Feature Generation Using Genetic Programming With Application to Fault Classification

Hong Guo, Lindsay B. Jack, and Asoke K. Nandi, *Senior Member, IEEE*

***Abstract*—One of the major challenges in pattern recognition problems is the feature extraction process which derives new features from existing features, or directly from raw data in order to reduce the cost of computation during the classification process, while improving classifier efficiency. Most current feature extraction techniques transform the original pattern vector into a new vector with increased discrimination capability but lower dimensionality. This is conducted within a predefined feature space, and thus, has limited searching power. Genetic programming (GP) can generate new features from the original dataset without prior knowledge of the probabilistic distribution. In this paper, a GP-based approach is developed for feature extraction from raw vibration data recorded from a rotating machine with six different conditions. The created features are then used as the inputs to a neural classifier for the identification of six bearing conditions. Experimental results demonstrate the ability of GP to discover autimatically the different bearing conditions using features expressed in the form of nonlinear functions. Furthermore, four sets of results—using GP extracted features with artificial neural networks (ANN) and support vector machines (SVM), as well as traditional features with ANN and SVM—have been obtained. This GP-based approach is used for bearing fault classification for the first time and exhibits superior searching power over other techniques. Additionaly, it significantly reduces the time for computation compared with genetic algorithm (GA), therefore, makes a more practical realization of the solution.**

***Index Terms*—Fault classification, feature generation, genetic programming (GP), machine condition monitoring (MCM).**

I. Introduction

MACHINE condition monitoring (MCM) is an area which is gaining increasing importance in the manufacturing industry. Maintenance costs can be reduced significantly by monitoring health of machinery. Potentially disastrous faults can be detected early, while enabling the implementation of condition-based maintenance rather than periodic, or responsive maintenance. Several conventional methods have been used to analyze the vibration signal in order to extract effective features for bearing fault detection. These include probabilistic analysis [1], [2], frequency analysis [3], time-domain [4], and finite-element analysis [1], [2].

Feature extraction is one of the most important factors in pattern recognition problems. The process derives new features from raw data in order to reduce the dimensionality of data presented to the classifier, while improving the classification efficiency. The choice of features can greatly affect the performance of classification. Any generated features will often be refined to try to achieve the desired level of performance. However, manually developing features can be very time-consuming and rely on the experience of the engineer. In many problems, generated features should have the ability to identify subtle, or complex relationships, within large datasets where the mapping from data to class labels is often obscure, or difficult for humans to identify. A variety of different machine learning techniques have been applied to the problem of automated feature generation including neural networks [5], [6], fuzzy systems [7], [8], and evolutionary algorithms [9], [10].

In recent years, the application of evolutionary learning algorithms to pattern recognition problem has become increasingly common. Evolution strategies [11], evolutionary programming (EP) [12], genetic algorithms (GA) [13]–[16], or genetic programming (GP) [17]–[22] have been used to solve complex problems. For instance, Raymer *et al.* [15] shows that a hybrid of KNN classifier and genetic algorithm can considerably improve the discrimination accuracy. GP was first introduced by Koza [17], and has been proposed as a machine learning method in different fields. In [18], the GP technique was used to develop a decision support system for vehicle dispatching considering a population of utility functions that evaluate candidate vehicles for servicing requests. GP was tested in six medical diagnosis problems [19] and the results were compared with those obtained by neural networks. In [20], the feasibility of applying GP to multicategory pattern classification problem was studied. Zhang *et al.* [21] applied GP for fault detection in MCM field. However, in all the above applications [18]–[21], GP was employed solely as a classifier based on manually developed features. In [22], GP-based feature extraction was used to improve the classification results and reduce the dimensionality of the data in the medical domain. GP exhibits pseudo-intelligent behavior by deciding whether to perform feature extraction or feature selection during the evolutionary process. Unfortunately, the system is unable to sample adequately the search space for high-dimensional problems and the main disadvantage lies in its computational complexity. Kotani *et al.* [23] performed feature extraction using GP with a KNN classifier on one artificial task and one acoustic diagnosis experiment with the conclusion that the GP is an effective tool for the feature extraction task.

Some techniques for feature extraction have been studied for MCM application to data. Genetic algorithm based feature selection was carried out in [14], for the classification of bearing faults using vibration signals. In [24], Chen *et al.* presented a GA-based method to automatically generate symptom parameter functions from rolling bearing data for the diagnosis

Manuscript received January 19, 2004; revised June 22, 2004. The work of H. Guo was supported by the Overseas Research Studentship Committee, U.K., and the University of Liverpool Graduates Association (HK). The work of L. B. Jack was supported by the Biotechnology and Biological Sciences Research Council. This paper was recommended by Associate Editor H. Takagi.

The authors are with the Signal Processing and Communications Group, Department of Electrical Engineering and Electronics, The University of Liverpool, Liverpool, L69 3GJ, U.K. (e-mail: a.nandi@liv.ac.uk).

Digital Object Identifier 10.1109/TSMCB.2004.841426

of machinery operating conditions. Experimental studies on rolling bearings using a feature extraction method which is a combination of wavelet and Fourier transformation are reported in [25]. As far as the authors are aware, GP has not yet been utilized for the purpose of feature generation in the field of bearing fault classification.

In this paper, GP is applied to generate features suitable for bearing fault classification in the area of MCM for the first time. It has the capability to extract features from raw vibration data and uses these features to improve the classification performance with the following advantages:

- All features within each generation are created automatically, thus, avoiding human influence or bias.
- Instead of using classification results to determine the fitness of a feature, usually requiring a large amount of computation and time, a novel fast method is proposed to evaluate the difference among classes and enhance the distribution of each class from others based on the well-known Fisher criterion.
- GP exhibits pseudo-intelligent behavior by deciding whether to perform feature extraction or feature selection during the evolutionary process, rather than the pure selection of features from a large number of candidate features using GA.

This paper is organized as follows. The data preparation using vibration signals for condition monitoring is addressed in Section II. The traditional feature extraction methods are presented at Section III. GP–based feature generation model is described in Section IV, while types of artificial neural network (ANN) used for the classification task in this work are described in Section V. Based on the model, a series of feature extraction experiments for the classification of vibration signals are conducted in Section VI, which also includes comparisons of classification performance using extracted features by GP and traditional methods. Section VII provides some discussions based on experimental results. Advantages and limitation of the GP-based feature extraction method are concluded in Section VIII.

II. Data Preparation

Rolling element bearings (see Fig. 1) are probably among the most widely used rotating machine components. It is of prime importance to be able to accurately detect the existence and severity of faults in machinery in certain areas of industry, as the machine may be safety or emergency-related in many cases. The work presented in this paper is the automatic extraction of features using GP and the employment of those features to solve the problem of bearing conditions monitoring. The six bearing conditions, each having their own distinguishing characteristics, are as follows.

1) normal bearing (NO);
2) worn normal bearing (NW);
3) inner race fault (IR);
4) outer race fault (OR);
5) rolling element fault (RE);
6) cage fault (CA).

The NO is a brand new bearing, which has been run in, but in otherwise perfect condition. The NW is in good condition, however, it has been running for some period of time and serves as an example of a bearing that has seen some usage. The IR fault

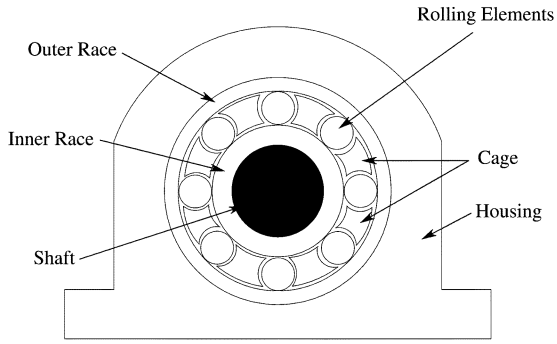

Fig. 1. Typical roller bearing, showing different component parts.

was created by first removing the cage, moving the elements to one side of the bearing, and then removing the inner race. A groove was then cut in the raceway of the inner race, using a small grinding stone, and the bearing was reassembled. The OR fault was created by removing the cage, pushing all the balls to one side, and then inserting a small grinding stone which created the outer race fault, and cutting a small groove in the outer raceway. The RE fault was induced by using an electrical etcher to mark the surface of one of the balls, simulating corrosion. The CA fault was simply created by removing the plastic cage from one of the bearings, cutting away a section of the cage, so that two of the balls were free to move, and not held at a regular spacing, as would normally be the case.

A. Data Acquisition

In order to simulate commonly occurring faults in rotating machinery, experimental data was collected from a test rig (Fig. 2), which consisted of a dc motor driving the shaft through a flexible coupling, with the shaft supported by two plummer bearing blocks. The damaged bearings were inserted into one of the plummer blocks, and the resultant vibrations in the horizontal and vertical planes were measured using two accelerometers. The output from the accelerometers were sampled at a rate of 24 kHz, giving a slight over-sampling of the data.

B. Time Domain Characteristics

By examining the actual vibration plots on a time series basis (Fig. 3), some characteristics can be found. First, signals from four conditions, including NO, NW, CA, and OR, look similar with amplitude not exceeding ± 100, while the other two conditions have periodic strong pulsations. These two groups can be easily differentiated by examining the amplitude. The two normal conditions look similar, though the signal from the worn condition is a bit noisier than that from a brand new bearing. The outer race fault and cage fault display little difference to the normal condition in terms of magnitude and noise level. Therefore, it will be difficult to identify these conditions solely by time series inspections.

C. Experimental Datasets

Experimental datasets were formed by running the machine (Fig. 2) over a series of sixteen different speeds and taking ten examples of data at each speed. Each example consists of 2000 data samples. This gives a total of 160 examples of each condition and a total of 960 raw data examples over six conditions to

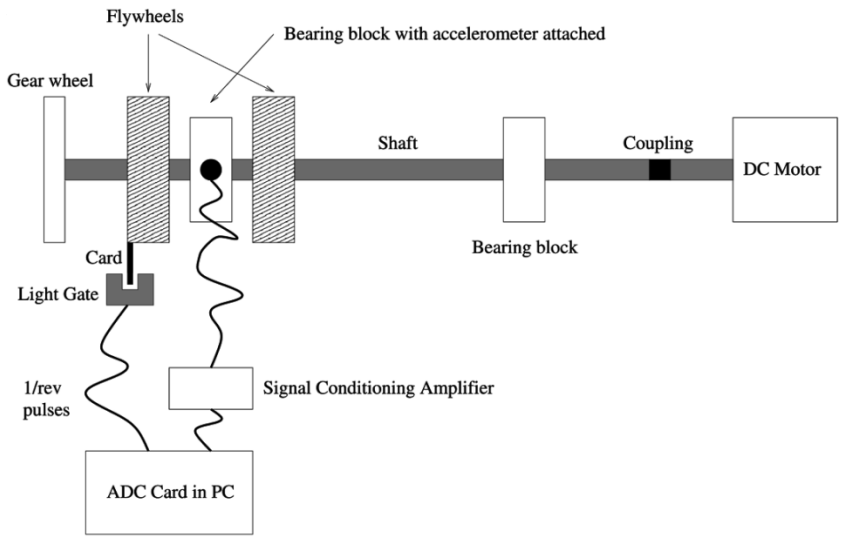

Fig. 2. Machine test rig used in experiments.

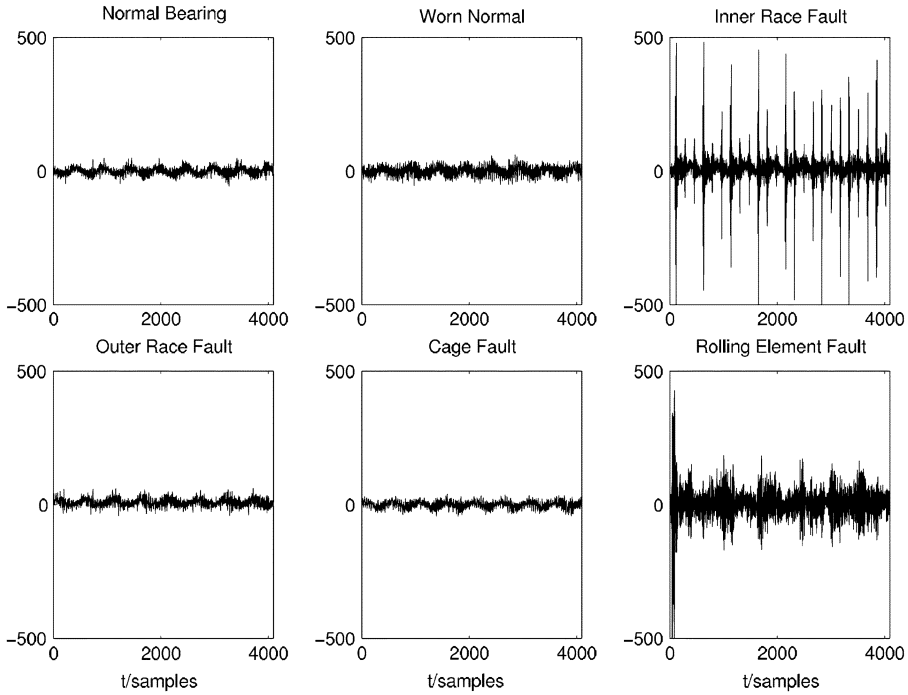

Fig. 3. Typical vibration signals for six conditions.

work with. The full input dataset creates a 960×2000 matrix as the training dataset. The other two 960×2000 matrices are the validation dataset and the test dataset, respectively. For each given vector in the input raw vibration datasets, a corresponding vector was created in a matrix containing the target information used during the experiment.

III. Traditional Feature Extraction Methods

Feature extraction is an important task in pattern recognition and classification problems, especially for MCM, where only well-chosen features provide discrimination information, and thus, can help identify subtle changes in the machine condition. Traditionally, to obtain an accurate measure of the condition, machine vibration signals are recorded and transformed into some indicator, which may be a measure of the energy, or the magnitude of a particular frequency due to a fault. This process is called the feature extraction. After evaluation and comparison, useful features are picked out for the further classification of different conditions. In this section, conventional features and a list of possibly useful features are described in detail.

A. Conventional Measures

A number of statistically-based performance indicators exist, which provide single figure assessments of the condition of rolling element bearings. These give an indication of whether a bearing is in a state of distress, or within normal operating parameters, and show the degree of distress that a bearing is

under. Conventionally, three most common measurements used are shock pulse (SP), crest factor (CF), and kurtosis [14].

The SP method is a signal processing technique used to measure impact and noise caused by metal to metal contact in the bearing. It is much more refined than other high frequency measurements. Shock pulse analysis relies on a specialized transducer, which has a resonant frequency of 32–36 kHz. The amplitude of the SP is relative to the velocity of the impact. For bearing conditions, carpet value and max value are two readings of the SP, and thus can be used as two components of four conventional features.

1) Carpet Value: Metal impacting on metal always occurs in rolling element bearings, even a brand new bearing under normal operating conditions. When there is no damage within the bearing, the metal-to-metal contact creates a background noise of SP. This is referred to as the carpet value. Carpet value decreases when the bearing is well lubricated. When damage occurs to the bearing there will be more metal to metal contact, which is reflected by an increasing carpet value. By examining the carpet values in the signal, information can be gleaned about the likelihood of the existence of a defect.

2) Max Value: Max value is another conventional parameter to identify the damage in the rolling element bearings. However, it is not capable of distinguishing different fault conditions of bearings. When a fault occurs within a bearing element, the rolling elements strike the defect periodically. These create a high amplitude SP. The max value increases as the bearing damage develops further. This peak from the carpet of background shock signals can be used to detect the damage in applications of bearing condition monitoring.

3) Crest Factor: The crest factor (CF) is a commonly used measure for the detection of bearing faults. CF is equal to the peak amplitude of the waveform divided by the root mean square (RMS) value. The analysis of the CF can give an idea of how much impacting is occurring in time domain. The impacting is associated with the rolling bearings. The CF is relatively high due to the amount of the impact occurring within the bearing and it works well while the fault is developing. However, as the degree of damage goes up, the high frequency component of a vibration signal increases. Hence, the RMS value increases with the result that the CF value decreases. Generally, the CF is trended over time to see whether the amount of the impacting is decreasing or not.

4) Kurtosis: Kurtosis is a time-domain analysis technique. It is commonly used as a measure of damage. The definition of kurtosis is given by

$$\text{kurt} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{x_j - \bar{x}}{\sigma}\right)^4 \tag{1}$$

where x_j represents a vibration sample, N is the total number of samples, $\bar{x}$ is the average of N vibration samples, and σ is their standard deviation. The kurtosis value emphasises the length of the "tails" of a distribution. Signals show a lot of sharp impacts when the rolling elements of a bearing strike a defect, the value of kurtosis will be high. The kurtosis value will be low while signals have little or no spike content.

B. Plain Statistics

It is well known that vibration signals depend mainly on the resonant frequencies of different parts of the machine. If the machine condition varies due to wear or damage, the resonant frequencies, and hence, the vibrations, will change. Also, it is generally not possible to classify the condition based upon an individual sample of the vibration, thus, some transformation of the recorded vibration time-series is required to extract time-invariant features. These are statistical moments and cumulants. For example, as the machine's condition deteriorates, the energy (mean square value) in the vibration signal is expected to increase. A number of different statistical features were generated using moments and cumulants of the vibration data. The nth-order moment is defined by (2). The four statistical features used here are the four (first to fourth order) moments. These are stored in a matrix of size 4×960.

$$m_x^{(k)} = \frac{1}{N}\sum_{i=1}^{N} x_i^k \tag{2}$$

for $k = 1, 2, 3,$ and 4.

C. Signal Difference and Sums

Differences highlight the high-frequency components in the signal, and the sums of the signal emphasise the low-frequency portions. The numerical derivative of each vibration signal was calculated by (3). The four plain statistical features were calculated from the derivatives. The results were saved in another 4×960 matrix. The numerical integral of vibration signal were given by (4). In the same way, this creates another 4×960 matrix.

$$d(n) = x(n) - x(n-1) \tag{3}$$

$$i(n) = \left\{x(n) - m_x^{(1)}\right\} + i(n-1). \tag{4}$$

D. High and Low Filtering

The four statistical features were calculated on data filtered using an eighth-order Butterworth IIR high pass filter with a cut-off frequency of 129 Hz; this gave another 4×960 matrix. A low-pass filter with the same cut-off frequency was used on the same datasets, and gave a 4×960 matrix.

E. Normalization

The importance of normalization to both the efficiency and accuracy has been demonstrated [26]. The normalization in experiments is based on (5)

$$f_i' = \frac{f_i - m_f^{(1)}}{\sigma_f} \tag{5}$$

where $m_f^{(1)}$ is the mean value of the feature vector f and σ_f is the standard deviation of the feature vector f.

IV. GP-Based Feature Generation Model

In this paper, GP, as a form of evolutionary algorithm and an extension of genetic algorithms, is proposed as the primary

method for the feature extraction/generation. The major difference between the GP and GA approaches lies in the way that each algorithm solves the problem under consideration. With a GA-based solution, the basic form of the solution is predefined; the GA is able to optimize parameters of the solution, however not the actual structure of the solution. GP by comparison has control over both the structure and the parameters of the solution to the problem.

Fig. 4 illustrates the system proposed in this paper. The block with a bold frame is the feature generator, which extracts the information from the raw vibration data to create features, based on the evolutionary algorithm. The surviving features from the feature generator are used as the inputs to the multilayer perceptron (MLP) for the classification of the six bearing conditions. Of course, some other classifier can be used as an alternative.

A. Process of GP

Since the aim of bearing fault classification is to identify different machine conditions from raw vibration signals accurately, the GP-based feature extractor is used to extract useful information from the raw vibration data in order to provide discriminating input features for the classifiers. The purpose of GP is to try to maximize the extra information content in the sample of the raw vibration signal, and it implicitly maximizes the separation between different conditions within the data. The evolutionary process of GP-based feature generation system is illustrated in Fig. 5. First, an initial population with a chosen number of individuals is generated on a random basis, meaning that there is no human influence, or bias, in the generation of original features. Raw bearing data recorded from experimental machines are fed as the inputs to the initial population. Each individual represents a transformation network, which tries to transform raw data into information for classification.

In terms of the usefulness of each individual for classification, a fitness value is assigned to each individual by fitness function. Therefore, the members with the best fitness values survive from the current generation and will be chosen as the origins of the next generation. In our design, only the elite will survive the natural selection. This mechanism allows the feature to evolve in a direction toward the best classification performance, thus achieving the automatic generation of features. At the beginning of the next generation, three operations—reproduction, crossover, and mutation—are conducted to produce new members based on the surviving member. If the termination criterion is met, the best solution is preserved.

B. Fitness Function

As one of the most important components of GP, the fitness function can greatly affect the performance of the system. A good fitness measure guarantees the improvement of solutions by rating the performance of each member and giving the stronger one a better chance of surviving. Traditionally, the classification results are used as the fitness value for multicategory classification problem; however, the computational demands are relatively high in training and validating a classifier for each individual.

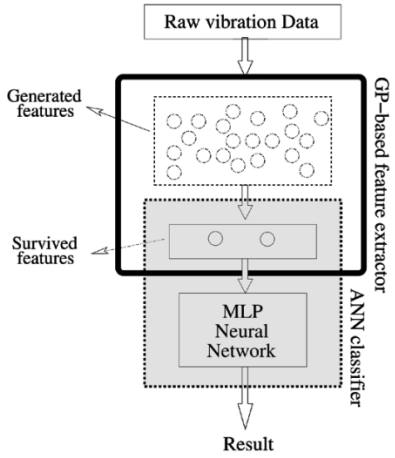

Fig. 4. Structure of the system.

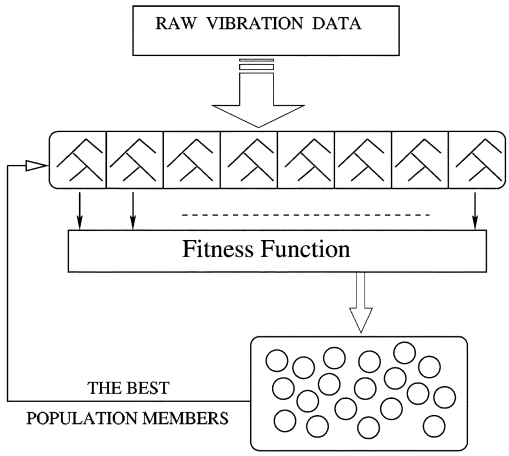

Fig. 5. Evolutionary process of GP-based feature generation system.

The Fisher criterion is adopted for the solution of the feature extraction problem based on the maximization of inter-class scatter over the intra-class scatter. In this paper, GP as a evolutionary method is proposed to maximize the degree of difference between two classes, analogous to the Fisher criterion, but in an iterative process. The following expression as the fitness function for two classes is obtained based on the Fisher criterion

$$f_{ij} = \frac{\left|\frac{1}{n}\sum_{i=1}^{n} S_i - \frac{1}{m}\sum_{j=1}^{m} S_j\right|}{\sqrt{\frac{\sum_{i=1}^{n}\left(S_i - \frac{1}{n}\sum_{i=1}^{n} S_i\right)^2}{n-1} + \frac{\sum_{j=1}^{m}\left(S_j - \frac{1}{m}\sum_{j=1}^{m} S_j\right)^2}{m-1}}} \tag{6}$$

where S represents the examples of bearing conditions, the numerator denotes the distance between class i and j while the denominator denotes the range of variance within class i and j.

Note that the n-class $(n > 2)$ Fisher criterion can be decomposed into $k(k = {}^{n}\mathrm{C}_2)$ two-class Fisher criteria. In order to solve the worst case in k two-class, the fitness value is defined to be predominantly determined by the minimum one of k two-class Fisher criteria, each of which measures the distribution of inter-class scatter over the classes intra-class scatter for any two-class. Considering the overall distribution of n classes, the weighted mean of k two-class Fisher criteria will contribute

to the fitness function by selecting the best feature for n classes. In an n-dimensional space, the fitness measure is defined as

$$F = \min(\vec{f}) + \lambda \cdot \frac{1}{k} \sum_{i=1}^{k} (\vec{f}) \tag{7}$$

where $\vec{f}$ denotes the vector, which consists of $k(k = {}^{n}_{1}\mathrm{C}_{2})$ two-class Fisher criteria among n classes, and λ is an empirical factor that accepts the contribution from mean value and at the same time diminishes the effect of too large a mean value. The purpose of taking the average is to take into account the distribution of conditions rather than the worst-separated two classes. Consequently, the feature with few large criteria values but small minimum value cannot compete with average criteria values but relatively large minimum value. On the other hand, if the minimum values for two features are similar, the one with the largest average of values will survive. Specifically, λ is chosen equal to 0.001. Overall, the individual having high fitness value means that difference between any two conditions, even the closest classes, is large.

C. Primitive Operations

GP evolves tree individuals representing possible solutions to the problem at hand. A population of such individuals is randomly created and then evolved by probability of genetic operations:

- Crossover: GP carries out a crossover operation to create new individuals with a probability P_c, which controls the occurrence of the crossover throughout generations. Two new individuals are generated by selecting compatible nodes randomly from each parent and swapping them, as illustrated in Fig. 6
- Mutation: The mutation operation is performed by the creation of a subtree at a randomly selected node with the probability P_m. First, for a given parent, there is an index assigned to each node for identification. A random index number is generated to indicate the place where mutation will happen. The node is located, then the tree downstream from this node is deleted and a new subtree is generated from this node (Fig. 7), exactly in the same way as growing initial population.
- Reproduction: The reproduction operation is performed by copying individuals to the next population without any change in terms of a certain probability P_r.

All these three operations happen within one generation based on three probabilities

$$P_c + P_m + P_r = 1. \tag{8}$$

D. Primitive Terminators

Terminators act as the interface between GP and the raw vibration signal. They are required to collect fault-related information as much as possible from the raw vibration data and to provide inputs to the feature extractor.

In our GP-based feature extractor, the terminator set is constructed by computing the estimate of four statistical moments.

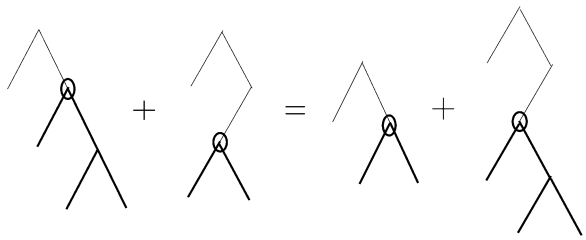

Fig. 6. Crossover operation.

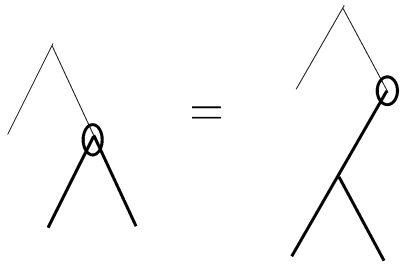

Fig. 7. Mutation operation.

TABLE I
OPERATOR SETS FOR THE GP

Symbol	No. of Inputs	Description
+, -	2	Addition, Subtraction
*, /	2	Multiplication, Division
square, sqrt	1	Square, Square Root
sin, cos	1	Trigonometric functions
asin, acos	1	Trigonometric functions
tan, tanh	1	Trigonometric functions
reciprocal	1	Reciprocal
log	1	Natural Logarithm
abs, negator	1	Absolute, Change Sign

(see Section III-B). The terminator receives the raw vibration data as the input and returns a single scalar value.

E. Primitive Operators

One of the main building blocks of the GP is the operator pool. The functions stored in the pool are mathematical, logical, or probabilistic operators that perform an operation on one or more inputs to give an output result. Table I lists the mathematical functions used as operators in this paper.

Note that any invalid input to an operator will result in a false flag being assigned to the fitness value in order to filter out individuals who cannot successfully complete the mathematical transforms. This will effectively exclude them from further consideration during the experiment.

F. Representation of Each Individual

Since expressions can be represented as trees ordered by operator precedence, GP systems in this paper, evolve programs using tree representation. Each member can be written as a polynomial expression consisting of several nonlinear functions up to a maximum specified depth. Using this function, each individual in the population is a mathematical formula that transforms the time series signals into a feature data. Formula $T\,\mathrm{Root} = \tanh(\mathrm{kurtosis}) + \mathrm{shewness}$ can be represented by the Fig. 8.

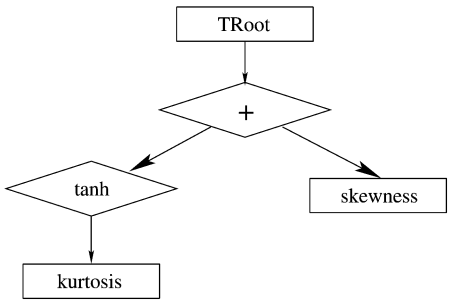

Fig. 8. Tree representation.

V. Classifiers

There are three possible ways of using classifiers.

- In the first case, raw vibration data can be used directly as the inputs of the classifier. We leave it to the classifiers for selection and extraction of information from the raw data by enhancing discriminating features and diminishing interfering data during the learning process. This is carried out by adjusting weights and the amount of data processed by the classifier will be enormous. Therefore, the useful information is difficult to maximize, while interference cannot be eliminated entirely.
- In the second case, a much more popular and effective realization, some useful features are prepared from the raw data for the inputs to the classifier. Based on the physical and mathematical analysis of the mechanism of rotating machinery in faulty conditions, features are computed to represent the fault information concealed within the raw vibration data. Therefore, classifier has much less difficulty than in the first realization.
- In the third case, rather than manual development, features are computed through an evolutionary process in order to avoid human negative influence and bias, and to conduct the feature extraction from a much larger space.

Obviously, the later two realizations are more effective and are examined and compared through a series of experiments in this paper. To demonstrate the robustness of extracted features, two classification algorithms are proposed, including artificial neural networks (ANNs) and support vector machines (SVMs).

A. ANN

ANNs are probably one of the most common classifiers in use today. This is mainly due to their ability to learn and identify patterns in the source data. For machine condition monitoring, where the training dataset is often sparse, and the classifier has to generalize to a certain extent, ANN is an ideal solution because of its nonlinearity, and many applications can be found regarding bearing fault detection [27]–[29]. The multilayer peceptron (MLP) is chosen here as the structure of the network for its overall performance over other configurations. The MLP used consists of one hidden layer and one output layer, with the hidden layer having a logistic activation function and the output layer using a linear activation function. For training procedure, the back propagation algorithm with adaptive learning and momentum is used. The learning algorithm is stopped when the classification performance of the validation set starts to diverge from that of the training set.

B. SVM

SVM, a new generation of learning algorithm based on advances in statistical learning theory, has gained considerable popularity recently in the field of machine learning. It can be characterized as a supervised learning algorithm capable of solving linear and nonlinear classification problems [30]–[32]. In this paper, a nonlinear SVM classifier with polynomial kernel is employed for the classification task.

VI. Results and Comparison

A. Feature Extraction Results

Two examples of extracted features described in this section are generated by GP with different stopping criteria.

1) Feature 1: Feature 1 was extracted by GP after evolving 1000 generations using the raw vibration data as the input. The maximum tree depth was chosen as five. There is always a possibility that two GP-generated features in one population are identical and this probability generally increases with the size of the population. Consequently, the population size does not need to be large for only four terminators and in this experiment is chosen as 10 to avoid unnecessary computations. The total time for computation is about five minutes for running 1000 generations. The extracted feature is given by

$$f1 = \log\left[\left(m_x^{(2)}\right)^4\right] - \left[\log\left(m_x^{(4)}\right) - m_x^{(1)} + \frac{m_x^{(1)}}{m_x^{(4)} - m_x^{(3)}}\right]. \quad (9)$$

Fig. 9 shows the feature-processed data for six different conditions. There are altogether 960 examples from the six conditions, with 160 examples for each class. Evidently, class IR, OR, and RE are well separated from each other, and from classes NO, NW, and CA as well, meaning that three faulty conditions—inner race fault, outer race fault and rolling element fault—are easily distinguishable with this feature.

On the other hand, conditions NO and CA overlap, while both of them are almost separated from NW. This implies that these three conditions are not easy to separate and these may be confused with each other to a large extent in the one-dimensional feature space. During the status change from normal condition to slightly worn condition in machine life, there is no significant defect occurring in the components. The physical nature of the bearing varies in a manner unnoticeable on visual inspection.

Also, it can be seen that the cage fault is confused with normal conditions. This may be improved by incorporating more features to solve the problem in multidimensional space and/or new terminators, or operators, which have discriminating ability especially for the cage fault.

2) Feature 2: Feature 2 was generated by GP after 10 000 generations with the population size of 14 and the maximum

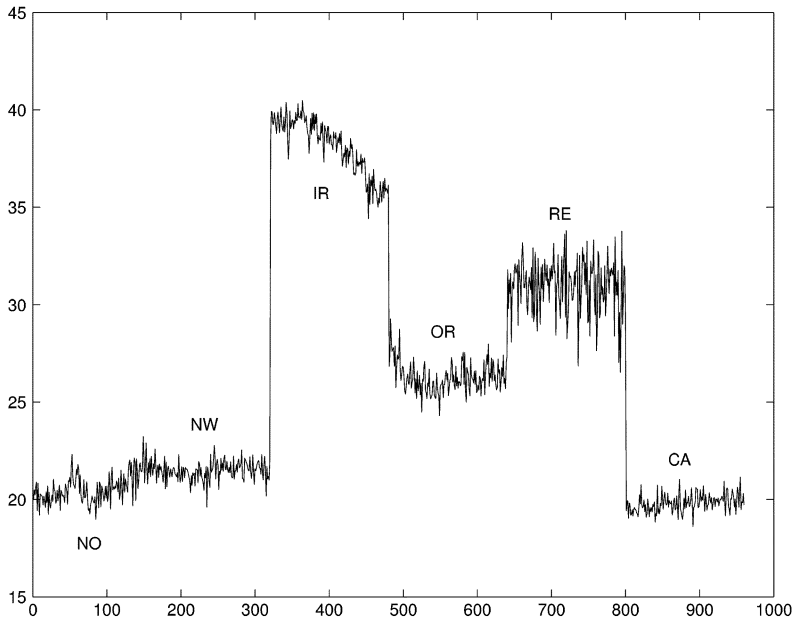

Fig. 9. Evolved feature 1.

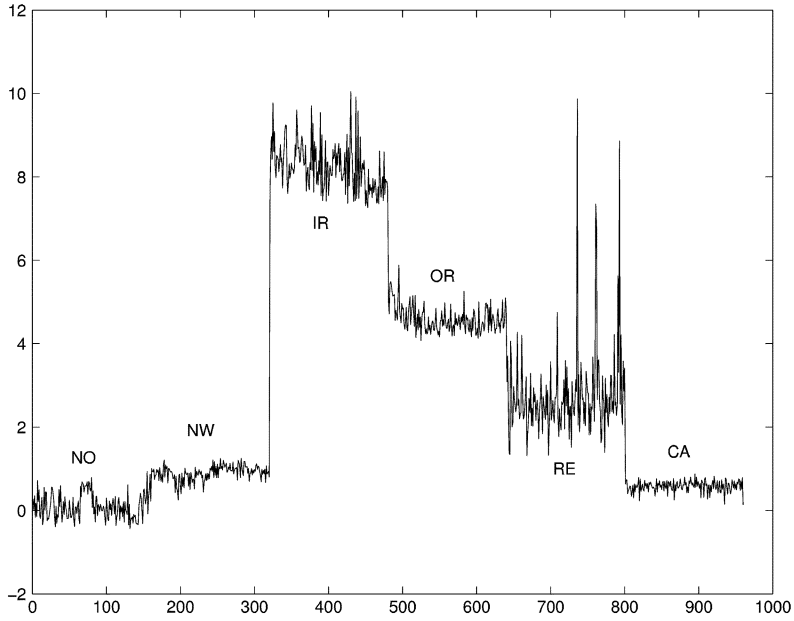

Fig. 10. Evolved feature 2.

depth of 10. It took about one hour for computation. The formula of the feature is given by

$$f2 = \cos\left(\log\left(\frac{m_x^{(2)}}{m_x^{(4)}}\right) - \tan\left[\mathrm{abs}\left(m_x^{(3)}\right)\right]\right) - \tan h\left(m_x^{(3)}\right) + \cos\left(m_x^{(1)}\right) + \frac{\log\left(m_x^{(2)}\right)}{m_x^{(2)} \Big/ m_x^{(4)}} + \mathrm{abs}\left(\frac{m_x^{(1)}}{m_x^{(4)} - m_x^{(1)}}\right) \quad (10)$$

Fig. 10 demonstrates that conditions IR, OR, and RE are well separated, although condition RE does not have as good a separation as with feature 1, mainly due to the fitness algorithm, which uses the smallest Fisher criterion value as the fitness, which gives the feature with better discriminating ability for closest classes, specifically condition NO, NW, and CA, more chance to survive. This can also be seen as a compensation among all classes in order to give the best overall performance. Clearly, classes NO (example 1–160) and NW (example 161–320) are better separated than with feature 1.

Overall, the GP-based feature extractor performed very well by separating different classes without any explicit knowledge of the statistical distribution of the data.

B. Classification Results

A number of experiments were carried out to evaluate the discriminating ability of features generated by GP and other classical feature extraction methods in term of classification performance, using ANN and SVM classifiers, respectively. The first set of results (Table II) were obtained from the combination of two GP-generated features and an MLP with one hidden layer. The second set of results (Table III) are achieved from combinations of each of 2, 3, 4, and 5 GP-generated features and an MLP with one hidden layer of each of 3–14 neurons. Table V presents the comparison results of classification success rate using GP generated features and original features. Finally, the classification performance using ANN and SVM with the non-normalized feature sets and normalized feature sets as the inputs are presented in Tables VI and VII.

1) Classification Result Using Two GP Extracted Features: Table II shows the confusion matrix of the classification performance for six conditions, using only two features extracted by GP. Each row of this Table shows the associated classifications results made by the MLP for a given condition. Each entry in the row shows what the perceived classification are, expressed as a percentage of the total number of cases for the condition. From Table II, it can be observed that condition IR, OR, and RE manage to achieve 100% accuracy, condition NW and CA achieving 95.6%. 10% of samples from normal condition are misclassified with the condition CA. It is also very clear from the Figs. 9 and 10, that most of the misclassification occur between classes NO and CA.

2) Classification Result Using Different Number of GP Extracted Features and Neurons: Table III shows the percentage of classification success for six bearing conditions with MLP, which consists of one hidden layer with neuron number from 3–14, using GP extracted features from 2 to 5. It is clear that the GP/ANN classification success rate is always more than 95%, with the lowest being 95% (for eight neurons with two features or eight neurons with three features) and the highest being 96.7% (for 14 neurons and five features). It can be seen that using the GP generated features the classification results are robust with respect to the choice of the number of neurons. The classification performance improvement by increasing the number of neurons is fairly small, ranging from 0.9% to 1.3% for each feature set.

The correlation coefficients of five GP extracted features are given in Table IV. It can be seen that all absolute values of coefficients involving feature 1 are 0.4 or less. Features 2–5 are fairly correlated, sometimes positively and sometimes negatively. Yet, the classification results obtained using two to five GP extracted features essentially do not vary with the number of neurons in the MLP (see Table III).

3) Classification Result Using ANN With the GP Generated Features and Four Plain Statistical Features: The percentages of correct classification are listed in Table V for the comparison of performance between four normalized statistical features and GP-extracted features with the variation of number of neurons used in the MLP hidden layer. The classification success rate (%) of GP/ANN is obtained by averaging over experiments using 2, 3, 4, and 5 GP-generated features. The classification success rate using GP-extracted features is higher than those using four plain statistical features, with improvements ranging from 1.9% in the case using 13 neurons to 17.4% in the case using three neurons. It is interesting to see that the increase in the number of neurons does not seem to help the classification much in cases using GP-extracted features while original features require more neurons to achieve better performance, with only a 78.4% success for one neuron and a maximum 94.3% success when using 13 neurons. However, only 79.4% classification success is obtained when 14 neurons are used. It should be noted that the data of four plain statistical features were normalized (see Table V) to help improve the classification accuracy; however, the GP results in this paper are based on unnormalized data and without the benefit of normalization enjoyed by the classifiers.

4) Comparison of Features Generated by GP and Classical Methods: The classification performance results displayed in Table VI were obtained using features generated by GP and classical methods. Both ANN and SVM classifiers are utilized in order to see the capability of different feature sets over different classification algorithms. In this experiment, features are directly used as the inputs to classifiers without normalization. As listed in Table VI, among classical methods, conventional features achieve the best classification performance for both ANN classifier with success rate at 91.5% and SVM classifier with success rate at 92%. The four low-pass filter features perform

TABLE II
CLASSIFICATION PERFORMANCE (%) FOR GP/ANN, USING TWO FEATURES EXTRACTED BY GP AND 12 NEURONS IN ANN

	NO	NW	IR	OR	RE	CA
NO (160 in total)	**85**	5	0	0	0	10
NW (160 in total)	0.6	**95.6**	0	0	0	3.8
IR (160 in total)	0	0	**100**	0	0	0
OR (160 in total)	0	0	0	**100**	0	0
RE (160 in total)	0	0	0	0	**100**	0
CA (160 in total)	3.1	1.3	0	0	0	**95.6**

TABLE III
CLASSIFICATION SUCCESS (%) WITH THREE TO 14 NEURONS IN ONE HIDDEN LAYER OF ANN, USING TWO TO FIVE DIFFERENT FEATURES EXTRACTED BY GP

	2 features test perf.(%)	3 features test perf.(%)	4 features test perf.(%)	5 features test perf.(%)
3 neurons	95.3	95.9	96.0	95.9
4 neurons	95.2	95.7	95.8	96.1
5 neurons	95.5	95.6	95.9	96.3
6 neurons	95.3	95.9	96.1	96.3
7 neurons	95.1	95.9	96.1	96.1
8 neurons	95.0	95.0	95.2	95.7
9 neurons	95.5	96.1	96.3	96.4
10 neurons	95.8	95.9	96.3	96.4
11 neurons	95.4	96.0	96.3	96.3
12 neurons	96.0	96.0	96.3	96.1
13 neurons	95.9	96.1	96.5	96.3
14 neurons	95.8	96.1	96.4	**96.7**

TABLE IV
CORRELATION COEFFICIENTS OF FIVE FEATURES EXTRACTED BY GP

	feature1	feature2	feature3	feature4	feature5
feature1	1.0000	-0.4084	0.4018	0.3576	-0.2709
feature2	-0.4084	1.0000	-0.8754	-0.9388	0.8182
feature3	0.4018	-0.8754	1.0000	0.8099	-0.9206
feature4	0.3576	-0.9388	0.8099	1.0000	-0.7729
feature5	-0.2709	0.8182	-0.9206	-0.7729	1.0000

TABLE V
CLASSIFICATION SUCCESS (%) WITH THREE TO 14 NEURONS IN ONE HIDDEN LAYER OF ANN, USING TWO TO FIVE DIFFERENT FEATURES EXTRACTED BY GP

	GP-Generated Features/ANN		Four Stat. Features/ANN
No. of Neurons	Test success(%) average	stdev	Test success(%) best
3 neurons	95.8	0.3	78.4
4 neurons	95.7	0.4	81.1
5 neurons	95.8	0.4	81.9
6 neurons	95.9	0.4	89.5
7 neurons	95.8	0.5	92.3
8 neurons	95.2	0.3	94.2
9 neurons	96.1	0.4	94.1
10 neurons	96.1	0.3	88.5
11 neurons	96.0	0.4	93.4
12 neurons	96.1	0.1	92.2
13 neurons	96.2	0.3	94.3
14 neurons	96.3	0.4	79.4

TABLE VI
CLASSIFICATION SUCCESS USING THE ANN AND SVM CLASSIFIERS WITH THE UN-NORMALISED DIFFERENT TYPE FEATURES

Features Type	ANN Best Perf.(%)	SVM Best Perf.(%)
4 Plain stat.	90.8	91.2
4 High pass Filter	85.4	85.0
4 Low pass Filter	87.1	16.7
4 Difference	45.4	35.6
4 Sum	65.1	90.5
4 Conventional	91.5	92
4 GP features	96.5	97.1

the worst with around 16.7% success in SVM classifier, which is mainly due to large variation in values in the unnormalized data. This may be addressed by using the ideas in [32]. When GP extracted features are used, the improvement is overwhelming in either ANN or SVM classifier. Overall, GP produces much more robust features for classifiers and has the ability to perform well without normalizing the data.

With normalized data, this experiment examines the classification performance in each scenario by the same feature extraction methods used in the proceeding experiment. As shown in Table VII, each scenario sees an improvement in classification performance, ranging from 0.6% to 68.9%, compared with those using un-normalized data. Evidently, the most significant enhancement occurs in the case of low pass filter features with SVM as the classifier, though this is mainly due to the normalization. For the same reason, two classifiers have much less difference in classification performance compared with those in the last experiment.

Summarizing the results from these two experiments, it can be said that when classical feature extraction methods are employed, the classification performance changes drastically with the variation of classification method and data, while GP-extracted features maintain a constantly high level of performance. The superior performance of GP against other methods is clearly demonstrated through the overall classification success.

VII. DISCUSSION

Based upon the experimental results, it can be said that using features generated by GP, both the ANN and SVM classifiers see a significant improvement in classification accuracy and robustness, compared with those using classically developed features. GP derives feature selection from GA, but the available feature space in GP is much larger than that in GA. Rather than the pure selection in GA, GP has been developed here to produce new features by choosing terminators and operators.

TABLE VII
CLASSIFICATION SUCCESS (%) USING THE ANN AND SVM CLASSIFIERS WITH THE NORMALIZED DIFFERENT TYPE FEATURES

Features Type	ANN Best Perf.(%)	SVM Best Perf.(%)
4 Plain stat.	94.3	94.6
4 High Pass Filter	91.4	94.5
4 Low Pass Filter	87.7	85.6
4 Difference	46.9	40.3
4 Sum	89.7	94.1
4 Conv. features	89.4	94.1

The computation cost of GP in the feature extraction process is slightly larger than that using manual approaches. However, GP requires less computation compared with GA for feature selection and generation. GA/ANN [14] takes a couple of days to work out a solution achieved by GP in only a few hours. The proposed method requires comparatively less computation since it does not involve wrapper type feature selection/extraction, instead it is based on a Fisher criterion. GP also has the ability to process the un-normalized datasets as the input, remove the systematic variation and bring the raw data onto the same ground for a fair comparison.

VIII. CONCLUSION AND FURTHER WORK

In this paper, a GP-based feature extractor is proposed for the generation/extraction of features from the raw vibration data for classification applied to the problem of bearing fault classification. GP is a powerful and efficient tool for the automatic feature generation directly from the raw data. Using features extracted by GP, the ANN and SVM classifier sees a significant improvement in classification results, compared with those using extracted features by classical methods.

It is also shown from the extraction results that GP is not only capable of enhancing the classification performance, but also reducing the dimensionality to describe the problem. Furthermore classification performance obtained from GP extracted features are very robust. Also, GP produces results in a tree representation, which allows an understanding of how it works.

So far only four statistical terminators and a few operators have been used in this paper in an attempt to examine the feasibility of the current scheme. In the next stage, it is necessary to implement more terminators and operators in order to increase the feature searching space, so as to give better discriminating performance and a more useful realization.

REFERENCES

[1] K. Ragulskis and A. Uurkauskas, *Vibration of Bearings*. Bristol, PA: Hemisphere, 1989.
[2] G. Lipovszky, K. Solyomvari, and G. Varga, *Vibration Testing of Machines and Their Maintenance*. Amsterdam, The Netherlands: Elsevier, 1990.
[3] T. A. Harris, *Rotting Bearing Analysis*. New York, NY: Wiley, 1991.
[4] S. Korablev, V. Shapin, and Y. Filatov, *Vibration Diagnostics in Precision Instruments*. Bristol, PA: Hemisphere, 1989.

[5] P. D. Heerman and N. Khazenie, "Classification of multispectral remote sensing data using a back propogation neural network," *IEEE Trans. Geosci. Remote Sens. E*, vol. 30, no. 1, pp. 81–88, Jan. 1992.
[6] M. J. Chang and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Trans, Neural Netw.*, vol. 6, no. 2, pp. 296–317, Mar. 1995.
[7] G. G. Yen and P. Meesad, "An effective neuro-fuzzy paradigm for machinery condition health monitoring," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 31, no. 4, pp. 523–536, Aug. 2001.
[8] K. P. Philip, E. L. Dove, D. D. McPherson, N. L. Gotteiner, W. Stanford, and K. B. Chandran, "The fuzzy hough transform-feature extraction in medical images," *IEEE Trans. Med. Imag.*, vol. 13, no. 2, pp. 235–240, Jun. 1994.
[9] M. M. Rizki, M. A. Zmuda, and L. A. Tamburino, "Evolving pattern recognition systems," *IEEE Trans. Evol. Comput.*, vol. 6, no. 6, pp. 594–609, Dec. 2002.
[10] M. L. Raymer, T. E. Doom, L. A. Kuhn, and W. F. Punch, "Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm raymer," *IEEE Trans Syst., Man, Cybern. B, Cybern.*, vol. 33, no. 5, pp. 802–813, Oct. 2003.
[11] T. Bach, *Evolutionary Algorithms in Theory and Practice*. London, U.K.: Oxford Univ. Press, 1996.
[12] D. B. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. Piscataway, NJ: IEEE Press, 1995.
[13] J. Yang and V. Honvar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, pp. 44–49, Mar./Apr. 1998.
[14] L. B. Jack and A. K. Nandi, "Genetic algorithms for feature selection in machine condition monitoring with vibration signals," *Proc. Instit. Elec. Eng. Vision, Image Signal Processing*, vol. 147, no. 3, pp. 205–212, Jun. 2000.
[15] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Trans. Evol. Comput.*, vol. 4, pp. 164–171, Apr. 2000.
[16] G. A. Rovithakis, M. Maniadakis, and M. Zervakis, "A hybrid neural network/genetic algorithm approach to optimizing feature extraction for signal classification," *IEEE Trans. Syst., Man, Cybern. Part.B*, vol. 34, no. 1, pp. 695–703, Feb. 2004.
[17] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.
[18] I. Benyahia and J. Potvin, "Decision support for vehicle dispatching using genetic programming," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 28, no. 3, pp. 306–314, May 1998.
[19] M. Brameier and W. Banzhaf, "A comparison of linear genetic programming and neural networks in medical data mining," *IEEE Trans. Evol. Comput.*, vol. 5, no. 1, pp. 17–26, Feb. 2001.
[20] J. K. Kishore, L. M. Patnaik, V. Mani, and V. K. Arawal, "Application of genetic programming for multicategory pattern classification," *IEEE Trans. Evol. Comput.*, vol. 4, no. 3, pp. 242–258, Sep. 2000.
[21] L. Zhang, L. B. Jack, and A. K. Nandi, "Fault detection using genetic progamming," *Mech. Syst. Sig. Process.*, 2003.
[22] J. R. Sherrah, R. E. Bogner, and A. Bouzerdoum, "The evolutionary pre-processor: Automatic feature extraction for supervised classification using genetic programming," in *Proc. 2nd Int. Conf. Genetic Programming*, 1997, pp. 304–312.
[23] M. Kotani, S. Ozawa, M. Nasak, and K. Akazawa, "Emergence of feature extraction function using genetic programming," in *Proc. 3rd Int. Conf. Knowledge-Based Intelligent Information Engineering Systems*, 1997, pp. 149–152.
[24] P. Chen, T. Toyota, and Z. He, "Automated function generation of symptom parameters and application to fault diagnosis of machinery under variable operating conditions," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 6, pp. 775–781, Nov. 2001.
[25] C. Wang and R. X. Gao, "Wavelet transform with spectral post processing for enhanced feature extraction," *IEEE Trans, Instrum. Meas.*, vol. 52, no. 4, pp. 1295–1301, Aug. 2003.
[26] C. M. Bishop, *Neural Networks for Pattern Recogniation*, Oxford, UK: Clarendon Press, 1995.
[27] L. B. Jack and A. K. Nandi, "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms," *Mechanical Systems and Signal Processing*, vol. 16, pp. 373–390, 2002.
[28] A. C. McCormick and A. K. Nandi, "Real time classification of rotating shaft loading conditions using artificial neural networks," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 748–757, May 1997.
[29] C. Kirkham, A. Long, O. Taylor, and C. Isbell, "Adaptive online systems for condition monitoring: The neural-maine project," in *Proc. COMADEM*. Kingham, U.K., 1999, pp. 317–325.
[30] S. T. Li, J. T. Kwok, H. L. Zhu, and Y. N. Wang, "Texture classification using the support vector machines," *Pattern Recognit.*, vol. 36, pp. 2883–2893, Dec. 2003.
[31] B. Heisele, T. Serre, S. Prentice, and T. Poggio, "Hierarchical classification and feature reduction for fast face detection with support vector machines," *Pattern Recognit.*, vol. 36, pp. 2007–2017, 2003.
[32] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

Hong Guo received the M.Sc. degree in microelectronic systems and telecommunications from the University of Liverpool, Liverpool, U.K., in 2003 and is currently pursuing the Ph.D. degree in signal processing and communication group from the same university.

Her research interests are in fields of feature extraction and pattern recognition using genetic programming and evolutionary algorithms.

Lindsay B. Jack received the B.Eng. degree in electrical and mechanical engineering from Strathclyde University, Strathclyde, U.K., in 1997 and the Ph.D. degree in electrical engineering from the University of Liverpool, Liverpool, U.K., in 2000.

His research interests include machine condition monitoring, biomedical signal processing, microarray analysis, neural networks, evolutionary programming, feature selection and feature extraction, and clustering algorithms.

Asoke K. Nandi (M'88–SM'96) received the Ph.D. degree from the University of Cambridge (Trinity College), Cambridge, U.K., in 1979.

He held several research positions in Rutherford Appleton Laboratory, U.K., European Organization for Nuclear Research, Switzerland, the Department of Physics, Queen Mary College, London, U.K., and the Department of Nuclear Physics, Oxford, U.K. In 1987, he joined the Imperial College, London, U.K., as the Solartron Lecturer in the Signal Processing Section of the Electrical Engineering Department. In 1991, he jointed the Signal Processing Division of the Electronic and Electrical Engineering Department in the University of Strathclyde, Glasgow, U.K., as a Senior Lecturer; subsequently, he was appointed a Reader in 1995, and a Professor in 1998. In March 1999 he moved to the University of Liverpool, Liverpool, U.K. to take up his appointment to the David Jardine Chair of Signal Processing in the Department of Electrical Engineering and Electronics. In 1983, he was a member of the UA1 team at CERN that discovered the three fundamental particles known as W^+, W^-, and Z° providing the evidence for the unification of the electromagnetic and weak forces, which was recognized by the Nobel Committee for Physics in 1984. Currently, he is the Head of the Signal Processing and Communications Research Group with interests in the areas of nonlinear systems, non-Gaussian signal processing, and communications research. With his group he has been carrying out research in machine condition monitoring, signal modeling, system identification, communication signal processing, biomedical signals, applications of artificial neural networks & support vector machines, ultrasonics, blind source separation, and blind deconvolution. He has authored or coauthored over 200 technical publications including two books "*Automatic Modulation Recognition of Communications Signals*" (Boston, MA: Kluwer Academic, 1996) and "*Blind Estimation Using Higher-Order Statistics*" (Boston, MA: Kluwer Academic, 1999).

Dr. Nandi was awarded the Mounbatten Premium Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers of the U.K. in 1998 and the Water Arbitration Prize of the Institution of Mechanical Engineers of the U.K. in 1999. He is a Fellow of the Cambridge Philosophical Society, the Institution of Electrical Engineers, the Institute of Mathematics and its applications, the Institute of Physics, and the Royal Society for Arts.