FAULT DIAGNOSIS AND PROGNOSIS IN MECHANICAL SYSTEM THROUGH WAVELET ARTIFICIAL NEURAL NETWORKS

Alexandre Carlos Eduardo, Prof. Dr.
Universidade Federal de Minas Gerais - UFMG
Departamento de Engenharia Mecânica
Laboratório de Acústica e Vibrações
CEP 31270-901 – Belo Horizonte – MG
aceduard2003@yahoo.com.br

Abstract. *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. This paper attempts to address this challenging problem with intelligence-oriented techniques, specifically hybrid method: static and dynamic wavelet neural networks. Cast as parallel processing structures and equipped with adaptive learning mechanisms, wavelet neural networks are capable of imitating human operators or experts to perform fault diagnostic tasks. 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 static and a dynamic wavelet neural network successfully diagnose and prognose a defective rotating system bearing with a crack in its inner race.*

Keywords: *Fault Diagnosis, Prognosis, Mechanical System, Wavelet Neural Network.*

1. 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 downtime, resulting in loss of productivity. Machine diagnostics/prognostics for condition-based maintenance involves an integrated system architecture with a diagnostic module – the diagnostic – 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 diagnostic 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 diagnostic and the prognostic module 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 (Konrad et al, 1996; Mylaraswamy and Venkatasubramanian, 1997). 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. 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 (Jardim-Goncalves et al, 1996), fuzzy pattern recognition principles (Frelicot, 1996), knowledge-intensive expert systems (Lembessis et al, 1989), nonlinear stochastic models of fatigue crack dynamics (Tangirala , 1996), polynomial neural networks (Parker et al, 1993) and other techniques have been introduced over the past years to address the diagnostic/prognostic problem. This paper attempts to address this diagnostic and prognostic issue by introducing a novel combination of a “virtual” sensor as a mapping tool between known measurements and “difficult-to-access” quantities, a static wavelet neural network as the “classifier”, i.e. the mechanism that classifies various fault modes of a monitored component, 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.

2. THE DIAGNOSTIC AND PROGNOSTIC SYSTEM ARCHITECTURE

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. 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 diagnostic. Thus, this scheme should reduce the uncertainty and improve the prediction accuracy as the accumulated evidence grows

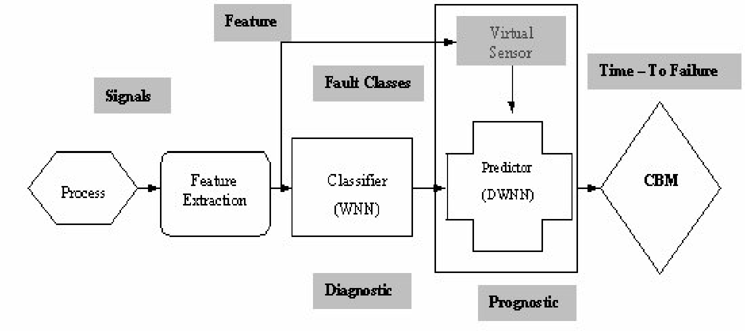


Figure 1 The overall architecture of the prognostic system

Figure 1 depicts the overall architecture of the proposed diagnostic and prognostic system. The diagnostic 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 diagnostic 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).

3. THE STATIC AND DYNAMIC 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.

A Wavelet Neural Network (WNN) can be formulated as:

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

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:

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

where x is the input row-vector; A the squashing matrix for the wavelet; b the translation vector; and T the transpose operator.

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.

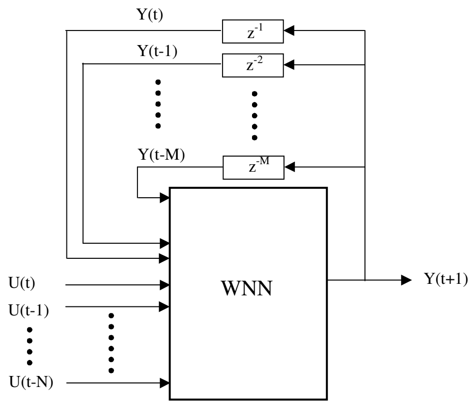


Figure 2 A dynamic wavelet neural network.

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:

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

4. THE DIAGNOSTIC

The core of the diagnostic is its WNN that serves as a nonlinear discriminator to classify impending faults. The diagnostic may also include a decision logic module that produces detection or identification outcomes from the output of the WNN. It is possible, however, to incorporate the decision logic module in the WNN network structure. A feature extractor assisted by a signal pre-processor and a feature database essentially provides the discriminator with refined or transformed information.

4.1 The Feature Extractor

The objective of this module is to determine and extract appropriate features for the fault or defect classification task. An additional objective is to reduce the search space and to speed up the computation. For example, a windowing operation can be applied to the 1-D signals in order to reduce the search space and facilitate the selection of appropriate features. Such features as the width and the height of the widowed signal could provide enough information to distinguish among fault types. To further distinguish among faults in each category, however, it is advisable to select more features in the time domain, the frequency domain, etc. For example, energy, area, geometric center, periodicity, peaks in the frequency spectrum, etc. Finally, a feature vector is defined as follows:

$$F(i) = [H_1(i)\, W_1(i)\, H_2(i)\, W_2(i)\, S_1(i)\, S_2(i)\, A_1(i)\, A_2(i) \quad \cdots\cdots] \tag{4}$$

where i is the signal index; F(i) the feature vector of the ith signal; H_1(i) the height of the first channel of the ith signal; W_1(i) the width of the first channel of the ith signal; H_2(i) the height of the second channel of the ith signal; and W_2(i) the width of the second channel of the ith signal; S_1(i) the bandwidth of the first channel of the ith signal; and S_2(i) the bandwidth of the second channel of the ith signal; A_1(i) the area of the first channel of the ith signal; and A_2(i) the area of the second channel of the ith signal; and so on. This feature vector may be tailored to accommodate different identification tasks.

4.2 The Classifier

The WNN is used as the classifier. Potential advantages of the WNN approach include: The resulting neural network is a universal approximator; the time - frequency localization property of wavelets leads to reduced networks at a given level of performance; WNNs offer a good compromise between robust implementations and efficient functional representations; the multi-resolution organization of wavelets provides a heuristic for neural network growth. The WNN is trained, thus, as a two-step process: the structure and the parameters of the network are determined iteratively until a performance metric is satisfied.

5.1 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 (Marko et al, 1996), 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.

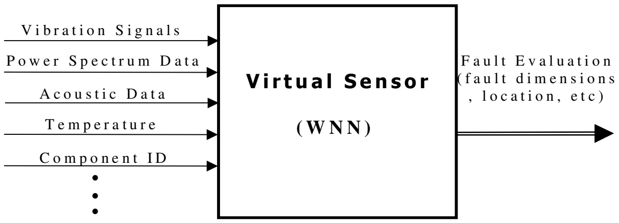


Figure 3 A schematic representation of the WNN as a virtual sensor

5.2 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.

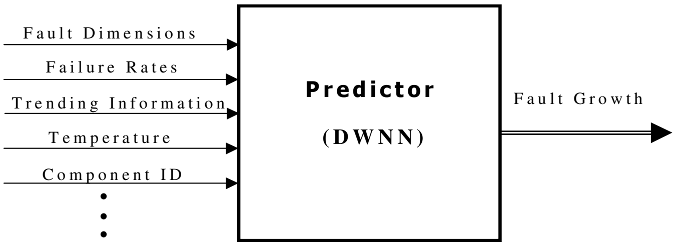


Figure 4 A schematic representation of the DWNN as the predictor

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.

5.3 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. 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 [TL T_U], due to data uncertainties.

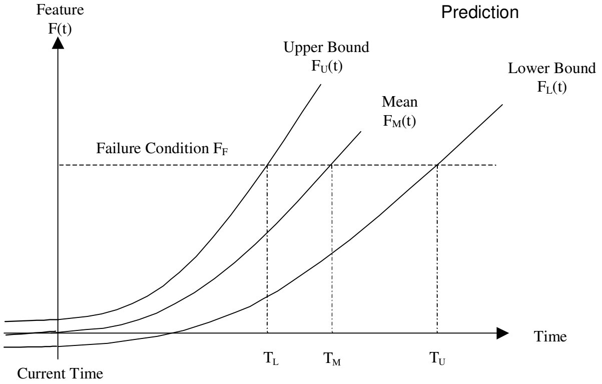


Figure 5 Uncertainty boundaries in a prognostic task

6. AN ILLUSTRATIVE EXAMPLE

Considerer a 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. 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. Training data sets relating to the crack sizes and to their 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.

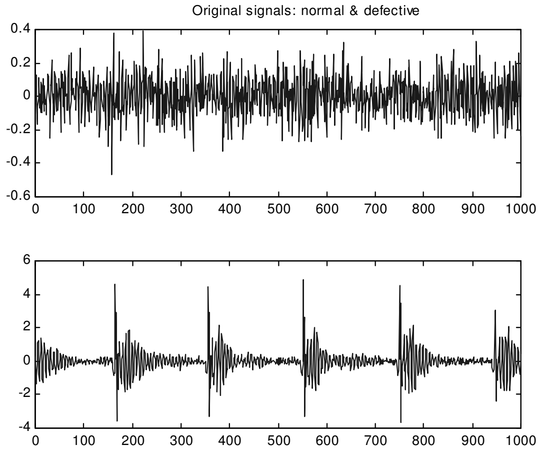


Figure 6 Vibration Signals from a good and a defective bearing

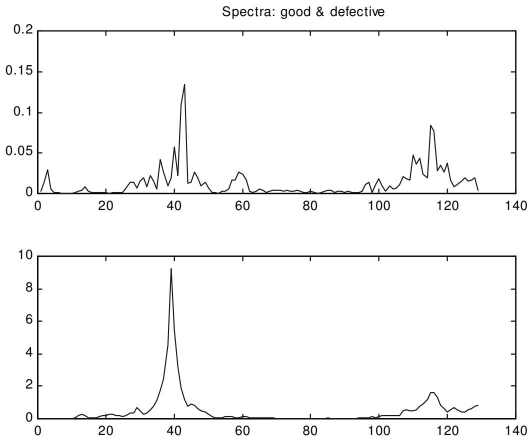
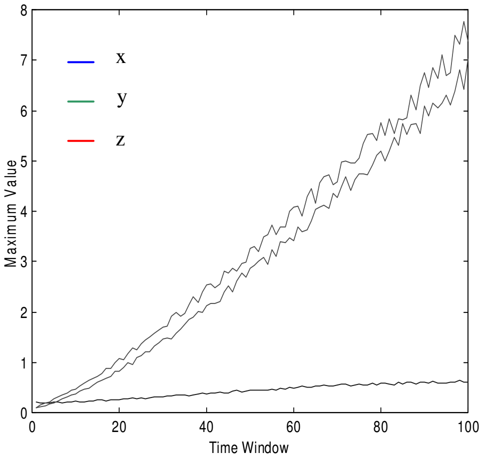


Figure 7 PSDs of the vibration signals in Figure 6

The task of bearing diagnosis is to examine if there are some faults on the bearing and what modes the faults are. Many bearing faults can be classified, such as those on races, rolling balls and lubrication materials. To detect or identify the bearing faults such as that indicated by Figure 6 and 7, the diagnostic was called in to distinguish between the vibration signal from the good bearing and that from the defective bearing. The peak of the signal amplitude and the peak of the signal PSD were chosen as the features.

Thus, the feature vector is [MaxS MaxPSD]. Assigning these feature values with appropriate fault classes, a training data set for diagnosis was obtained. The two classes for bearing faults are displayed in Figure 8. The diagnostic system proposed was trained using the data set, as shown in Figure 9. The trained diagnostic system is employed to perform the diagnosis. New signals from a normal bearing and a defective one are those shown in Figure 6. Their PSD are shown in Figure 7 in correspondence.



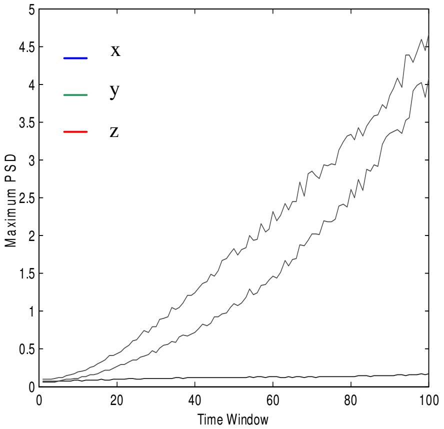
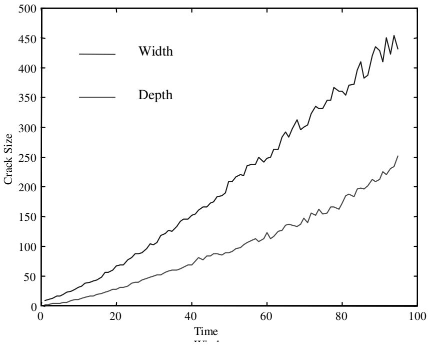


Figure 8. The peak values of the original signals

Figure 9. The maximum PSDs of the original signals

This result can easily be extended to include the cases in which multiple faults classes are concerned.

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 10. The PSDs of the windowed vibration signals were calculated and their peak values extracted as depicted in Figure 11.



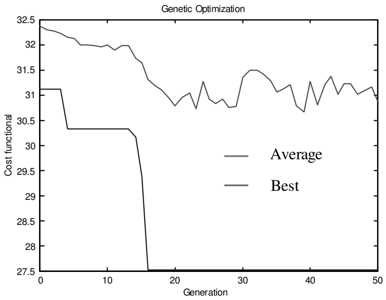


Figure 10. The original crack sizes

Figure 11. The crack sizes measured by the trained virtual sensor

Figure 12 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 12 demonstrates the crack growth as a function of time. The model is first trained using fault data up to the 100th 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 13.

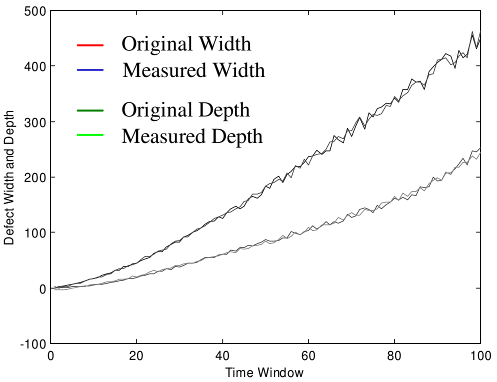


Figure 12. The crack sizes measured by the trained virtual sensor

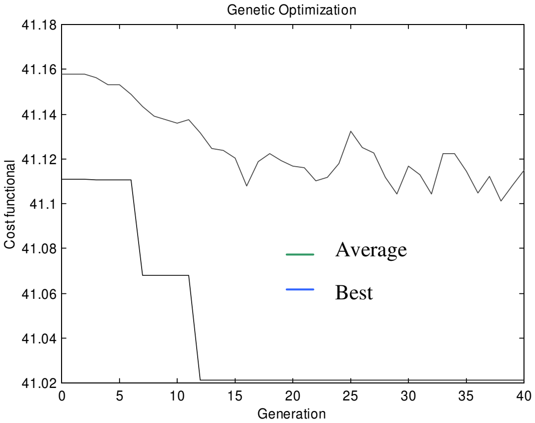
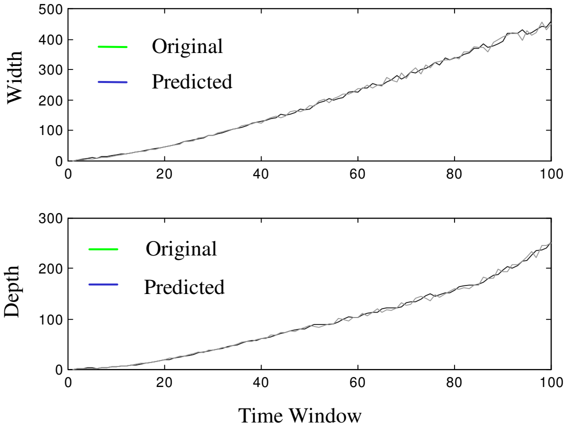


Figure 13 The training of the predictor

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 14. 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 15.



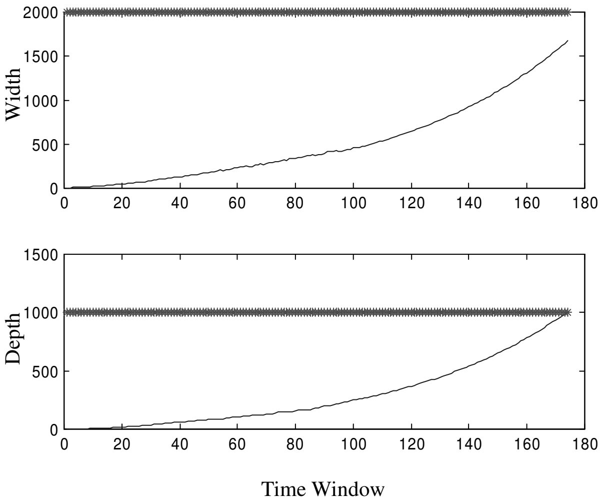


Figure 14 The crack growth predicted by the trained predictor within 100th time window

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

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 174th 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.

7. CONCLUSIONS

A fault diagnosis and prognosis architecture consisting of a static wavelet neural network, a virtual sensor and a dynamic wavelet neural network has been developed. The proposed model demonstrates its effectiveness and efficiency in diagnosis of machine or component faults. More importantly, 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. 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.

REFERENCES

Barbera, F., Schneider, H., and Kelle, P., 1996. "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.

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

Jardim-Goncalves, R., Martins-Barata, M., Alvaro Assis-Lopes, J., and Steiger-Garcao, 1996 “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 .

Konrad, H. and Isermann, R., 1996. “Diagnosis of different faults in milling using drive signals and process models”, Proceedings of the 13th IFAC World Congress, Vol.B., Manufacturing, p.91-6.

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

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

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

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., 1993. “Helicopter gearbox diagnostics and prognostics using vibration signature analysis”, Proceedings of the SPIE - The International Society for Optical Engineering, vol.1965, p.531-42.