Bearing Fault Diagnosis based on Neural Network Classification and Wavelet Transform

OMAR JOSÉ LARA CASTRO
CRISTINA CASTEJÓN SISAMÓN
JUAN CARLOS GARCÍA PRADA
Mechanical Engineering Department
Universidad Carlos III de Madrid
Avda. de la Universidad nº 30. 28911. Leganés. Madrid
SPAIN

Abstract: - Automated fault classification has been an important pattern recognition problem for decades. In the performance of all motor driven systems, bearings play an important role. The purpose of condition monitoring and fault diagnostics are to detect and distinguish faults occurring in machinery, in order to provide a significant improvement in plant economy, reduce operational and maintenance costs and improve the level of safety. This paper addresses a recent study to employ Wavelet decomposition to process the accelerometer signals and identify band patterns features. Selected features are extracted from the vibration signatures so obtained and these are used as inputs to three types of artificial networks trained to identify the bearing conditions at three different rotational speeds. Vibration signals for normal bearings, bearing with inner race fault, outer race faults and ball faults were acquired from a motor-driven experimental system. The experimental results are presented and compared with those of currently best-performing on this field. Later sections explain some of the artificial intelligence methods design considerations such as network architecture, performance and implementation.The results demonstrate that the developed diagnostic method can reliably detect and classify four different bearing fault conditions into distinct groups.

Key-Words: - Wavelets, Artificial networks, Fault diagnosis, Signal processing, Pattern classification

1 Introduction

Most of machinery used in the modern world operates by means of motors and rotary parts which can develop faults. The monitoring of the operative conditions of a rotary machine provides a great economic improvement by reducing the operational and maintenance costs, as well as improving the safety level. As a part of the machine maintenance task, it is necessary to analyze the external relevant information in order to evaluate the internal components state which, generally, are inaccessible without dismantle the machine.

From the beginning of the 20th century, different technologies have been used in order to study the general dynamical behavior of machines and mechanisms elements. From Fourier's classic analysis [1] and its variations [2], which gave a great step in this field, up today, where the Wavelet analysis [3] plays the main role of a new time - frequency evaluation method for the failures detection in rotary parts. The recent history of the Wavelet analysis has the origin in a narrow collaboration among several scientists from different areas in the middle of last century, and it consists of the next logical step of the processing signal evolution. The main Wavelet characteristic is the skill of evaluating the transitory signals in different scales and resolutions.

There are two important phases to implement in the rotary parts faults detection process: the first consists of performing a signal processing, in order to extract the typical features or "patterns" and diminish the signal noise, and the second phase consists of the signal classification based on the characteristics obtained in the previous step; these tasks are not direct for every type of signal, since, to manage them with relative success, some knowledge and experience in the field is needed.

Faults detection in mechanical rotary parts, requires increasing the current efficiency and performance. For this reason, a real-time on line monitoring of the machine in study, and a faults automatic detection would be a great advance in this field. For this type of tasks, it is necessary to use more reliable and faster mathematical methods.

The application of the Wavelet Transform (WT), for faults diagnosis in machines, has been widely developed in the last 15 years. In 1990, Leducq [4] used the WT to analyze the hydraulic noise of a centrifugal pump. This was, probably, the first work related to Wavelet application for the failures diagnosis. In addition, the majority of WT applications have been performed on gears. In 1993, Wang and McFadden [5] were the pioneers in the application of the WT to the vibration signal analysis in gears, results showed that Wavelet is able to detect both mechanical incipient failures and different types of faults simultaneously.

In 1999, Newland [6] introduced the WT in the engineering field, with several calculation methods and examples of his application in the vibratory signal analysis; also, the author proposed a Harmonic Wavelet and identified successfully peaks and phases of transitory signals. Newland's work made the Wavelet popular in engineering applications, especially for vibration analysis. Wang et al. [7] experimentally investigated the sensitivity and robustness of the currently well-accepted techniques for gear damage monitoring, and the results shows that the WT is reliable technique for gear health condition monitoring, which is more robust than other methods.

Liu et al. [8] proposed a method based on Wavelet packets, for the diagnosis of failures in ball bearings. The coefficients of the Wavelet packets were used as typical features, and the results show that they possess a high sensibility to failures. Aminian et al. [9] developed a system of failures diagnosis based on a Bayessian neural network which was using the WT, normalization and main components of the signal as pre-processors. Loparo et al. [1], used a fuzzy classifier to diagnose faults in bearings, based on the WT, achieving to generate typical vectors, with the standard deviation of the DWT coefficients. Farías et al. [10] analyzed the effects that the DWT application has, as a processing signal tool in the performance of the images classification and they determined the WT parameters which present the best results for the workspace. Peng et al. [11] carried out a bibliographical review about the WT application in the monitoring and the fault diagnosis in machines.

The evident uncertainty for knowing which is the type of Wavelet most adapted to process signals is object of many studies, among them, Liu et al. [12] applied a new methodology to the failures study of a Diesel engine. They obtained a value of mutual information that was used as indicator of the degree of relative information to the fault that a Wavelet possesses. This methodology is an extension of the Mallat's search correspondence technology [2]. Goumas et al. [13], carried out a study for the faults detection in washing machines, based on the DWT and faults statistical classifiers as standard deviation and Bayes's theory, major rates of success were 87 %. In summary, many kinds of fault features can be obtained, principally with the Wavelet coefficients or the Wavelet energy. Since the Wavelet coefficients will highlight the changes in signals, which often predicate the occurrence of the fault, the Wavelet coefficients-based features are suitable for fault detection. However, because the slight changes in signals often have small energy these changes will be easy masked in the Wavelet energy-based features. Therefore, the Wavelet energy-based features are often not able to detect the early faults.

2 Experimental Setup

Experiments were performed on the single ball bearing FAG 7206 B, having 13 rolling elements, and contact angle of 40º. An image of the test lab bench and its accessories is presented in Fig. 1; from the right hand side, the following elements are visible: axial and radial pneumatic pistons, assembly which contains the analyzed bearing, accelerometer Brüel & Kjaer 4383 with bandwidth of 8.5 KHz, two single row taper roller bearings SKF 32007X which acts as a support for the machine, photo tachometer device to measure the RPM's, and a transmission pulley directly connected to the motor by a V-belt.

The acquisition system consists of an amplifier NEXUS of B&K, 140 KHz. bandwidth, an acquisition card Keithley DAS-1200 with a maximum sampling frequency of 1 MHz. The sampling rate is 5000 Hz, each acquired signal has a length of 5120 points.

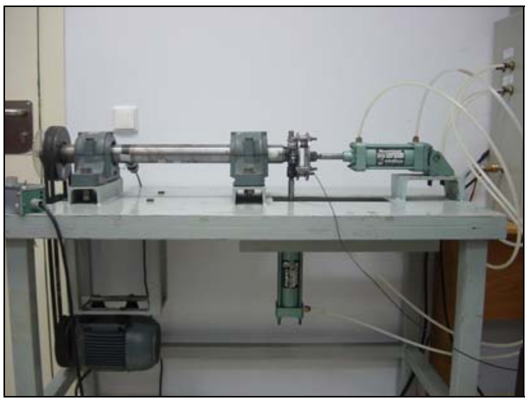

Fig. 1: Bearing Test lab bench.

2.1 Description of the Tests

Each tested bearing was trying at 600, 1200 and 1800 RPM, and four sets of data were obtained from the experimental system: (i) under normal conditions; (ii) inner race faults; (iii) outer race faults and (iv) ball fault. A pit of 2 mm. long was artificially produced by an electric pen to simulate the damage. In the rolling ball case, multiple slots in the surface were performed, to simulate the flacking phenomenon. All cases are shown in the figure 2.

a b c

Fig. 2: Flacking in bearing parts (a) Inner race, (b) Outer race and (c) Ball.

3 Data Processing

By examining the magnitude of the vibration data under operating conditions with incipient bearing faults, it is possible to distinguish the normal data from different types of fault data (see Fig. 3). However, this is not always applicable, because the signal morphology that results from a fault changes over time as the fault progresses from initiation to failure. Thus, some faults will be undetectable until failure is imminent.

Because the early detection and isolation of faults is important for condition-based maintenance, a more sophisticated signal processing approach is necessary. The first step in our approach is to process the test data performing the Wavelet analysis.

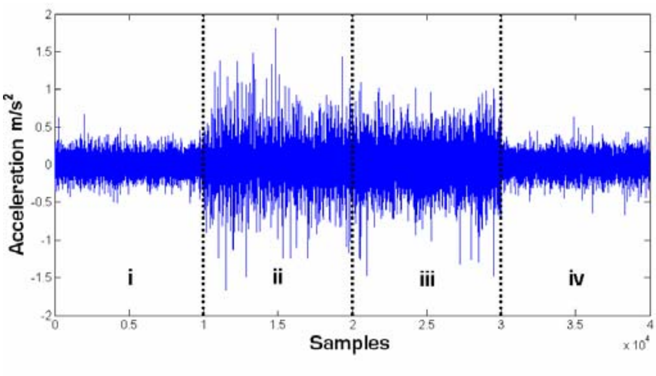

Fig. 3. Acquired signals of each condition at 600 RPM.(i) Normal conditions; (ii) inner race faults; (iii) outer race faults and (iv) ball fault.

3.1 Brief review of Wavelet Theory

The Fourier's transform changes a signal in the time domain to a frequency domain, using trigonometric functions as basis functions. On the other hand, the Wavelet Transform (WT) changes a signal in the space domain to the scale domain by means of basis functions of finite energy, called Wavelets [2].

In general, the WT is a mathematical function which multiplies the signal during all its length, with elongated and compressed versions of a mother Wavelet as is shown in Figure 4. In this way, the coefficients of the Wavelet basis functions are generated. WT cannot be evaluated in a practical way using analytical equations, integrals, etc. Therefore, a WT discretization is needed, and MRA (Multiresolution Analysis) is the most suitable form to perform it. MRA consists of apply recursive filters to calculate the Wavelet decomposition [3].

Given f [n], the signal to be analyzed, a discrete function, the DWT from this signal is given in Eq. (1):

$$c(j,k) = \sum_{Z} f[n]\psi_{j,k}[n] \quad (j,k) \in \mathbb{Z}^2 \tag{1}$$

Parameters τ and s are defined according to the dyadic scale, as appears in Eq. (2), where j represents the decomposition level and k is the translation factor.

$$\tau = 2^j \,; s = k2^j \quad \tau > 0;\ s > 1 \quad (j,k) \in \mathbb{Z}^2 \tag{2}$$

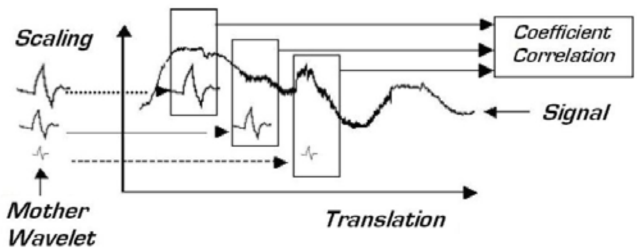

Fig. 4. Illustration of the Wavelet Transform execution.

The filtering process of the signal f[n], resulting from the Wavelet decomposition, begins by applying to the signal a half band low pass filter, which possesses an impulse response h[n]. The filter fits a mathematical convolution operation with the signal and the filter impulse response. Low pass filter eliminates all the frequency components over the half of the maximum signal frequency; for example, if the signal f[n] possesses maximum frequency components over 2500 Hz, the low band filter will eliminate all the frequency components in the 1250-2500 Hz band and will give the 1st detail coefficients (cD1).

The DWT analyzes the signal in different frequency bands, but with different resolutions, having separated the signal in two different data: approximation information and detail information. Two functional groups are used to represent the information, they are called scaled functions (ϕ) and Wavelet functions (ψ), which are associated with the low band and high band filters respectively. After filtering, the same number of samples will be achieved, but in a half of frequency band. Therefore, and applying to the Nyquist rule [3], half of the samples is justified to be eliminate without loosing information. This procedure constitutes the first level of decomposition (level 1) and it is possible to be expressed mathematically in the followings equations (3, 4).

$$y_{high}[k] = \sum_n x[n]\, g[2k-n] \qquad (3)$$

$$y_{low}[k] = \sum_n x[n]\, h[2k-n] \qquad (4)$$

yhigh and ylow are the low (h) and high (g) band output filter respectively. The term 2k determines the subsampling by a factor of two, and n is the amount of original samples. This procedure, called subband coding, can be repeated for additional decompositions. This process can be represented by the figure 5.

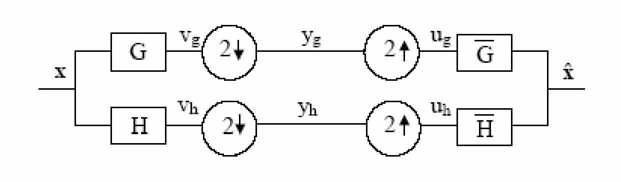

Fig. 5. Wavelet Decomposition by means of two channels

Due to the basis functions used on the WT are compactly supported, Wavelets have good properties of energy concentration. Wavelet coefficients are generally very small, and can have an error without affecting significantly the representation of the signal. Another advantage of the WT is that is able to represent the signal with a limited number of coefficients, which are used as typical faults features.

3.2 Feature Extraction

We began our examination by processing the 25 signals per condition, the Wavelet Daubechies-6 was selected for signal analysis and synthesis. Note that there is not a particular reason to select it, and the results could vary if another one were chosen. Fig. 6 shows the comparison of a signal under normal conditions and ball fault. In these figures, the approximation (a5), and five levels of details (d1-d5) are chosen for each signal. Based on these plots, the inner race fault can be separated from the normal condition because, for example, d3, d4 and d5 are quite different in magnitude between the two conditions, and for a ball fault almost all bands are quite different.

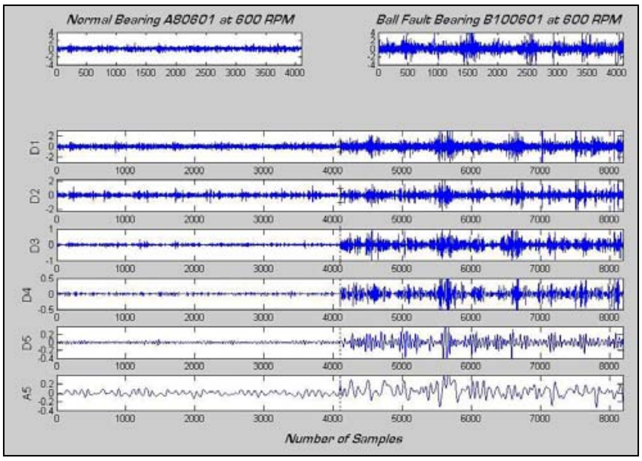

Fig. 5. Wavelet decomposition comparison at fifth level. Normal Bearing (A80601) and a Ball Fault Bearing (B100601) Upper, the original signals; a5-the 5th level approximation (0-78.125 Hz), d1-d5: the five details, d1 (1250-2500 Hz), d2 (625-1250 Hz), d3 (312.5-625 Hz), d4 (156.25-312.5 Hz) and d5 (78.125-156.25 Hz).

3.3 Feature Vector Formation

The more appropriated coefficients provided for the Wavelet decomposition are the 5th level detail coefficients (cD5) for all conditions. Because of there are information in less coefficients (42) and they are visually different. The number of coefficients is given by the decomposition level. The DWT feature vectors are calculated for the test data and post processed by normalizing in the range [0 1] and belong to the frequency band (78.125-156.25 Hz). In general, for each discrete signal f[n] we have a characteristic vector Vc(f[n]) which contains the 5th level detail coefficients (cD5) of the signal before mentioned. We can define it as the Eq. (5):

$$Vc_{f_i[n]} = cD5_{f_i[n]} \qquad (5)$$

4 Artificial Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. An ANN can be configured for a specific application, such as pattern recognition or data classification, through a learning process.

Two phases in all NN application exist: the phase of learning or training and the phase of test. Once trained the model with the type of selected learning, the phase of test is followed, in which the representative features of the inputs, called training patterns, are processed. After calculated the weights of the network, the values of the last layer neurons, they are compared with the wished output to state the suitability of the design.

4.1 Neural Networks for Classification

By using the characteristic vectors as inputs in an Artificial Neural Network, it is possible to classify the signals in its condition. We compared three neural Networks for classification widely used: MLP, PNN and RBF. The 70% of all signals are used for training, and leave the rest of the signals for testing.

In order to select a good Neural Network configuration, there are several factors to take into consideration. The major points of interest regarding the ANN topology selection are related to (1) network design, (2) training and (3) some practical considerations.

1) Network design points such as determining the input and output variables, the number of input and output nodes to be used, the number of hidden layers in the neural network, the number of hidden nodes in each hidden layer, and the size of the training data and testing data sets. 2) Network training considerations such as initialing the network weights, choosing proper training parameter values (such as the learning rate and momentum rate), and selecting the training termination criteria; and 3) practical considerations such as network accuracy, network robustness, and implementation feasibility. Even though selecting these parameters is partly a trial-error process, there are some guidelines that can be used in choosing these values. For readers who are interested in a detailed description, please refer to [14, 15]

The input layer is conformed by the characteristic vector developed; the number of nodes is the length of the vector. The output variables are four (4), exactly the number of conditions that we want to classify, and the targets for training are the followings:

Normal conditions	[0 0 0 1]
Inner race faults	[0 0 1 0]
Outer race faults	[0 1 0 0]
Ball fault	[1 0 0 0]

The numbers of Hidden Layers and Hidden nodes affect both the network accuracy and the time required training the neural network. In this paper, we limit ourselves to a three-layer net, because this structure is the most popular one and has been proven to be able to learn arbitrarily complicated continuous functions [16]. In order to realize an objective comparison of various Neural Networks, it is necessary to set the number of hidden nodes in 10, 20 and 30 and for different tests.

It is a common practice to choose a set of training data and a set of testing data that are statistically significant to represent the system under consideration. The training data set is used to train the neural network, and they include validation, while the testing data set is used to test the networks accuracy after the network has been successfully trained. We will use the pre-processed sets extracted from the experimental setup. The normalization of the input values will increase the numerical stability of the neural network processing, while the normalization of the output values is necessary because of the characteristics of the activation function that is used in processing units of the neural network. We will normalize the inputs between [0 1] and the output between [0.1 0.9].

After the network configuration and input-target training data have been determined, we need to select the training parameters, which include the initial network weights, the learning rate, the momentum rate and the criteria to stop network training. The choice of these values is sometimes critical to the success of the neural network training process. Unfortunately, the choice of these values is generally problem dependent. There is no generic formula that can be used to choose these parameter values. Nevertheless, some guidelines, which are described below, can be followed as an initial trial. After a few trials, the neural network designer should have enough experience to set the appropriate parameters for any given problem.

A critical parameter is the speed of convergence, which is determined by the learning coefficient. In general, it is desirable to have fast learning, but not so fast as to cause instability of learning iterations. Starting with a large coefficient and reducing it as the learning process proceeds, results in both fast learning and stable iterations

Different measures can be used to train and monitor the Training Process. We use the two-norm error measure to train and monitor the performance of the neural network, as described in Eq. 6

$$E=\left\|\hat{y}(x,w)-y(x)\right\|_2=\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{y}(x_p,w)-y(x_p)\right)^2},\ p=1,...,P \quad (6)$$

Where:

$\hat{y}(x,w)$ Is the network output

$y(x_p)$ Is the actual system output with given the same input x_p, and w are the current weights of the network

Popular criteria used to terminate network training are: a) sufficiently small mean-square training error and b) sufficiently small changes in training error. How sufficiently small is usually up to the network designer and based on the desired level of accuracy of the neural network. We terminate the network training when either the root mean-square error of the training data set or the change in network error is less than 0.005.

Information is stored in a feedforward net in its weights. The more weights a network has, the more information it can store. The number of weights is a function of the number of hidden nodes for a three-layer feedforward net. Thus, the more hidden nodes a network has, the more information the net can store. Furthermore, the network is able to learn faster [14].

The initial weights of the neural network play a significant role in the convergence of the training method. Without a priori information about the final network weights, it is common practice to initialize all weights with random numbers of small absolute values between [-0.1, 0.1]. In linear vector quantization and derived techniques, it is usually required to renormalize the weights at every training epoch.

4.1.1 Multilayer Perceptron (MLP)

An MLP is a network of simple neurons called perceptrons. The basic concept of a single perceptron was introduced by Rosenblatt in 1958 [15]. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. An example Network is shown in Fig 6.

They are supervised networks, so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification. MLP can approximate virtually any input-output map. They have been shown to approximate the performance of optimal statistical classifiers in difficult problems. The most NN applications involve MLPs.

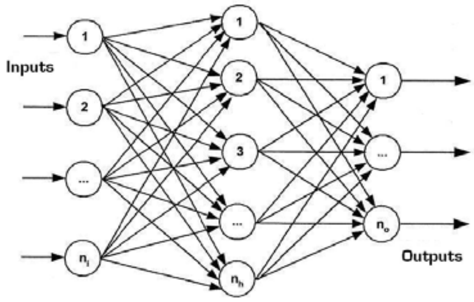

Fig. 6. Typical MLP Structure

4.1.2 Radial basis Function (RBF)

Radial basis function (RBF) networks have a static Gaussian function as the nonlinearity for the hidden layer processing elements. The Gaussian function responds only to a small region of the input space where the Gaussian is centered. The key to a successful implementation of these networks is to find the suitable centers for the Gaussian functions. This can be done with supervised learning. The structure of a RBF is simple; it contains an input layer, a hidden layer with nonlinear activation functions and an output layer with linear activation functions. Figure 7 shows the architecture of a RBF.

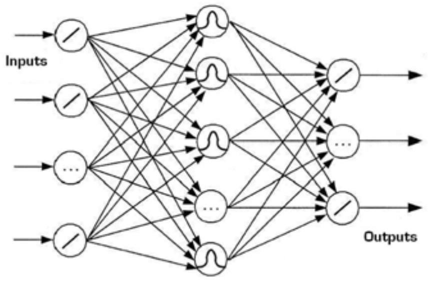

Fig. 7. Typical RBF Structure

4.1.3 Probabilistic Neural Networks (PNN)

This network provides a general solution to pattern classification problems by following an approach developed in statistics, called Bayesian classifiers. Bayes theory, developed in the 1950's, takes into account the relative likelihood of events and uses a priori information to improve prediction. The network paradigm also uses Parzen Estimators which were developed to construct the probability density functions required by Bayes theory.

The probabilistic neural network uses a supervised training set to develop distribution functions within a pattern layer. These functions are used to estimate the likelihood of an input feature vector being part of a learned category. The learned patterns can also be combined, or weighted, with the a priori probability, also called the relative frequency, of each category to determine the most likely class for a given input vector. If the relative frequency of the categories is unknown, then all categories can be assumed to be equally likely and the determination of category is solely based on the closeness of the input feature vector to the distribution function of a class. An example Network is shown in Fig 8.

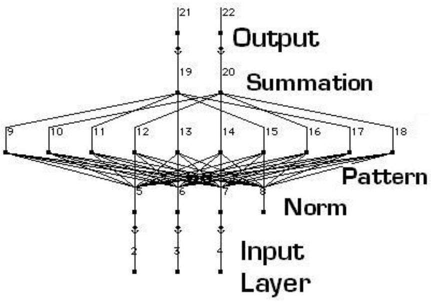

Fig. 8. Typical RBF Structure

5 Discussion

Figures 9-13 shows the training and testing success of the three Neural Networks studied, it can be seen that the PNN Network gives better results and can predict reliably the faults in bearing with a success rate upper to 92%. The success rate is a function of the number of Hidden neurons or "processing units" in the hidden layer, because of the learning process in a Neural Network is more complex while more neurons have. It is necessary to clarify the use of three different amounts of hidden neurons, so this approach is based on the traditional design of neural networks, and there is not a specific guideline to obtain the "best" number of hidden neurons, and the trial & error technique is fundamental to begin the work, it can be seen that 30 neurons results better than 10 or 20, but it has to be considered that the processing time is bigger too.

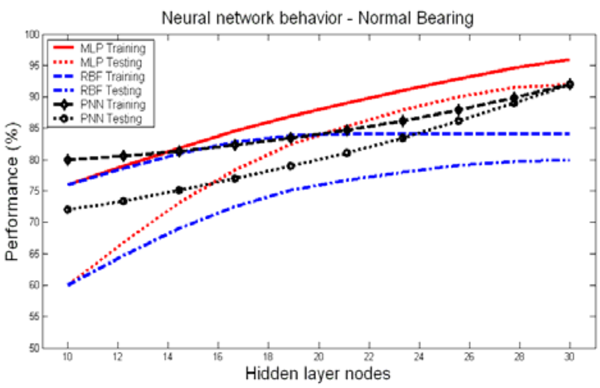

Fig. 9. Accuracy of the Neural Networks for predicting the normal bearing condition for different number of hidden nodes.

The Neural Network behavior is deeply affected for a lot of variables, and there is not a particular methodology for each engineering problem. One of the most important variables are the training algorithm, which is not discussed in this paper, and the will be mentioned in the near future. Other tests not included here have shown that it exist an overtraining phenomenon for 40 and more neurons. It may happen sometimes because the network learns all the training datasets and lose its ability to generalize with the test dataset.

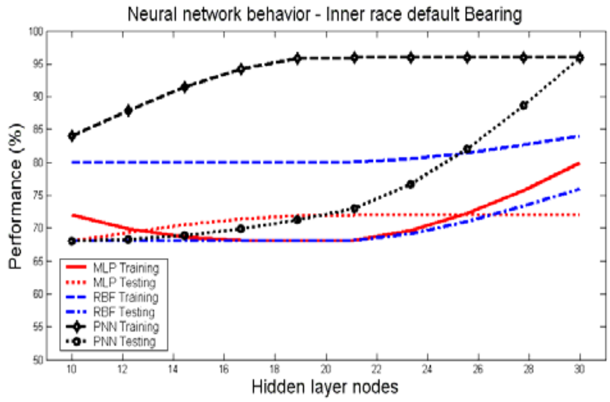

Fig. 10. Accuracy of the Neural Networks for predicting the Inner race fault bearing condition for different number of hidden nodes.

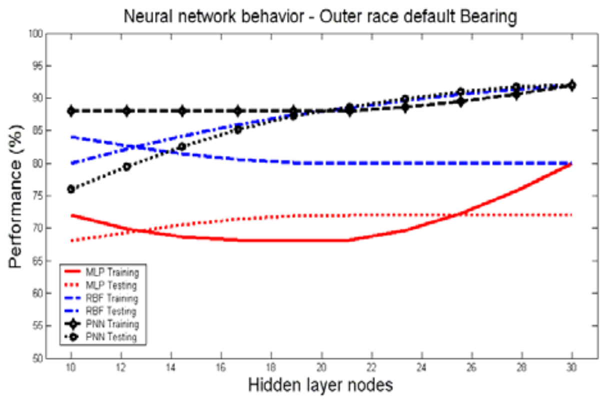

Fig. 11. Accuracy of the Neural Networks for predicting the Outer race fault bearing condition for different number of hidden nodes.

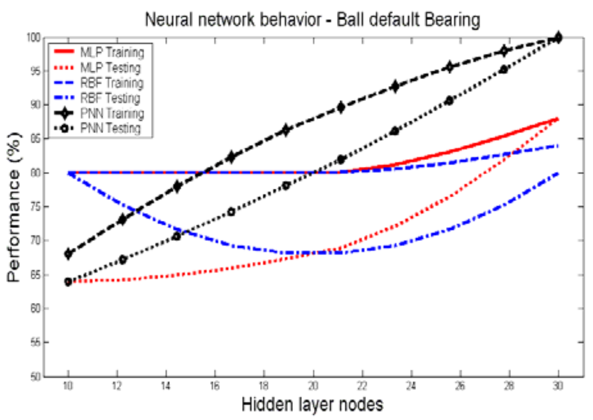

Fig. 12. Accuracy of the Neural Networks for predicting the Ball fault bearing condition for different number of hidden nodes.

6 Conclusions

This work has presented an automatic fault classification technique based on DWT and three different Neural Networks, MLP, RBF and PNN. The following conclusions are drawn:

Vibration data can be classified based on the DWT and Neural Networks.

The three Neural Networks can be used to classify the signals upper 70% of rate success.

The probabilistic Neural Network shows the best results for our experiments, and the proposed network is superior to the traditional methods in classifying the fault characteristic of bearings.

A novel Neural Networks comparison study has been developed in the fault classification field.

Acknlowledgement

The authors wish sincerely thanks the Spanish Government for financing provided through the MCYT project DPI-2003-084790-C02-01.

References:

[1] Loparo, K. Bearing fault diagnosis based on wavelet transform and fuzzy inference. Mechanical Systems and Signal Processing, Vol. 18. 2004, pp. 1077-1095.

[2] Mallat, S. A wavelet tour of signal processing. Academic Press, 1998.

[3] Polikar, R. The Wavelet tutorial. Rowan University. College of Engineering. Available in Internet:http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html. 1999.

[4] Leducq D. Hidraulic noise diagnostic using Wavelet analysis. Proceedings of the international conference on noise control engineering. 1990, pp. 997-1000.

[5] Wang, W.J., McFadden, P.D. Application of the Wavelet transform to gearbox vibration analysis. American society of mechanical engineers. Petroleum division. 1993, pp. 13-20.

[6] Newland, D.E. Ridge and phase identification in the frequency analysis of transient signals by harmonic Wavelets. Journal of Vibration and acoustic. Transactions of the ASME. 1999, pp. 149-155.

[7] Wang, W.Q., Ismail, F. Golnaragh, M.F. Assessment of gear damage monitoring techniques using vibration measurements. Mechanical Systems and Signal Processing., Vol. 15. 2001, pp. 905-922.

[8] Liu, B., Ling, S.F. Machinery diagnostic based on Wavelet Packets. Journal of vibration and control, Vol. 3. 1997, pp. 5-17.

[9] Aminian, F., Aminian, M. Fault diagnostic of analog circuits using bayesian neural networks with Wavelet transform as preprocessor. Journal of electronic testing, Vol. 17. 2001, pp. 29-36.

[10] Farias, G. Santos, M., Marron, F. Determinación de Parámetros de la Transformada Wavelet para la Clasificación de señales. XXV Jornadas de Automática, U. C. L. M. 2004.

[11] Peng, Z. K. Application of the Wavelet transform in machine condition monitoring and fault diagnostics. Mechanical Systems and Signal Processing, Vol. 18. 2004, pp. 199-221.

[12] Liu, B., Ling, S.F. On the selection of informative Wavelets for machinery diagnosis. Mechanical Systems and Signal Processing, Vol. 13. 1999, pp. 145-162

[13] Goumas, S.K., Zervakis, M.E. Classification of washing machines vibration signals using DWT for feature extraction. IEEE Transactions on instrumentation, Vol.51, No.3. 2002, pp. 497-508.

[14] Chow, M. Methodologies of Using Neural Network and Fuzzy-Logic Technologies for Motor Incipient Fault Detection. World Scientific. 1997.

[15] Godoy, M., Meireles, M., Almeida, P. A comprehensive review for industrial applicability of artificial neural networks. IEEE transactions on industrial electronics, Vol. 50. 2003, pp. 585-601.

[16] Cybenko, G. Approximation by superpositions of a sigmoidal function. Mathematics of control, Signals and Systems, Vol. 2, pp. 303-314.