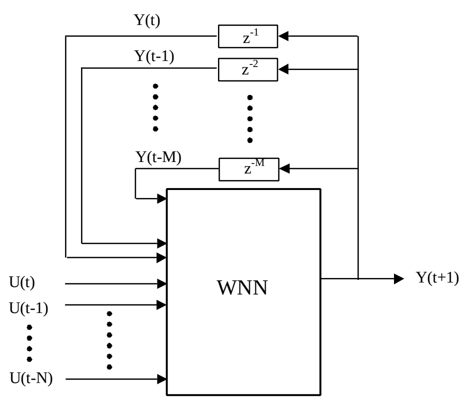

Figure 2 A dynamic wavelet neural network.

The basic structure of a DWNN is shown in Figure 2. Delayed versions of the input and output augment now the input feature vector and the resulting construct can be formulated as:

$$Y(t+1) = WNN(Y(t),\cdots,Y(t-M),U(t),\cdots,U(t-N)) \quad (3)$$

where U is the external input; Y is the output; M is the number of outputs; N is the number of external inputs; and WNN stands for the static WNN. The DWNN described by (3) 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. In addition, such fundamental performance concerns as stability can be examined using system-theoretic concepts, for example, Lyapunov stability theory.

The Virtual Sensor

It is often true that machine or component faults are not directly accessible for monitoring their growth behavioral patterns. Consider, for example, the case of a bearing fault. No direct measurement of the crack dimensions is possible when the bearing is in an operational state. That is, there is no such device as a “fault meter” capable of providing direct measurements of the fault evolution. Examples of a similar nature abound. In [16], the authors report on the development of a neural net based virtual or ideal sensor used to diagnose engine combustion failures, known as misfire detection. Their technique employs a recurrent neural net as the classifier that takes such inputs as crankshaft acceleration, engine speed, engine load and engine ID and produces a misfire diagnostic evaluation as the output. In the present study, the same concept is exploited to design a virtual sensor which takes as inputs measurable quantities or features and outputs the time evolution of the fault pattern. A schematic representation of the WNN as a virtual sensor is shown in Figure 3.

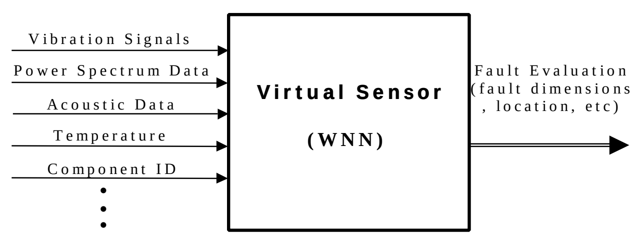

Figure 3 A schematic representation of the WNN as a virtual sensor

The Predictor

Prediction of the course in which a fault could develop can be looked into from two different viewpoints: one view is to locate the fault value at a certain time moment and the other is to find the time moment when the fault reaches a given value, i.e. the fault dimensions reach a pre-specified threshold. The latter appears to be more meaningful because it concentrates on revealing the critical time without requiring estimation of the whole time interval, thus resulting in a more efficient algorithm. The notion of Time-To-Failure (TTF) is the most important measure in prognosis. In fact, prognosis can be accomplished in either the time or frequency or even the event domain, since all of these domains are made up of ordered points.

A fault predictor based on the DWNN is illustrated in Figure 4. The process is monitored real-time using appropriate sensors. Here, virtual sensors can also be employed to measure signals or their derivatives that are difficult to record on-line and on-site. Data obtained from measurements are continuously processed and features extracted on a time scale. The features are organized into a time-stamped feature vector that serves as the input to the DWNN. Consequently, the DWNN performs as a dynamic classifier or identifier. The data used to train the predictor must be recorded with time information, which is the basis for the prognosis-oriented prediction task. In the case of a bearing fault, the predictor could take the fault dimensions, failure rates, trending information, temperature, component ID, etc. as its inputs and generate the fault growth as the output. Feature extraction can be performed periodically for the processes under prognosis. It should be noted that features are extracted in temporal series and are dynamic in the sense that the DWNN processes them in a dynamic fashion. Then, the obtained features are fused into the time-dependent feature vector that characterizes the process at the designated time instants. Feature selection is based on criteria that distinguish a fault signature from normal operating conditions and one particular fault mode from another. Such other criteria as computational cost may be included.

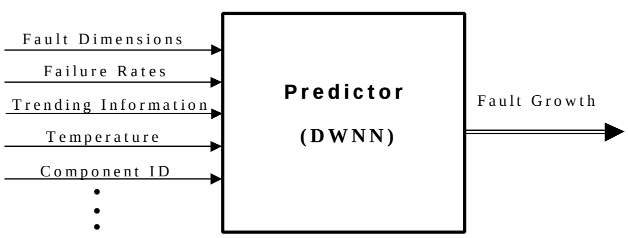

Figure 4 A schematic representation of the DWNN as the predictor

The DWNN must be trained and validated before any on-line implementation and use. Such algorithms as the Back-Propagation or Genetic Algorithm can be used to train the network. Once trained, the DWNN, along with the TTF calculation mechanism, can act as an on-line prognostic operator. It is worth reiterating that the results from the diagnosis serve as the input to the prognosis. Thus, the fidelity and accuracy of the diagnostician bears a direct impact on the reliability of the prognosticator. Predictions can be substantially improved as more fault data become available. The diagnostic/prognostic operation is viewed, therefore, as a dynamic, “evolving” mechanism with adaptive observation and prediction windows; more accurate predictions resulting from the utility of additional historical information. The DWNN is, indeed, updated on-line in a real-time fashion.

Uncertainty Management

The basic features of the proposed prognostic architecture will be illustrated via an application example. The case at hand refers to a rolling-elements bearing failure. Such components are common in industrial equipment and their failure may result in severe damage of critical processes. Micro-cracks may grow in size over time as local and other operating conditions stress the constituent elements of the bearing. Uncertainty and ambiguity are the rule rather that the exception in the diagnosis and prognosis of failure modes in such systems. They manifest themselves at various levels of abstraction: at the data level, the feature level, the decision level and classification levels. As the prediction window increases, so does the uncertainty resulting from the levels of the data processing hierarchy. There are many potential root causes of uncertainty associated with fault conditions: Faults exhibit varying signatures depending upon the location, cause, prevailing operating conditions and the state of the component materials. Detection and identification at an early stage of an incipient failure mode requires reliable and robust techniques for accurate declaration without false alarms. Prediction of the future behavior of a fault is much more demanding – essentially taxing severely the available means to quantify uncertainty. Prediction algorithms, therefore, must incorporate possibilistic (or probabilistic) quantifiers that inform the user of the expected time-to-failure as well as its anticipated variance (in terms of the earliest and latest time estimates). Fuzzy notions, such as fuzzy membership functions, are known to capture well uncertainty estimates and Dempster-Shafer theory may prove useful in combining conflicting evidence and supporting upper and lower bounds (plausibility and belief metrics) in these estimates.

Difficulties in uncertainty management for fault prognosis are due to the fact that fault prognosis involves subjective as well as objective uncertainties and operates over the time horizon from the past, through the present and to the future. Thus, two essential investigation steps are deemed to be necessary: identifying uncertainty sources and devising uncertainty management schemes. For a process fault prognosis task, uncertainty sources can be broken down into four types: uncertainties in the historic data, uncertainties in the prognostic method, uncertainties in the process itself, and uncertainties in the operator or designer’s opinion. Correspondingly, these uncertainties are named (a) data uncertainties, (b) process uncertainties, (c) method uncertainties, and (d) designer uncertainties. For process fault prognosis, identification of these uncertainty sources allows us to target different mathematical tools at different types of uncertainties. Normally, there are two mathematical tools that can be used for the uncertainty management problem: probability theory and possibility theory.

For simplicity, this paper deals only with data uncertainties and uses uncertainty boundaries for reporting prognostic results. This results in the so-called interval prediction, compared to point predictions. An uncertainty interval can readily be generated by estimation of a lower and an upper bound of the prediction window. As shown in Figure 5, a fault indicated by the feature F(t) would evolve along its mean $F_M(t)$ and within its lower bound $F_L(t)$ and upper bound $F_U(t)$. Thus, the prognostic result can be reported as that the remaining useful lifetime has a mean T_M and bounded in [T_L T_U], due to data uncertainties.

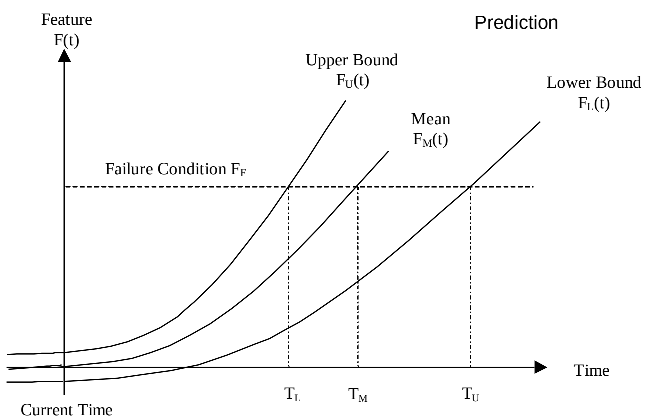

Figure 5 Uncertainty boundaries in a prognostic task

An Illustrative Example

Industrial chillers are typical processes found in many critical applications. These devices support electronics, communications, etc. on a navy ship, computing and communication in commercial enterprises, refrigeration and other functions in food processing, etc. Of special interest is the fact that their design incorporates a diverse assemblage of common and vital components, i.e. pumps, motors, compressors, etc. A rich variety of failure modes are observed on such equipment ranging from vibration-induced faults to electrical failures and a multitude of process-related failure events. Most chillers are well instrumented monitoring vibrations, temperature, pressure, flow, etc., and many mechanical faults exhibit symptoms that are sensed via vibration measurements. For example, a water pump will vibrate if its motor bearing is defective, if its shaft is misaligned or if its mounting is somewhat loose. A rolling-element bearing fault is used in this study to demonstrate the feasibility of the prognostic algorithms.

Defective bearings or loose mounting bolts would cause a pump to vibrate abnormally. The vibrations are normally monitored by an accelerometer. The measured signals are transferred to a data acquisition unit via a high-quality co-axial cable. Tri-axial vibration signals originating from a bearing with a crack in its inner race have been collected [17]. An initial crack was seeded in the bearing and the experiment was run for a period of time and vibration data were recorded during that period. The set-up was then stopped and the crack size was increased followed by a second run. This procedure was repeated until the bearing failed. The crack sizes were organized in an ascending order while time information was assumed uniformly distributed among the crack sizes. A training data set relating to the crack growth was thus obtained. Time segments of vibration signals from a good bearing and a defective one are shown in Figure 6. Their corresponding power spectral densities (PSD) are shown in Figure 7. The original signals were windowed with each window containing 1000 time points. The maximum values of the vibration signals in each window were also recorded as shown in Figure 8. The PSDs of the windowed vibration signals were calculated and their peak values extracted as depicted in Figure 9. Figure 10 shows the corresponding crack sizes. Crack size information at intermediate points was generated via interpolation to avoid a large number of repeated experiments. There are 100 data points for each curve in the figures. The features chosen for prognosis were the maximum signal values and the maximum signal PSDs for all three axes, i.e., (MaxSx MaxSy MaxSz) and (MaxPSDx MaxPSDy MaxPSDz).

Figure 10 demonstrates the crack growth as a function of time. The model is first trained using fault data up to the 100^{th} time window. from then on, it predicts the crack evolution until the final bearing failure. The virtual sensor, implemented as a WNN with seven hidden nodes or neurons is trained through the process of Figure 11. This virtual sensor "measures" the crack size on the basis of the maximum signal amplitude and the maximum signal PSDs as inputs. The training results are depicted in Figure 12. It is observed that 100 data points employed for training lead to very satisfactory results. The DWNN, acting as the predictor, is trained next. The optimized training procedure results in a DWNN of eight hidden neurons. The training results are shown in Figure 14. Training is deemed satisfactory when 100 data points are used. The trained predictor is employed next to predict the future crack development, as shown in Figure 15. A failure hazard threshold was established on the basis of empirical evidence corresponding to Crack_Width = 2000 microns or Crack_Depth = 1000 microns. The crack reaches this hazard condition at the 174^{th} time window. The Crack_Width criterion is reached first. These results are preliminary and intended only to illustrate the proposed prognostic architecture. A substantially large data base is required for feature extraction, training, validation and optimization. Such a data base will permit a series of sensitivity studies that may lead to more conclusive results as to the capabilities and the effectiveness of the proposed approach.

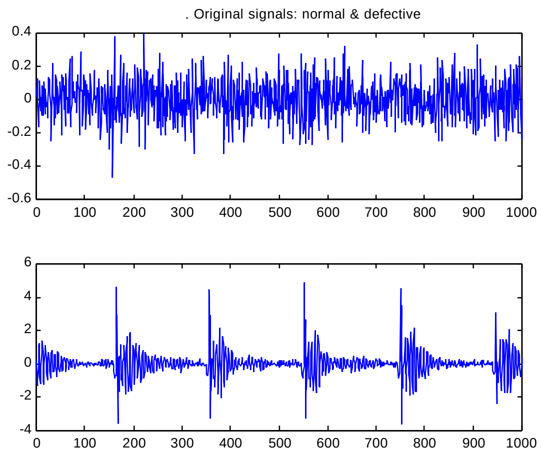

Figure 6 Vibration Signals from a good and a defective bearing

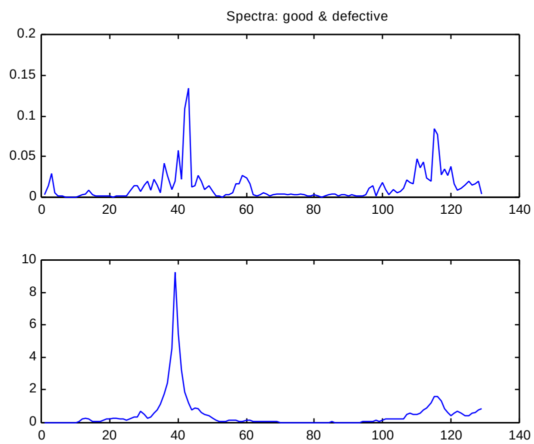

Figure 7 PSDs of the vibration signals in Figure 6

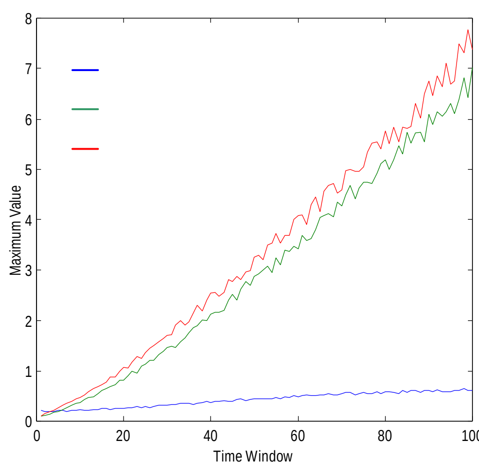

Figure 8 The peak values of the original signal

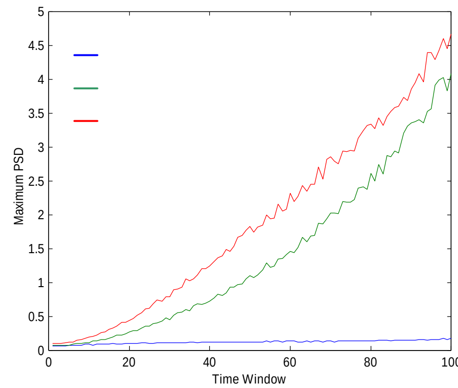

Figure 9 The maximum PSDs of the original signals

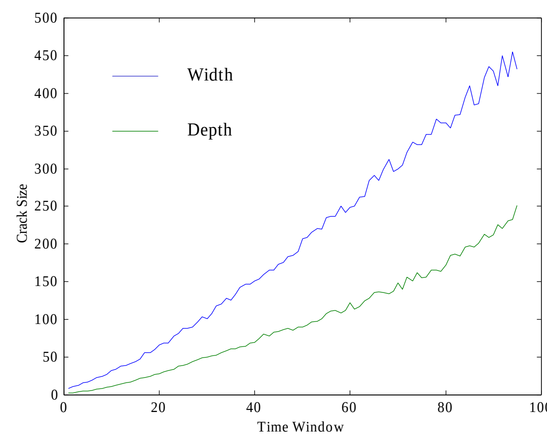

Figure 10 The original crack sizes

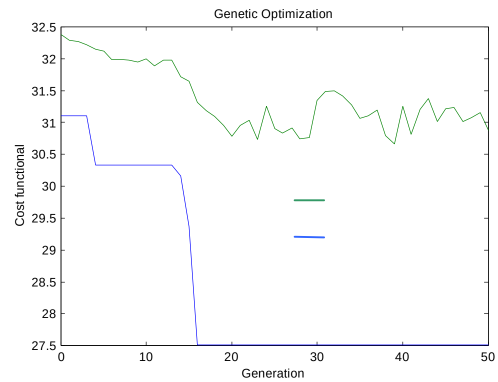

Figure 11 The training of the virtual sensor

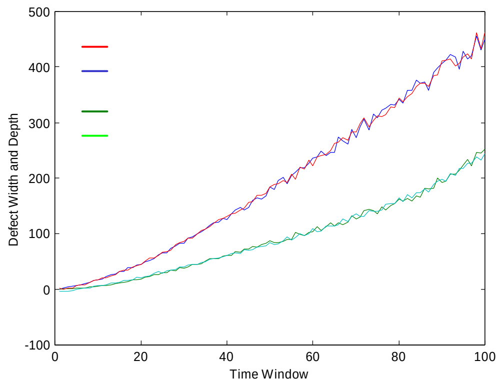

Figure 12 The crack sizes measured by the trained virtual sensor

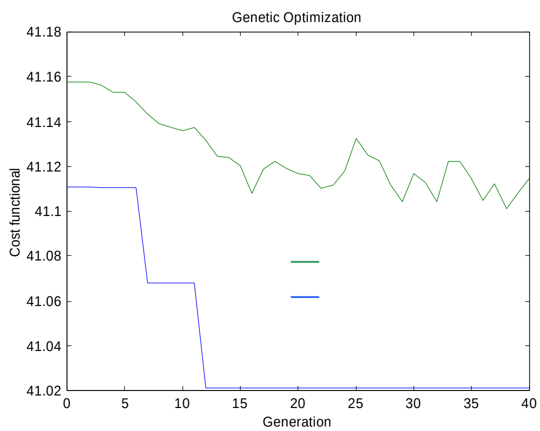

Figure 13 The training of the predictor

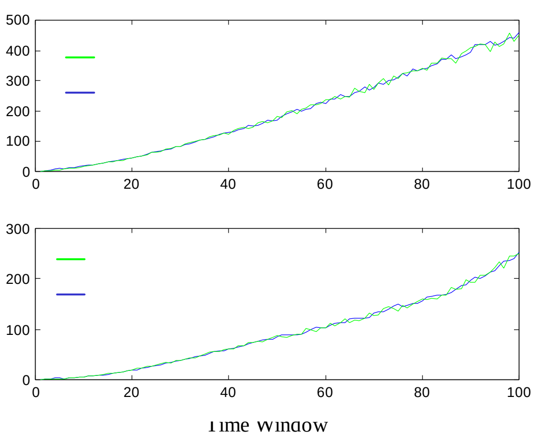

Figure 14 The crack growth predicted by the trained predictor within 100th time window

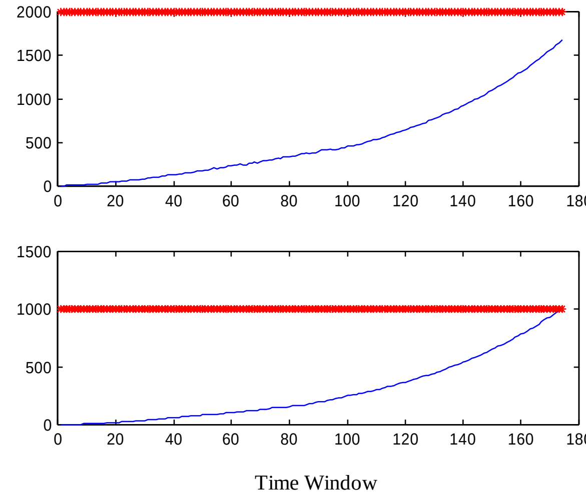

Figure 15 The crack growth predicted by the trained predictor beyond 100th time window

Conclusions

A fault prognosis architecture consisting of a virtual sensor and a dynamic wavelet neural network has been developed. The proposed model addresses two challenging issues relating to prognosis of machine or component failures: How do we "measures" the growth of a fault and how do we predict the remaining useful lifetime of such a failing component or machine? Reliable answers to these questions are bound to assist maintenance personnel in the conduct of condition-based maintenance so that uptime is maximized and the useful life of critical assets is prolonged. Simulation studies of the virtual sensor – predictor configuration, based on a limited experimental data set, show promise. More extensive failure data – difficult to obtain in critical processes – are required to draw firm and comparative conclusions. The proposed architecture provides a generic and open platform that can be easily modified and augmented as new failure evidence becomes available. The WNN construct (in both the static and dynamic versions) is amenable to accommodating learning routines (on-line and off-line) so that the algorithm can be improved with time. Uncertainty, a dominant influence in diagnostics and prognostics, must be accommodated and managed. A neuro-fuzzy version of the basic WNN and DWNN can assist in this direction when coupled with notions from Dempster-Shafer theory. This paper, therefore, serves as a motivation to encourage further research in those challenges areas of data collection and management, modeling, validation and verification, implementation and assessment that are crucial to a successful penetration of these technologies in the industrial and manufacturing sectors of our economy.

Acknowledgement

This research was partially supported by Honeywell Inc. and the Office of Naval Research under contract No. E-21-K30. Their support and assistance in the conduct of this work is greatly appreciated.

References

[1] Konrad, H. and Isermann, R., “Diagnosis of different faults in milling using drive signals and process models”, Proceedings of the 13th World Congress, International Federation of Automatic Control, Vol.B., Manufacturing, p.91-6, 30 June-5 July 1996.

[2] Mylaraswamy, D., and Venkatasubramanian, V., “A hybrid framework for large scale process fault diagnosis”, Computers & Chemical Engineering, vol.21, suppl. issue, p. S935-40, 25-29 May 1997.

[3] Barbera, F., Schneider, H., and Kelle, P., "A condition based maintenance model with exponential failures and fixed inspection intervals", Journal of the Operational Research Society, vol.47, no.8, p.1037-45, 1996.

[4] Makis, V., Jiang, X., and Jardine, A.K.S., "A condition-based maintenance model", IMA Journal of mathematics Applied in Business and Industry, vol.9, no.2, pp.201-210, 1998.

[5] Prickett, P. and Eavery, S., "The case for condition based maintenance", Integrated Manufacturing Systems, vol.2, no.3, pp.19-24, 1991.

[6] Taylor, N.B., Stedman’s Medical Dictionary, 18th Edition, The Williams & Wilkins Company, Baltimore, 1953.

[7] Jardim-Goncalves, R., Martins-Barata, M., Alvaro Assis-Lopes, J., and Steiger-Garcao, “Application of stochastic modelling to support predictive maintenance for industrial environments”, Proceedings of 1996 IEEE International Conference on Systems, Man and Cybernetics, Information Intelligence and Systems, p.117-22, vol.1, 14-17 Oct. 1996.

[8] Frelicot, C., “A fuzzy-based prognostic adaptive system”, RAIRO-APII-JESA, Journal Europeen des Systemes Automatises, vol.30, no.2-3, p.281-99, 1996.

[9] Lembessis, E., Antonopoulos, G., King, R.E., Halatsis, C., and Torres, J., “'CASSANDRA': an on-line expert system for fault prognosis”, Proceedings of the 5th CIM Europe Conference on Computer Integrated Manufacturing, p.371-7, 17-19 May 1989.

[10] Ray, A. and Tangirala, S., “Stochastic modeling of fatigue crack dynamics for on-line failure prognostics”, IEEE Transactions on Control Systems Technology, vol.4, no.4, p.443-51, July 1996.

[11] Parker, B.E., Jr., Nigro, T.M., Carley, M.P., Barron, R.L., Ward, D.G., Poor, H.V., Rock, D., and DuBois, T.A., “Helicopter gearbox diagnostics and prognostics using vibration signature analysis”, Proceedings of the SPIE - The International Society for Optical Engineering, vol.1965, p.531-42, 1993.

[12] Schauz,J.R., Wavelet Neural Networks for EEG Modeling and Classification, PhD Thesis, Georgia Institute of Technology, 1996.

[13] Tsui, F.C., Sun, M.G., Li, C.C., Sclabassi, R.J., “A wavelet based neural network for prediction of ICP signal” Proceedings of 1995 IEEE Engineering in Medicine and Biology 17th Annual Conference and 21 Canadian Medical and Biological Engineering, p.1045-6, vol.2, 20-23 Sept. 1995.

[14] He, S.C. and He, Z.Y., “Blind equalization of nonlinear communication channels using recurrent wavelet neural networks”, 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, p.3305-8, vol.4, 21-24 April 1997.

[15] Rao, S.S. and Kumthekar, B., “Recurrent wavelet networks”, Proceedings of the 1994 IEEE International Conference on Neural Networks, p 3143-3147, Orlando, FL, USA, 1994.

[16] Marko, K.A., James, J.V., Feldkamp, T.M., Puskorius, C.V., Feldkamp, J.A., and Roller, D., “Applications of neural networks to the construction of "virtual" sensors and model-based diagnostics”, Proceedings of ISATA 29th International Symposium on Automotive Technology and Automation, p.133-8, 3-6 June 1996.

[17] Shiroishi, J., Li, Y., Liang, S., Kurfess, T., and Danyluk, S., “Bearing condition diagnostics via vibration and acoustic emission measurements”, Mechanical Systems and Signal Processing, vol.11, no.5, p.693-705, Sept. 1997.