

## Chapter 8

# Wavelet Packet-Transform for Defect Severity Classification

Once a defect is detected, the next question that comes up naturally is how severe the defect is. Since machine downtime is physically rooted in the progressive degradation of defects within the machine's components, accurate assessment of the severity of defect is critically important in terms of providing input to adjusting the maintenance schedule and minimizing machine downtime. This chapter describes how wavelet packet transform (WPT)-based techniques can classify machine defect severity, with specific application to rolling bearings.

### 8.1 Subband Feature Extraction

Because of the complex nature of machines and the intricacy of related parameters, it is generally difficult to assess the status of a machine directly from the measured time domain data. The general practice is to extract “features” to identify characteristics and patterns embedded in the data series that are indicative of status changes of the machine being monitored. The advent of wavelet transform has provided an effective tool for feature extraction from various time-varying signals, such as washing machines (Goumas et al. 2002), rolling bearings (Mori et al. 1996; Prabhakar et al. 2002), and machine tools (Lee and Tang 1999; Li et al. 2000a, 2000b). As an extension of the wavelet transform, WPT provides more flexible time–frequency decomposition, especially in the high-frequency region, when compared with the wavelet transform. In particular, WPT allows for feature extraction (e.g., energy content or kurtosis value) from subfrequency bands of the decomposed signal where the features are concentrated, thereby directing the computation to where it is most needed. Prior efforts have studied different sets of wavelet packet vectors to represent bearing vibration under different defect conditions (Liu et al. 1997). Altmann and Mathew (2001) found that features extracted from wavelet packets that cover the multiple subfrequency bands yield a higher signal-to-noise (S/N) ratio than those from a conventional band-pass filter. For multistage gearbox vibration analysis, the Hilbert transform and WPT were combined to enable gear defect detection at the incipient stage (Fan and Zuo 2006).

Given a time-domain signal  $x(t)$ , the WPT decomposes it into a number of subbands, as expressed by the resulting wavelet packet coefficients:

$$x(t) = \sum_{n=0}^{2^j-1} x_j^n(t) \quad (8.1)$$

In (8.1), the term  $x_j^n(t)$  denotes the wavelet coefficients at the  $j$  level,  $n$  subband. From these coefficients, “features” will be extracted, at each subband, to provide information on the condition of the machine being monitored.

### 8.1.1 Energy Feature

The energy content of a signal provides a quantitative measure for the signal, which uniquely characterizes the signal. The amount of energy contained in a signal  $x(t)$  is expressed as:

$$E_{x(t)} = \int |x(t)|^2 dt \quad (8.2)$$

The energy content of a signal can also be calculated from the coefficients of the signal’s transform. In the case of a WPT, the coefficients  $x_j^n(t)$  quantify the amount of energy associated with each of the subbands. The total amount of energy contained in the signal is equal to the sum of the energy in each subband and expressed as:

$$E_{x(t)} = \sum_{n=0}^{2^j-1} \int |x_j^n(t)|^2 dt \quad (8.3)$$

Since the energy content of each subband of the signal is directly related to the severity of the defect, it presents an indicator or *feature* of the machine’s condition. From (8.3), the energy feature in each subband is defined as:

$$E_j^n = \int |x_j^n(t)|^2 dt \quad (8.4)$$

Similarly, when a signal is represented by discrete, sampled values  $x(i)$  ( $i = 1, 2, \dots, M$ ), the total energy feature in the subbands is calculated as:

$$E_j^n = \sum_{i=1}^M x_j^n(i)^2 \quad (8.5)$$

### 8.1.2 Kurtosis

Kurtosis is a dimensionless, statistical measure that characterizes the flatness of a signal's probability density function. An impulsive signal that is peaked has a larger kurtosis value than a signal that is flat and varies with time slowly, as illustrated in Fig. 8.1.

Mathematically, the kurtosis of a signal is defined as its fourth-order moment:

$$K_{x(t)} = \frac{E[(x(t) - \mu_{x(t)})^4]}{\sigma_{x(t)}^4} \quad (8.6)$$

where  $\mu_{x(t)}$  and  $\sigma_{x(t)}$  denotes the mean value and standard deviation of the signal  $x(t)$ , respectively. The symbol  $E[\bullet]$  denotes the expectation operation. Table 8.1 lists the kurtosis values of several representative signals. It should be noted that the value of kurtosis does not depend on the amplitude of a signal.

For the wavelet packet coefficients in each subband, the corresponding kurtosis value is defined as:

$$K_j^n = \frac{E[(x_j^n(t) - \mu_{x_j^n(t)})^4]}{\sigma_{x_j^n(t)}^4} \quad (8.7)$$

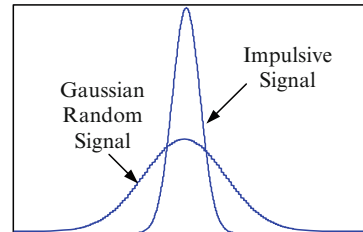
where the symbols  $\mu_{x_j^n(t)}$  and  $\sigma_{x_j^n(t)}$  are the mean and standard deviation of the wavelet packet coefficients  $x_j^n(t)$ , respectively.

When the wavelet packet coefficients are sampled as  $x_j^n(i)$ , the kurtosis value is calculated as:

$$K_j^n = \frac{\sum_{i=1}^N [x_j^n(i) - \mu_{x_j^n(i)}]^4}{N\sigma_{x_j^n(i)}^4} \quad (8.8)$$

where the symbols  $\mu_{x_j^n(i)}$  and  $\sigma_{x_j^n(i)}$  are the mean and standard deviation of the wavelet packet coefficients  $x_j^n(i)$ , respectively.

Since the energy content of a signal provides a robust indicator of the signal, but is not sensitive in characterizing incipient defects, whereas the kurtosis



**Fig. 8.1** Illustration of probability density functions of signals

**Table 8.1** Kurtosis values of several typical signals

Signal	Kurtosis
Square signal	1.0
Sinusoidal signal	1.5
Gaussian signal	3.0
Pulse signal	>3.0

value has high sensitivity to incipient defects but has low stability (Yan and Gao 2004), these two features can be combined instead of being used alone to improve the signal characterization. Suppose we decompose a signal into  $j$  levels (e.g.,  $j = 4$ ), which generates  $2^j$  or  $2^4 = 16$  subbands. Given that the energy and kurtosis values are calculated from each subband, there will be  $2 \times 2^j$  or  $2 \times 2^4 = 32$  features extracted from the signal. These features can be expressed in a feature vector as:

$$FV = [E_j^0, E_j^1, \dots, E_j^{2^{j-1}}, K_j^0, K_j^1, \dots, K_j^{2^{j-1}}]^T \quad (8.9)$$

For simplicity, (8.9) can be expressed as:

$$FV = \{f_l | l = 1, 2, \dots, p\}, \quad p = 2^{j+1} \quad (8.10)$$

where  $f_1 = E_j^0, \dots, f_{2^j} = E_j^{2^j-1}, f_{2^j+1} = K_j^0, \dots, f_p = K_j^{2^j-1}$ .

Determining which features shown in (8.10) are most effective for characterizing machine defect is generally not a simple, straightforward process because the usefulness of features may be affected by factors such as the specific location of the sensors, and consequently, the quality of the signal measured in terms of the S/N ratio or signal contamination. Furthermore, using more features may not necessarily improve the effectiveness of defect severity estimation, while increasing the computation cost (Malhi and Gao 2004). Since defect-induced signals are typically reflected in the variation of the characteristic frequencies (e.g., characteristic defect frequency shifts as the defect size grows), degradation of machine conditions is predominantly reflected in certain subfrequency bands, whereas other subfrequency bands contain information unrelated to the defect. This indicates that feature selection is needed for identifying significant features from the pool of WPT-based feature set.

## 8.2 Key Feature Selection

This section introduces two feature selection methods: Fisher linear discriminant (FLD) analysis and principal component analysis (PCA). The goal is to differentiate the signals (which represent different machine defect severity) by examining only

those subbands with distinct feature discrimination than others, thus improving the efficiency while not missing critical information related to the signal, to ensure reliability of the diagnostic operation.

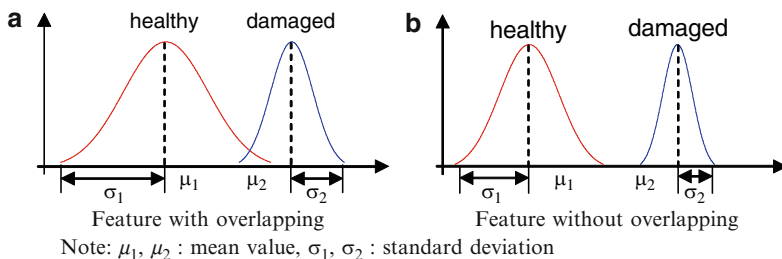
### 8.2.1 Fisher Linear Discriminant Analysis

Distance measures, such as the Bhattacharyya distance, Kolmogorov distance, and FLD (Fukunaga 1990; Yen and Lin 2000; Duda et al. 2001), have been applied to components differentiation within a class pair. Generally, the greater the distance between two feature components within a class pair is, the easier it will be to separate them. Illustrated in Fig. 8.2 are two feature components representing the two signals from a healthy and a defective bearing, respectively. The features in the right-hand section of the figure are easier to be separated than those in the left-hand section because the probability distributions of the two components do not overlap, because of their relatively smaller standard distributions and larger distance between the mean values, compared with the two features in the left section. The result is that this pair of features has a higher discriminant power than the other pair of features.

The approach introduced here for efficient feature selection is to evaluate the discriminant power of each individual feature within a class pair. Features that have a low discriminant power are excluded from the data analysis process, as they contain little useful information. Such an approach can be realized by examining the rank order (Kittler 1975) of the feature vector  $FV = \{f_l | l = 1, 2, \dots, p\}$ , shown in (8.10), as:

$$J(f_1) \geq J(f_2) \geq \dots \geq J(f_d) \geq \dots \geq J(f_p) \quad (8.11)$$

where  $J(\bullet)$  is a criterion function for evaluating the discriminant power of a specific feature. Here the utility of the FLD is introduced (Duda et al. 2000), where the criterion function for differentiating a class pair is given by:

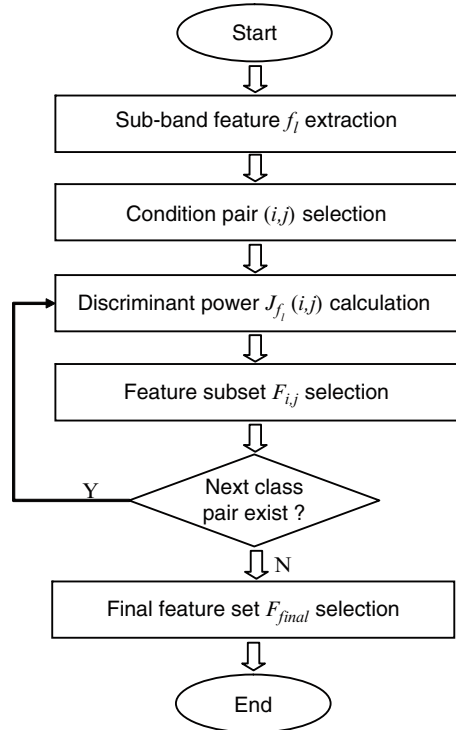


**Fig. 8.2** Feature discrimination based on the distance between constituent components. (a) Feature with overlapping. (b) Feature without overlapping. Note:  $\mu_1, \mu_2$  : mean value,  $\sigma_1, \sigma_2$  : standard deviation

$$J_{f_l}(i, j) = \frac{|\mu_{i,f_l} - \mu_{j,f_l}|^2}{\sigma_{i,f_l}^2 + \sigma_{j,f_l}^2} \quad (8.12)$$

The symbols  $\mu_{i,f_l}$ ,  $\mu_{j,f_l}$  and  $\sigma_{i,f_l}^2$ ,  $\sigma_{j,f_l}^2$  represent the mean values and the variances of the  $l$ th feature,  $f_l$ , and for the classes  $i$  and  $j$ , respectively. Since typically more than two defect severity levels need to be evaluated in a machine defect diagnosis system, a  $k$ -class,  $p$  feature component problem with  $k(k-1)/2$  class pairs is investigated for generality. The process for feature selection, based on the FLD analysis method, is illustrated in Fig. 8.3.

Features (i.e., subband energy and kurtosis values) are first extracted by means of the WPT from the signals measured on the machine (e.g., a milling machine, a spindle), under various operating conditions (e.g., speed and load). The mean values and variances of each individual feature  $f_l$  corresponding to each machine status are then calculated, for each set of operation conditions (e.g., 1,200 rpm, 3.6 kN radial load). For each possible class pair  $\{(i, j) | i = 1, 2, \dots, k-1, j = i+1, i+2, \dots, k\}$  formed from two different machine states (e.g., health vs. light defect), the discriminant power measure  $J_{f_l}(i, j)$  for each feature  $f_l$ , is calculated, using (8.12). Descending sorting  $J_{f_l}(i, j)$  yields:



**Fig. 8.3** Flowchart of the FLD feature selection process

$$J_{f_1}(i, j) \geq J_{f_2}(i, j) \geq \cdots \geq J_{f_d}(i, j) \geq \cdots \geq J_{f_p}(i, j) \quad (8.13)$$

The first group of  $d$  features that have the highest relative  $J_{f_i}(i, j)$  values are chosen to form the feature subset  $F_{i,j}$ , for each class pair  $(i, j)$  :

$$F_{i,j} = \{ f_l | l = 1, 2, \dots, d \}, \quad i = 1, 2, \dots, k-1; j = i+1, i+2, \dots, k \quad (8.14)$$

The final feature set is obtained by taking the union of each feature subset across all the class pairs as:

$$F_{final} = \left\{ \bigcup_{i=1}^{L-1} \bigcup_{j=i+1}^L F_{i,j} \right\} \quad (8.15)$$

This feature set is subsequently selected for the machine defect severity classification.

### 8.2.2 Principal Component Analysis

PCA, as a multivariate statistical technique, has been intensively studied and utilized as an effective tool for process monitoring (Kano et al. 2001), structural damage identification (De Boe and Golival 2003), and machine health diagnosis (Baydar et al. 2001; He et al. 2008). This is due to its ability in dimension reduction and pattern classification. In general, the PCA technique seeks to determine a series of new variables, called the principal components, which indicates the maximal amount of variability in the data with a minimal loss of information (Jolliffe 1986), to best represent the data in a least square sense.

Suppose there are  $m$  feature vectors  $FV_i (i = 1, 2, \dots, m)$  extracted from  $m$  signals, respectively. A  $p \times m$  feature matrix  $X$  can then be formulated as:

$$X = [FV_1, FV_2, \dots, FV_m] \quad (8.16)$$

where the symbol FV denotes a  $p$ -dimensional feature vector as shown in (8.10). Depending on the decomposition level  $j$  of the WPT, the dimension of the feature vector is determined as  $p = 2^{j+1}$ . Correspondingly, a scatter matrix  $S$  is constructed from the feature matrix  $X$  as:

$$S = E[(X - \bar{X})(X - \bar{X})^T] \quad (8.17)$$

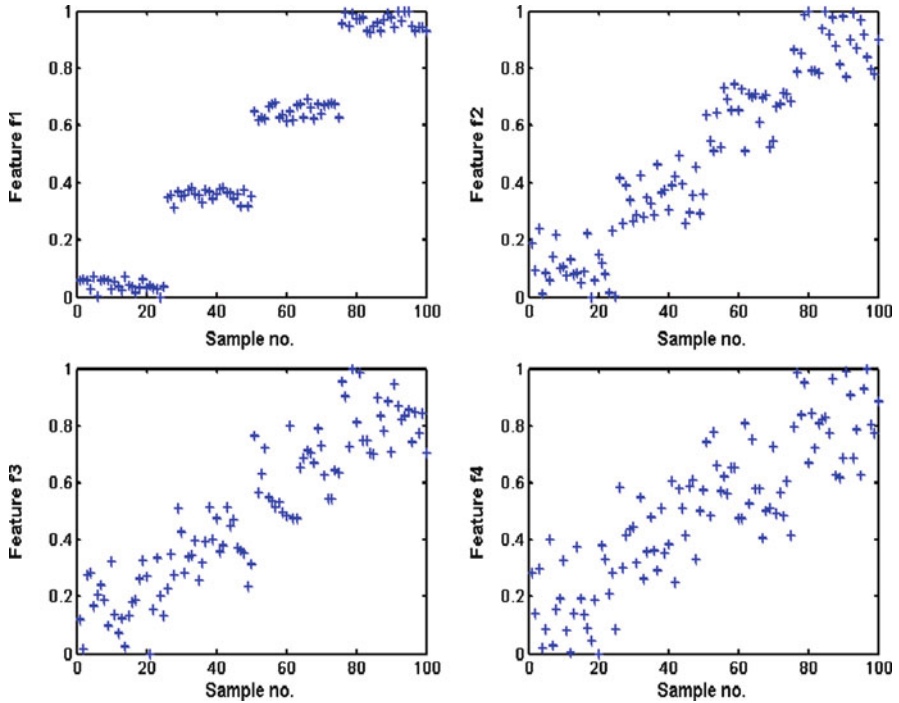
Where  $\bar{X}$  is the mean value of  $X$ , and  $E[\bullet]$  is the statistical expectation operation (Duda et al. 2000). Performing singular value decomposition on the scatter matrix leads to:

$$S = A\Lambda A^T \quad (8.18)$$

where  $A$  is a  $p \times p$  matrix whose columns are the orthonormal eigenvectors of the scatter matrix, and  $A^T A = I_p$ . The symbol  $\Lambda$  is a diagonal matrix whose diagonal elements  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  are the eigenvalues of the scatter matrix. Since the eigenvector in matrix  $A$  with the highest eigenvalue (i.e.,  $\lambda_1$  in the diagonal matrix  $\Lambda$ ) is the first principle component of the  $p$ -dimensional feature vectors, it is better-suited than any other feature vectors as the representative feature that identifies the condition of the machine being monitored, for example, defect severity of a bearing. As a result, PCA ranks the order of eigenvectors by means of their respective eigenvalues, from the highest to the lowest. Such a ranking sequence reflects upon the order of significance of the corresponding components. By examining the accumulated variance (e.g., 90%) of the principle components, which is defined as:

$$\text{var} = \left( \frac{\sum_{i=1}^q \lambda_i}{\sum_{j=1}^p \lambda_j} \right) \times 100\% \quad (8.19)$$

a lower-dimensional feature vectors  $Y$  can be constructed as:



**Fig. 8.4** Simulated data of  $(f_1, f_2, f_3, f_4)$  for developing a feature selection scheme



$$Y = A_{p \times q}^T X \quad (8.20)$$

where  $q < p$ , and  $A_{p \times q}$  is the first  $q$  columns of  $A$ .

Given that the features transformed by the principal components are not directly connected to the physical nature of the defect, the eigenvectors in  $A_{p \times q}$  for the transformed features are only used as the basis for choosing the most significant features from the original  $p$ -dimensional feature vectors. This is explained by means of a numerical simulation. As illustrated in Fig. 8.4, four normalized feature vectors,  $f_1, f_2, f_3$ , and  $f_4$ , are constructed with each of them forming clusters around four distinct levels of magnitudes. A total of 100 samples are considered for each of the four features, hence each feature is a 100-by-1 vector. The four features are simulated to have random variations from the same mean for each of the four clusters. This is similar in principle to the variation of a measured data feature for four different defect severities. Each of the four clusters for each feature contained 25 data points. The four features become less clearly differentiated from  $f_1$  to  $f_4$ , as overlap between the clusters increases.

It is evident that a suitable feature selection scheme should be able to rank  $f_1, f_2, f_3$ , and  $f_4$  in the same order as shown in Fig. 8.4. To derive the principal components for the simulated data set, the four normalized features are collected in a 4-by-100 matrix  $X$ :

$$X = [f_1, f_2, f_3, f_4]^T \quad (8.21)$$

The eigenvalues and the eigenvectors are calculated from the scatter matrix  $S$ . The matrix of eigenvectors can be represented as  $A = [a_{ij}]$ , where  $i = 1$  to 4, and  $j = 1$  to 4. The eigenvector  $a_4$  consists of four components from the fourth column of the matrix  $A$  as  $a_4 = [a_{1,4} \ a_{2,4} \ a_{3,4} \ a_{4,4}]$ . Similar arrangement applies to  $a_1, a_2$ , and  $a_3$  (i.e.,  $a_1 = [a_{1,1} \ a_{2,1} \ a_{3,1} \ a_{4,1}]$ ,  $a_2 = [a_{1,2} \ a_{2,2} \ a_{3,2} \ a_{4,2}]$ , and  $a_3 = [a_{1,3} \ a_{2,3} \ a_{3,3} \ a_{4,3}]$ ), respectively. The matrix  $A$  is a  $4 \times 4$  square matrix because of the presence of the four features  $f_1$ – $f_4$ . The eigenvector corresponding to the eigenvalue with the largest magnitude is chosen. As shown in Table 8.2, one of the four eigenvalues of the data set is much larger than the other three, indicating that most of the variance is concentrated in one direction. Table 8.3 lists the component magnitudes for the eigenvector corresponding to the largest eigenvalue. Since this corresponds to  $a_4$ , the feature that is responsible for the maximum variance in the data is thus identified.

Subsequently, the magnitudes of the four components of  $e_4$  are examined. As shown in Table 8.3,  $|a_{1,4}| > |a_{2,4}| > |a_{3,4}| > |a_{4,4}|$ . This result can be interpreted in terms of the directionality of the eigenvector ( $a_4$ ) in the original feature space. If the unit vectors for the original feature space are represented as  $u_1, u_2, u_3$ , and  $u_4$  (where  $u_1 = [1 \ 0 \ 0 \ 0]^T$ ,  $u_2 = [0 \ 1 \ 0 \ 0]^T$ , etc.), then a higher magnitude of  $a_{i,4}$  denotes the similarity in direction of the eigenvector  $a_4$  with  $u_i$ , when compared with the other unit vectors forming the basis for the original feature space. For the simulated data, the component  $a_{1,4}$  has the largest magnitude, followed by  $a_{2,4}$ ,  $a_{3,4}$ , and  $a_{4,4}$ .

**Table 8.2** Eigenvalues for simulated data

$\lambda_1$	0.241
$\lambda_2$	0.775
$\lambda_3$	1.318
$\lambda_4$	32.023

**Table 8.3** The fourth eigenvector component magnitudes for simulated data

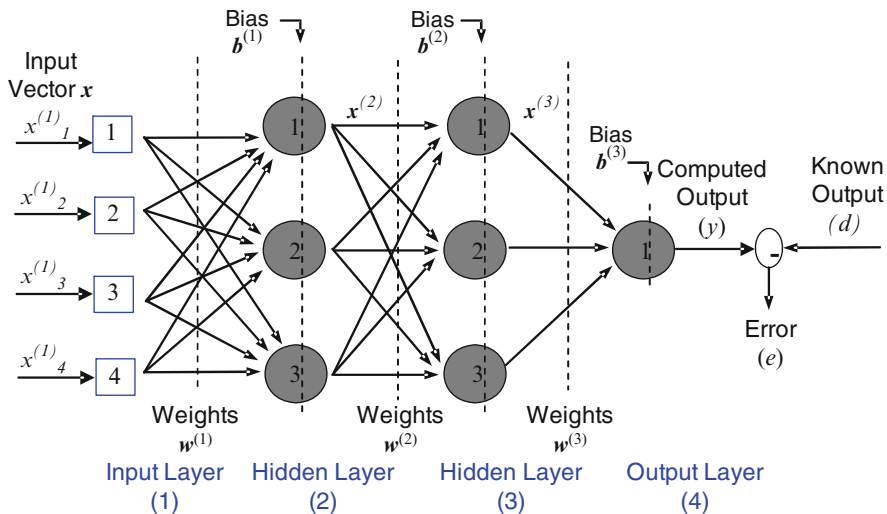
Component	Magnitude
$a_{1,4}$	0.599
$a_{2,4}$	0.523
$a_{3,4}$	0.439
$a_{4,4}$	0.417

Thus, the feature represented along  $u_1$  is the most sensitive, followed by those along  $u_2$ ,  $u_3$ , and  $u_4$ . As a result, the PCA-based scheme ranks the four features  $f_1$ – $f_4$  as desired and selects most representative features.

### 8.3 Neural-Network Classifier

Once a suitable feature set (e.g., 6) is chosen from the extracted features (e.g., 32), the machine defect severity levels can be evaluated by means of a status classifier. Neural network as a classifier has been applied to machine health diagnosis, for example, for classifying rotating machines with imbalance and rub faults (McCormick and Nandi 1997), bearing faults (Li et al. 2000), and tank reactor operation states (Maki and Loparo 1997). In general, a neural network consists of multiple layers of nodes or neurons, and each layer has a number of parallel nodes that are connected to all the nodes in the succeeding layer through different weights (Haykin 1994). Using a training algorithm, the weights are adjusted such that the neural network responds to the inputs with outputs corresponding to the severity of a structural defect at the output layer. Figure 8.5 illustrates the architecture of a feed-forward neural network. For the  $i$ th layer of links, the symbols  $\mathbf{w}^{(i)}$ ,  $\mathbf{b}^{(i)}$ ,  $\mathbf{x}^{(i)}$ , and  $\mathbf{y}^{(i)}$  represent a vector of weights between the layers, node biases of a layer, inputs of nodes at one layer, and output at the output layer, respectively. At the output layer, a linear neuron is used to produce an output to indicate the machine defect severity level.

The weights of a multilayer feed-forward neural network are continuously updated, while it is trained with training data consisting of a set of machine defect feature input vectors ( $\mathbf{x}$ ) and known output ( $d$ ). This is realized by minimizing the error between the computed output of the network and the known output in the training process. Consider  $n$  pairs of input and output training data  $\{(\mathbf{x}_p, d_p)\}$ ,  $p = 1, 2, \dots, n$ . For the  $p$ th pair data  $\{\mathbf{x}_p, d_p\}$ , the mean square error (MSE) of the network output  $y_p$  is expressed as:



**Fig. 8.5** Structure of a multilayer feed-forward neural network

$$e_p = \sum_{m=1}^j (d_{pm} - y_{pm})^2 \quad (8.22)$$

where  $m$  is the number of nodes at the output layer. Assuming that each input vector corresponds to a single severity value, the value of  $m$  is 1. For the entire training data set, the total error  $Err$ , i.e., the learning error, is expressed as:

$$Err = \sum_{p=1}^n e_p = \sum_{p=1}^n \sum_{m=1}^j (d_{pm} - y_{pm})^2 \quad (8.23)$$

In the training process, the learning error is minimized through continuously updating the connection weights in its structure with certain learning rule. After training with the training data, the designed network with the resulting connection weights generalize the relationship between the input and output to correctly classify new input data. When input feature vectors associated with a defective measurement occur, for which the network is however not trained, the neural network will interpolate a defect severity by the location of the new input in the space spanned by the training data.

There are several gradient-based learning rules to minimize the learning error  $Err$  by changing the connection weight  $w$  of the multilayer feed-forward neural network. Different learning rules differ in how they use the gradients to update the weights  $w$  in training. Steepest decent with fixed learning rate is the traditional learning rule of the neural network, in which the weight  $w^{(k)}$  between the  $k$ th layer and  $(k + 1)$ th layer is tuned for each epoch, along the gradient direction by an amount:



in each subband, key feature selection process is performed to determine the most significant features from the feature set, which are subsequently used as input to a neural network-based classifier for defect severity classification. Figure 8.6 illustrates how the developed technique is realized. The left side of Fig. 8.6 depicts the training process of the hybrid technique in a manner of supervised learning (i.e., based on the available reference data, denoted as signal 1, ...,  $n$ ). In addition to providing inputs for constructing the neural-network classifier model, the results from the feature selection process are used to guide the feature vector selection in the evaluation process. The right side of Fig. 8.6 describes an evaluation process of the WPT-based algorithm. An input signal is passed through the process of signal decomposition, feature extraction, and selection. Eventually, the corresponding defect severity level is determined through the neural network classifier.

## 8.5 Case Studies

The application of the above described wavelet packet-based machine defect severity classification algorithm is described through two case studies.

### 8.5.1 Case Study I: Roller Bearing Defect Severity Evaluation

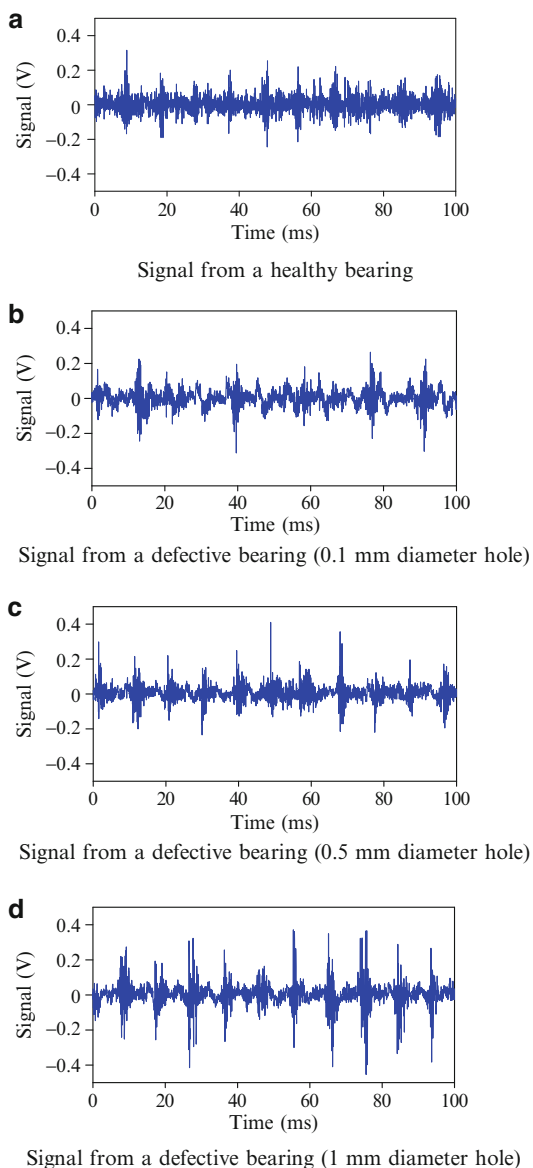
The first case study is to evaluate the defect severity level of a set of roller bearings (N205 ECP) with and without seeded defects. Specifically, vibration data were measured from both a new, “healthy” bearing that served as a reference base and three defective bearings containing localized defects of different sizes:

- (a) one 0.1 mm diameter hole in the outer raceway
- (b) one 0.5 mm diameter hole in the outer raceway
- (c) one 1 mm diameter hole in the outer raceway

In Fig. 8.7, segments of vibration signals measured from the healthy and defective bearings are shown.

To provide sufficient training and testing data sets to the neural-network classifier, a total of 240 vibration data sets were collected under a bearing rotating speed of 1,200 rpm and a radial load of 3,600 N. For each operation condition, 60 data sets were collected. Each data set was first decomposed by the WPT. Analysis has shown that features extracted from a four-level decomposition provided adequate information to differentiate the four defect conditions from each other (Gao and Yan 2007). On the basis of the information collected in the 16 subfrequency bands (since  $2^4 = 16$ ), a feature vector was subsequently constructed, which contained 32 feature elements (i.e., 16 subband energy values and 16 subband kurtosis values).

**Fig. 8.7** Vibration signals measured from roller bearings with different conditions. (a) Signal from a healthy bearing. (b) Signal from a defective bearing (0.1-mm diameter hole). (c) Signal from a defective bearing (0.5-mm diameter hole). (d) Signal from a defective bearing (1-mm diameter hole)



FLD analysis is then applied for feature selection. The means and variances of the feature element,  $f_i$ , are obtained for each of the four bearing conditions. Table 8.4 summarizes the discriminant power of the extracted features for different condition pairs, based on the Fisher discriminant criterion. The first three key features within each condition pair, for example,  $E_4^{12}$ ,  $E_4^{13}$ , and  $E_4^{14}$  for the condition pair (healthy, light defect), are selected, and the final feature set is obtained through a union

**Table 8.4** Discriminant power of the extracted features for various condition pairs in different subbands

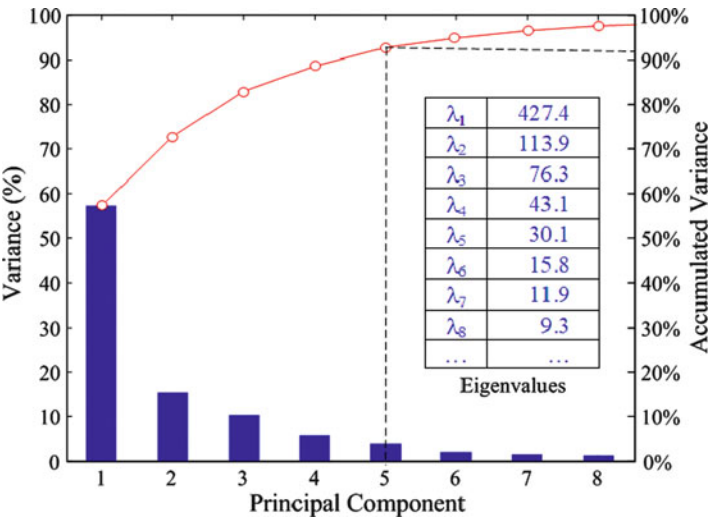
Subband features	Healthy vs. light	Healthy vs. medium	Healthy vs. severe	Light vs. medium	Light vs. severe	Medium vs. severe
$E_4^0$	0.52	2.55	1.80	2.58	2.19	2.15
$E_4^1$	0.06	48.91	8.57	0.43	0.05	10.49
$E_4^2$	18.29	695.40	118.62	41.28	8.37	193.98
$E_4^3$	0.01	110.73	8.86	1.06	0.97	27.86
$E_4^4$	1,222.60	92.10	89.62	1,696.00	216.26	52.50
$E_4^5$	45.20	212.92	125.51	374.81	260.53	11.15
$E_4^6$	41.60	346.97	61.40	2.88	0.08	5.64
$E_4^7$	226.86	440.07	191.23	308.11	88.32	71.01
$E_4^8$	2.41	172.13	7.43	173.16	0.01	183.39
$E_4^9$	87.38	47.61	4.48	69.18	204.47	52.60
$E_4^{10}$	466.15	8.60	211.70	916.98	170.90	340.01
$E_4^{11}$	80.74	14.86	69.41	60.15	15.69	44.79
$E_4^{12}$	5,118.80	1,238.00	133.33	7,946.10	873.16	12.21
$E_4^{13}$	12,702.00	308.19	2,280.10	7,397.60	3,293.60	788.82
$E_4^{14}$	3,652.80	36.25	2,374.40	2,414.90	707.70	1,414.70
$E_4^{15}$	1,863.40	56.36	12,956.60	730.50	86.52	1,150.90
$K_4^0$	2.54	57.50	41.34	0.33	0.10	7.88
$K_4^1$	0.01	0.33	4.04	0.01	0.01	2.30
$K_4^2$	0.02	17.60	7.80	0.03	0.01	16.22
$K_4^3$	0.04	6.98	37.80	0.04	0.03	51.64
$K_4^4$	0.91	39.95	9.99	0.03	0.25	3.07
$K_4^5$	0.34	74.66	50.31	0.18	0.15	1.06
$K_4^6$	0.07	2.33	7.47	0.09	0.10	8.72
$K_4^7$	7.81	30.13	21.94	1.84	1.90	10.41
$K_4^8$	0.98	51.73	6.09	0.05	0.14	0.36
$K_4^9$	0.40	74.16	32.09	0.10	0.18	7.43
$K_4^{10}$	0.25	14.71	9.05	0.02	0.06	0.51
$K_4^{11}$	0.89	15.32	38.21	0.11	0.38	3.09
$K_4^{12}$	0.30	4.62	11.99	0.04	0.02	0.93
$K_4^{13}$	0.02	9.95	2.88	0.01	0.02	0.92
$K_4^{14}$	0.03	16.09	3.15	0.02	0.03	2.39
$K_4^{15}$	0.01	9.42	4.90	0.01	0.01	1.51

operation from all the six condition pairs as listed in Table 8.5, where the energy features  $E_4^2$ ,  $E_4^7$ ,  $E_4^{12}$ ,  $E_4^{13}$ ,  $E_4^{14}$ , and  $E_4^{15}$  are selected as the most representative features, because they possess higher discriminant power than others as listed in Table 8.4.

Next, the PCA technique was performed on the feature vectors. As seen in Fig. 8.8, the first five principal components represent over 90% variance, which preserves most of the information contained in the original feature set (Jolliffe 1986). This is

**Table 8.5** Final feature set obtained through a union operation by Fisher linear discriminant analysis

Subband features	Healthy vs. light	Healthy vs. medium	Healthy vs. severe	Light vs. medium	Light vs. severe	Medium vs. severe	Final feature set
$E_4^2$		✓					✓
$E_4^7$		✓					✓
$E_4^{12}$	✓	✓		✓	✓		✓
$E_4^{13}$	✓		✓	✓	✓	✓	✓
$E_4^{14}$	✓		✓	✓	✓	✓	✓
$E_4^{15}$			✓			✓	✓



**Fig. 8.8** Accumulated variance of principal components for the tested bearings

considered sufficient for constructing the corresponding subspace principle component matrix (based on their corresponding eigenvectors) for choosing the features from the original feature vector. Table 8.6 lists the eigenvectors that correspond to the first five principal components. By searching for those components with the largest magnitude in each eigenvector, the corresponding energy values  $E_4^0$  and  $E_4^3$ , kurtosis values  $K_4^1$ ,  $K_4^3$ , and  $K_4^{15}$  are identified as the most representative features.

The selected feature set was input to a multiplayer perception (MLP) neural network for bearing defect severity classification. Since different ratios (e.g., 70–30 or 50–50) for the training and testing data were suggested for neural network-based classifier in the literature (Paya et al. 1997; Jack and Nandi 2001), but no single fixed ratio has been preferred, two thirds of the data sets corresponding to each condition were used for training the classifier, and the remaining one third for performance checking, from a total of 240 collected data sets. This was aimed at providing sufficient training data to ensure accuracy of the classifier. The classification rates are listed in Table 8.7. When the FLD-selected



**Table 8.6** The first five eigenvectors of the extracted features for the roller bearing

Subband features	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$
$E_4^0$	0.382	0.106	0.605	0.222	0.069
$E_4^1$	0.039	-0.128	0.186	-0.013	-0.039
$E_4^2$	0.060	-0.115	-0.028	-0.388	-0.079
$E_4^3$	0.030	-0.163	0.052	-0.501	0.002
$E_4^4$	-0.045	-0.025	-0.052	-0.133	-0.035
$E_4^5$	-0.078	0.015	-0.194	-0.030	-0.011
$E_4^6$	0.030	0.013	-0.117	-0.094	-0.016
$E_4^7$	-0.042	0.046	-0.195	0.044	0.014
$E_4^8$	-0.068	0.100	-0.225	0.270	0.044
$E_4^9$	-0.098	0.103	-0.252	0.284	0.045
$E_4^{10}$	-0.046	0.012	0.032	0.069	0.001
$E_4^{11}$	-0.085	0.019	0.177	0.185	0.004
$E_4^{12}$	-0.026	0.001	-0.033	-0.012	-0.005
$E_4^{13}$	-0.021	0.001	0.012	0.016	-0.002
$E_4^{14}$	-0.021	0.008	0.015	0.040	0.004
$E_4^{15}$	-0.012	0.009	0.017	0.041	0.004
$K_4^0$	-0.012	-0.063	0.236	0.067	-0.010
$K_4^1$	0.482	-0.445	-0.138	0.337	-0.422
$K_4^2$	0.241	-0.273	-0.088	-0.132	-0.347
$K_4^3$	0.459	-0.156	-0.081	-0.101	0.745
$K_4^4$	0.080	0.017	-0.225	0.100	-0.023
$K_4^5$	0.193	-0.104	-0.176	0.045	0.156
$K_4^6$	0.030	-0.108	-0.080	-0.288	-0.031
$K_4^7$	0.016	-0.035	-0.073	-0.133	-0.010
$K_4^8$	0.036	-0.011	-0.114	0.014	-0.005
$K_4^9$	0.085	-0.001	-0.120	0.044	0.038
$K_4^{10}$	0.059	0.041	-0.136	0.025	-0.019
$K_4^{11}$	0.056	0.063	-0.093	0.010	0.049
$K_4^{12}$	0.033	0.039	-0.295	0.063	0.078
$K_4^{13}$	0.294	0.361	-0.124	-0.113	-0.139
$K_4^{14}$	0.276	0.297	-0.128	0.050	0.056
$K_4^{15}$	0.295	0.598	0.028	-0.208	-0.263

**Table 8.7** Neural-network classifier results of the roller bearing

Classification rate	WPT features with FLD (%)	WPT features with PCA (%)	WPT features only (%)	Raw data features (%)
No defect	100	95	90	85
0.1-mm hole	95	80	70	60
0.5-mm hole	100	95	95	95
1-mm hole	100	100	100	95
Overall	99	92	88	83

feature set was used as input to the MLP classifier, only 5% of measured data with the 0.1-mm hole in the bearing outer raceway is misclassified, out of the whole test data. This led to 98% overall classification success. When the PCA-selected feature set was used as the MLP input, a classification rate of 92% was identified, which is lower than the FLD-selected feature set. The rate, on the contrary, is still higher than the rates obtained using WPT features only as the input to MLP (88%) and using raw data features as the input (83%). This illustrates that the WPT-based feature extraction and selection method is effective in defect severity classification.

### 8.5.2 Case Study II: Ball Bearing Defect Severity Evaluation

For the second case study, a run-to-failure experiment was conducted on a deep groove ball bearing (type 1100KR) of 52 mm outer diameter, under a radial load of 5,498 N. The bearing contained a 0.27-mm wide groove across its outer raceway as an embedded defect, and was continually run under a rotational speed of 2,000 rpm. Upon reaching approximately 2.7 million revolutions, the defect has propagated throughout the entire raceway and rendered the bearing practically nonfunctional. This case study was designed to investigate the effect of continuous degradation of the defect, whereas case study I discussed above concerns with the effect of discrete defects.

Vibration signals were collected during the experiment at an interval of every 7 min. Figure 8.9 illustrates the trend of the vibration amplitude along the process of defect propagation. Three vibration signals are also shown in Fig. 8.10, and they are measured right after the bearing is physically examined at different test stages. For

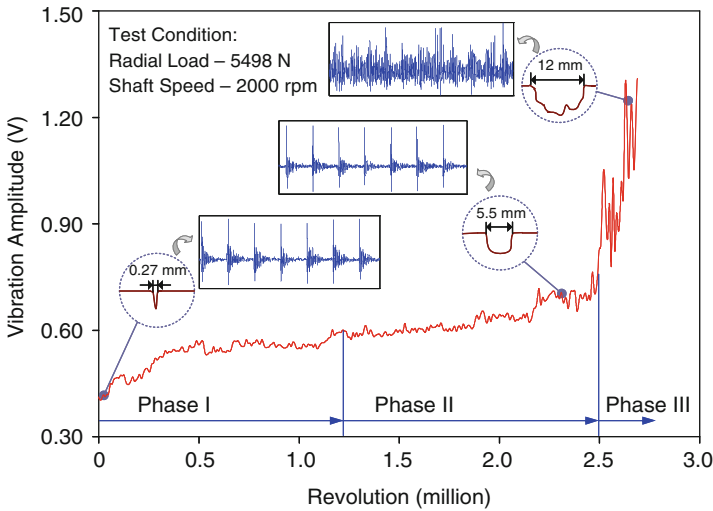
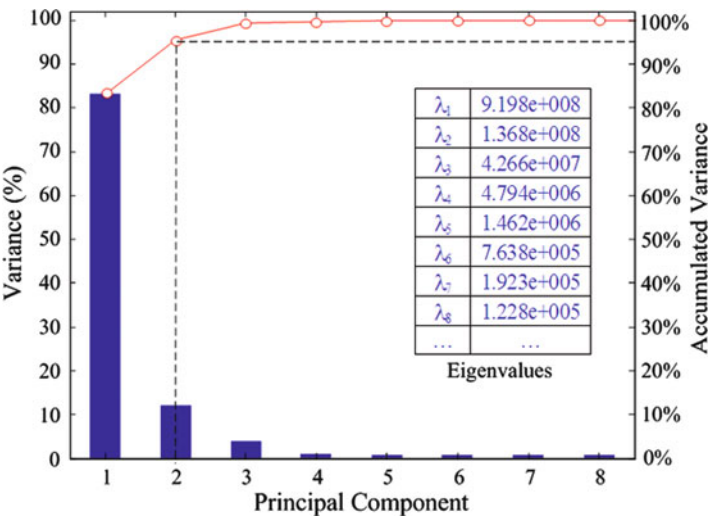


Fig. 8.9 Amplitude as a function of the ball bearing revolution



**Fig. 8.10** Accumulated variance of principal components for the ball bearing 1100KR

purpose of defect severity evaluation, all the collected vibration data sets are divided into three sections with the threshold values of the amplitude being set at 0.6 and 0.9 V, respectively. As shown in Fig. 8.9, the three sections correspond to three different defect propagation phases during the run-to-failure test. It should be noted that, since no prior knowledge was available regarding the relationship between the vibration amplitude and the defect severity level, the choice of three phases investigated here is empirical.

The vibration signals were first decomposed into 16 subbands. The energy and kurtosis features were then calculated from the wavelet packet coefficients in each subband to formulate the feature vectors. FLD analysis was then used for feature selection. The means and variances of the feature element,  $f_i$ , were obtained for each of the four bearing initial conditions. Table 8.8 summarizes the discriminant power of the extracted features for different phase pairs, based on the Fisher discriminant criterion. The first three key features within each phase pair were selected and the final feature set was obtained through a union operation among different phase pairs (i.e., phase I, phase II; phase I, phase III; and phase II, phase III). As listed in Table 8.9, the kurtosis features  $K_4^0$ ,  $K_4^1$ ,  $K_4^2$ ,  $K_4^6$ , and  $K_4^{10}$  are selected as the most representative features.

When the PCA was performed on the extracted subband feature vectors, the first two principal components shown in Fig. 8.10 represent over 90% variance, which was subsequently used as the reference features from the original feature set. Table 8.10 lists the eigenvectors corresponding to the first two principal components. The highest magnitude in the first eigenvector was found to be associated with the first component, and the highest magnitude in the second eigenvector was seen to be related to the third component. Accordingly, the

**Table 8.8** Discriminant power of the extracted features for the ball bearing phase pair in various subbands

Subband features	Phase I vs. phase II	Phase I vs. phase III	Phase II vs. phase III
$E_4^0$	$1.08 \times 10^{-8}$	$1.25 \times 10^{-8}$	$2.66 \times 10^{-9}$
$E_4^1$	$1.30 \times 10^{-8}$	$1.15 \times 10^{-8}$	$2.65 \times 10^{-8}$
$E_4^2$	$2.23 \times 10^{-9}$	$1.72 \times 10^{-8}$	$5.18 \times 10^{-8}$
$E_4^3$	$6.13 \times 10^{-7}$	$6.81 \times 10^{-7}$	$4.38 \times 10^{-9}$
$E_4^4$	$3.54 \times 10^{-7}$	$4.37 \times 10^{-8}$	$4.05 \times 10^{-10}$
$E_4^5$	$5.54 \times 10^{-7}$	$2.54 \times 10^{-7}$	$2.90 \times 10^{-10}$
$E_4^6$	$1.81 \times 10^{-8}$	$5.13 \times 10^{-9}$	$7.44 \times 10^{-8}$
$E_4^7$	$1.82 \times 10^{-5}$	$1.42 \times 10^{-6}$	$7.26 \times 10^{-7}$
$E_4^8$	$9.94 \times 10^{-7}$	$5.05 \times 10^{-7}$	$9.64 \times 10^{-7}$
$E_4^9$	$5.54 \times 10^{-6}$	$3.23 \times 10^{-6}$	$5.07 \times 10^{-6}$
$E_4^{10}$	$5.59 \times 10^{-4}$	$1.13 \times 10^{-6}$	$3.62 \times 10^{-7}$
$E_4^{11}$	$2.72 \times 10^{-3}$	$1.49 \times 10^{-5}$	$1.69 \times 10^{-6}$
$E_4^{12}$	$2.15 \times 10^{-6}$	$1.80 \times 10^{-7}$	$8.29 \times 10^{-9}$
$E_4^{13}$	$5.39 \times 10^{-6}$	$8.77 \times 10^{-7}$	$9.99 \times 10^{-8}$
$E_4^{14}$	$7.55 \times 10^{-4}$	$3.40 \times 10^{-6}$	$2.26 \times 10^{-6}$
$E_4^{15}$	$1.24 \times 10^{-3}$	$1.75 \times 10^{-5}$	$9.36 \times 10^{-6}$
$K_4^0$	$2.07 \times 10^{-1}$	$2.54\text{E}+01$	$1.75\text{E}+00$
$K_4^1$	$8.18 \times 10^{-2}$	$4.69 \times 10^{-1}$	$5.16 \times 10^{-1}$
$K_4^2$	$2.35 \times 10^{-3}$	$1.39\text{E}+00$	$3.14 \times 10^{-1}$
$K_4^3$	$9.96 \times 10^{-5}$	$4.10 \times 10^{-2}$	$4.11 \times 10^{-2}$
$K_4^4$	$7.98 \times 10^{-4}$	$1.93 \times 10^{-5}$	$1.34 \times 10^{-5}$
$K_4^5$	$3.72 \times 10^{-4}$	$4.19 \times 10^{-3}$	$4.61 \times 10^{-3}$
$K_4^6$	$1.22 \times 10^{-3}$	$9.56 \times 10^{-1}$	$2.37 \times 10^{-1}$
$K_4^7$	$6.30 \times 10^{-6}$	$4.94 \times 10^{-2}$	$9.81 \times 10^{-3}$
$K_4^8$	$3.26 \times 10^{-4}$	$2.18 \times 10^{-6}$	$2.76 \times 10^{-6}$
$K_4^9$	$2.85 \times 10^{-5}$	$9.04 \times 10^{-7}$	$9.47 \times 10^{-7}$
$K_4^{10}$	$3.05 \times 10^{-3}$	$1.05 \times 10^{-6}$	$8.54 \times 10^{-7}$
$K_4^{11}$	$1.57 \times 10^{-4}$	$6.45 \times 10^{-7}$	$7.02 \times 10^{-7}$
$K_4^{12}$	$1.13 \times 10^{-3}$	$3.02 \times 10^{-5}$	$2.23 \times 10^{-5}$
$K_4^{13}$	$5.42 \times 10^{-4}$	$3.45 \times 10^{-6}$	$2.04 \times 10^{-7}$
$K_4^{14}$	$2.64 \times 10^{-4}$	$5.86 \times 10^{-6}$	$3.50 \times 10^{-6}$
$K_4^{15}$	$7.76 \times 10^{-4}$	$6.13 \times 10^{-5}$	$9.54 \times 10^{-6}$

**Table 8.9** Final feature set obtained for the ball bearing through a union operation by FLD

Subband features	Phase I vs. phase II	Phase I vs. phase III	Phase II vs. phase III	Final feature set
$K_4^0$	✓	✓	✓	✓
$K_4^1$	✓		✓	✓
$K_4^2$		✓	✓	✓
$K_4^6$		✓		✓
$K_4^{10}$	✓			✓

**Table 8.10** The first two eigenvectors calculated the extracted features for the ball bearing

Subband features	$a_1$	$a_2$
$E_4^0$	-0.997	-0.022
$E_4^1$	-0.033	-0.229
$E_4^2$	0.024	-0.942
$E_4^3$	0.027	-0.037
$E_4^4$	0.027	0.012
$E_4^5$	0.023	-0.001
$E_4^6$	0.029	-0.241
$E_4^7$	0.021	0.004
$E_4^8$	-0.001	0.001
$E_4^9$	0.001	-0.001
$E_4^{10}$	0.003	-0.002
$E_4^{11}$	0.001	-0.001
$E_4^{12}$	0.012	0.006
$E_4^{13}$	0.006	0.004
$E_4^{14}$	0.003	0.002
$E_4^{15}$	0.003	0.002
$K_4^0$	-0.001	-0.001
$K_4^1$	-0.001	-0.001
$K_4^2$	-0.001	-0.001
$K_4^3$	-0.001	-0.001
$K_4^4$	0.001	0.001
$K_4^5$	-0.001	-0.001
$K_4^6$	-0.001	-0.001
$K_4^7$	-0.001	-0.001
$K_4^8$	0.001	0.001
$K_4^9$	0.001	0.001
$K_4^{10}$	0.001	0.001
$K_4^{11}$	0.001	0.001
$K_4^{12}$	0.001	0.001
$K_4^{13}$	-0.001	-0.001
$K_4^{14}$	0.001	-0.001
$K_4^{15}$	-0.001	-0.001

**Table 8.11** Results of neural-network classification rate of results of the ball bearing

Classification rate	WPT features with FLD (%)	WPT features with PCA (%)	WPT features only (%)	Raw data features (%)
Phase I	92	87	82	79
Phase II	91	84	81	77
Phase III	94	88	88	82
Overall	92	86	82	78

energy value at subbands 1 and 3 were (denoted as  $E_4^0$  and  $E_4^2$ ) identified as the most representative features.

Following the same procedure as described in case study I, 2/3 of the data sets corresponding to each defect propagation phase are used for training the MLP classifier, and the remaining 1/3 data points are used for performance checking. As shown in Table 8.11, when the features selected from the FLD approach were used as input to the MLP classifier, the classification rate for each phase is found to be 92%, 91%, and 94%, respectively. This led to the overall classification rate of 92%. In comparison, when features selected using PCA technique were used as input to the MLP, the classification rates were lower, 87%, 84%, and 88%, respectively. Furthermore, when feature set extracted from each subband and raw data was directly used as the MLP input, the classification rates dropped down to even lower values (e.g., 82% overall classification rate for WPT features only, and 78% overall classification rate for raw data features). This indicates again the effectiveness of the presented approach for defect severity classification.

## 8.6 Summary

This chapter introduces a wavelet packet-based signal processing approach for machine defect severity classification. After the subband energy and kurtosis features are extracted from realistic vibration signals using the wavelet-packet coefficients, the most representative features are chosen using the Fisher discriminant criterion and principal feature analysis, respectively. These features are used as inputs to the neural-network classifiers to evaluate the machine defect severity. The effectiveness of the approach has been experimentally verified through two case studies for rolling bearing defect severity classification. It is shown that the introduced approach provides a practical way for feature extraction and selection. In addition to bearing defect severity classification, this approach is applicable to classifying the working states of other machines and machine components, thus providing a useful tool for machine condition monitoring and diagnosis.

## 8.7 References

- Altmann J, Mathew J (2001) Multiple band-pass autoregressive demodulation for rolling-element bearing fault diagnosis. *Mech Syst Signal Process* 15:963–977
- Baydar N, Chen Q, Ball A, Kruger U (2001) Detection of incipient tooth defect in helical gears using multivariate statistics. *Mech Syst Signal Process* 15:303–321
- De Boe P, Golinval JC (2003) Principal component analysis of a piezo-sensor array for damage localization. *Int J Struct Health Monit* 2(2):137–144
- Duda R, Hart P, Stork D (2000) *Pattern classification*. Wiley-Interscience, New York.
- Fan X, Zuo MJ (2006) Gearbox fault detection using Hilbert and wavelet packet transform. *Mech Syst Signal Process* 20:966–982

- Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic, New York
- Gao R, Yan R (2007) Wavelet packet transform-based hybrid signal processing for machine health monitoring and diagnosis. In: The 6th international workshop on structural health monitoring, Stanford, CA, pp 598–605
- Goumas SK, Zervakis ME, Stavrakakis GS (2002) Classification of washing machine vibration signals using discrete wavelet analysis for feature extraction. *IEEE Trans Instrum Meas* 51 (3):497–508
- Haykin, S (1994) Neural networks. Macmillan Publishing Company, New York
- He Q, Yan R, Kong F, Du R (2008) Machine condition monitoring using principle component representation. *Mech Syst Signal Process.* 23(2):446–466
- Jack LB, Nandi AK (2001) Support vector machines for detection and characterization of rolling element bearing faults. *Proc Inst Mech Eng* 215:1065–1074
- Jolliffe IT (1986) Principal component analysis. Springer-Verlag New York Inc, New York
- Kano M, Hasebe S, Hashimoto I (2001) A new multivariate statistical process monitoring method using principal component analysis. *Comput Chem Eng* 25:1103–1113
- Kittler J (1975) Mathematical methods of feature selection in pattern recognition. *Int J Man Mach Stud* 7(5):609–637
- Lee BY, Tang YS (1999) Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current. *Int J Adv Manuf Technol* 15(4):238–243
- Li B, Chow M, Tipsuwan Y, Hung JC (2000a) Neural-network-based motor rolling bearing fault diagnosis. *IEEE Trans Ind Electron* 47(5):1060–1069
- Li XL, Tso SK, Wang J (2000b) Real-time tool condition monitoring using wavelet transforms and fuzzy techniques. *IEEE Trans Syst Man Cybern C Appl Rev* 30(3):352–357
- Liu B, Ling SF, Meng Q (1997) Machinery diagnosis based on wavelet packets. *J Vib Control* 3:5–17
- Maki Y, Loparo KA (1997) A neural-network approach to fault detection and diagnosis in industrial processes. *IEEE Trans Control Syst Technol* 5(6):529–541
- Malhi A, Gao R. (2004) PCA-based feature selection scheme for machine defect classification. *IEEE Trans Instrum Meas* 53(6):1517–1525
- McCormick AC, Nandi AK (1997) Classification of the rotating machine condition using artificial neural networks. *Proc Inst Mech Eng C* 211:439–450
- Mori K, Kasashima N, Yoshioka T, Ueno Y (1996) Prediction of spalling on a ball bearing by applying the discrete wavelet transform to vibration signals. *Wear* 195:162–168
- Paya BA, Esat II, Badi MNM (1997) Artificial neural network based fault diagnosis of rotating machinery using wavelet transforms as a preprocessor. *Mech Syst Signal Process* 11 (5):751–765
- Prabhakar S, Mohanty AR, Sekhar AS (2002) Application of discrete wavelet transform for detection of ball bearing race faults. *Tribol Int* 35(12):793–800
- Yan R, Gao R (2004) Harmonic wavelet packet transform for on-line system health diagnosis. SPIE international symposium on sensors and smart structures technologies for civil, mechanical and aerospace systems, San Diego, CA, pp 512–522
- Yen G, Lin K (2000) Wavelet packet feature extraction for vibration monitoring. *IEEE Trans Ind Electron* 47(3):650–667