Med Bio Eng Comput (2006) 44:683–694
DOI 10.1007/s11517-006-0077-6

ORIGINAL ARTICLE

Classification of breast masses in mammograms using genetic programming and feature selection

R. J. Nandi · A. K. Nandi ·
R. M. Rangayyan · D. Scutt

Received: 16 December 2005 / Accepted: 12 May 2006 / Published online: 21 July 2006

Abstract Mammography is a widely used screening tool and is the gold standard for the early detection of breast cancer. The classification of breast masses into the benign and malignant categories is an important problem in the area of computer-aided diagnosis of breast cancer. A small dataset of 57 breast mass images, each with 22 features computed, was used in this investigation; the same dataset has been previously used in other studies. The extracted features relate to edge-sharpness, shape, and texture. The novelty of this paper is the adaptation and application of the classification technique called genetic programming (GP), which possesses feature selection implicitly. To refine the pool of features available to the GP classifier, we used feature-selection methods, including the introduction of three statistical measures—Student's *t* test, Kolmogorov–Smirnov test, and Kullback–Leibler divergence. Both the training and test accuracies obtained were high: above 99.5% for training and typically above 98% for test experiments. A leave-one-out experiment showed 97.3% success in the classification of benign masses and 95.0% success in the classification of malignant tumors. A shape feature known as fractional concavity was found to be the most important among those tested, since it was automatically selected by the GP classifier in almost every experiment.

Keywords Pattern classification · Breast masses · Breast tumors · Breast cancer · Computer-aided diagnosis · Genetic programming · Feature selection

1 Introduction

Breast cancer is the most common cancer in England and Wales and the most common cause of cancer death in women [1]. Early detection of breast cancer is a key factor in prognosis, and consequently plays a major role in reducing mortality. The detection of breast cancer is currently performed by radiologists using mammography, with a significant human element to the diagnosis. Techniques are being developed to introduce computer-aided diagnosis (CAD) procedures for efficient screening and detection of breast cancer [2, 3, 4]. Several attempts have been made to detect and classify breast masses in mammograms. Brzakovic et al. [5] employed a fuzzy pyramid linking technique for mass localization and shape analysis for false-positive elimination. Their method was applied on a small dataset of 25 mammograms and produced a classification accuracy of 85%. Kegelmeyer et al. [6] employed four of Laws' texture measures [7] and a new feature sensitive to stellate patterns for the detection of spiculated lesions. They applied their algorithm to a database of 85 cases, consisting of 49 normal cases and 36 positive cases. They achieved a sensitivity of 97% at 0.28 false positives per image. Rangayyan et al. [8]

R. J. Nandi · A. K. Nandi (✉)
Department of Electrical Engineering and Electronics,
The University of Liverpool,
Brownlow Hill, Liverpool, L69 3GJ, UK
e-mail: a.nandi@liverpool.ac.uk

R. M. Rangayyan
Department of Electrical and Computer Engineering,
Schulich School of Engineering, University of Calgary,
Calgary, AB, Canada T2N 1N4

D. Scutt
School of Health Sciences, The University of Liverpool,
Thompson Yates Building, Liverpool, L69 3GB, UK

introduced two new shape factors, spiculation index and fractional concavity. Their combined use of these two factors and the measure of compactness yielded a benign-versus-malignant classification accuracy of 81.5%. Sahiner et al. [9, 10] defined a "rubber-band straightening transform" (RBST) to map ribbons around breast masses in mammograms into rectangular arrays, and then computed Haralick's texture measures [11, 12]. The texture measures individually provided classification accuracies of up to only 66%, whereas the Fourier-descriptor-based shape factor defined by Shen et al. [13] gave an accuracy of 82%. Rangayyan et al. [14] proposed the use of shape factors and a measure of edge-sharpness known as acutance for the classification of manually segmented masses as benign or malignant, and as spiculated or circumscribed. They obtained an overall accuracy of 95% with a database of 54 mammographic images.

A problem encountered in studies such as those described above lies in the selection of the best subset of features from those available so as to facilitate efficient pattern classification. Sahiner et al. [15] studied this problem, including the practical situation related to the availability of a small dataset, in the context of CAD of breast cancer. Alto et al. [16] used stepwise logistic regression to select subsets of features from a total set of 22 features of breast masses, including measures of texture, shape, and edge-sharpness. The selected features were used for content-based retrieval of breast masses, as well as for pattern classification. The separation of feature selection from the pattern classification step raises questions regarding the potential optimality of the feature subset selected for some classifiers and not others. The exhaustive testing of all possible combinations of features is impractical when the number of features available is large. Well-known techniques for feature selection such as sequential forward or backward selection [17] lead to suboptimal results, depending upon the sequence of selection of the features. The popular technique of principal component analysis (PCA) [18] does not identify the result of feature selection explicitly, resulting in a void in the practical interpretation of the results.

In this paper, a novel technique called genetic programming (GP) is introduced and adapted for classification of breast masses into the benign and malignant categories. In Sect. 2 are presented details regarding the data and the features used in this paper. The technique of GP is introduced in Sect. 3. In Sect. 4 is discussed feature selection, and the background to testing is presented in Sect. 5. Experiments and results are detailed in Sect. 6, and the paper is concluded in Sect. 7.

2 The data and the features

The data were extracted from mammograms obtained from Screen test: Alberta Program for the Early Detection of Breast Cancer [19], with 37 regions of interest (ROIs) related to benign masses and 20 ROIs related to malignant tumors [16]. The images were digitized to a resolution of 50 μm with 12 bits per pixel; however, texture features were extracted after resampling to 200 μm and requantization to 8 bits per pixel. The diagnosis of each case was proven by biopsy. Benign and malignant ROIs were manually identified, and contours were drawn by a radiologist experienced in screening mammography.

Benign masses generally possess smooth and round contours, whereas malignant tumors typically exhibit rough contours with spiculations and concavities. Also, benign masses generally have homogeneous internal texture and sharp or well-circumscribed margins, whereas malignant tumors typically exhibit heterogeneous texture and ill-defined or blurred margins. Shown in Fig. 1 are examples of four breast masses and tumors, selected from the data used in the present work, illustrating their typical characteristics as described above.

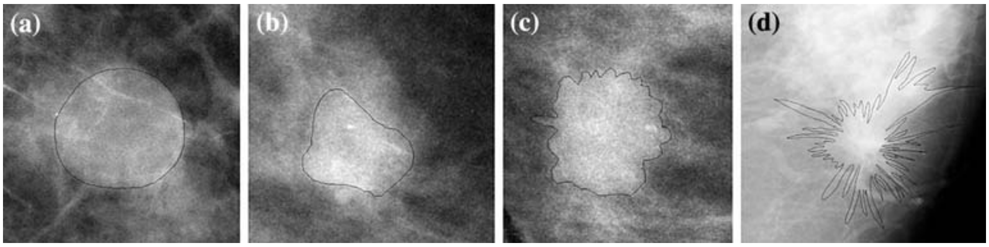

Fig. 1 Examples of benign breast masses and malignant tumors: **a** a well-circumscribed benign mass with a round contour; **b** a macrolobulated benign mass; **c** a microlobulated malignant tumor; and **d** a spiculated malignant tumor. The contour of the mass or tumor, as drawn by a radiologist experienced in mammography, is overlaid on each image

A set of 22 features was extracted for each ROI [16]. The features include four edge-sharpness measures [14, 20], four shape factors [8, 14], and 14 statistical texture features [11, 12, 20, 21].

2.1 Edge-sharpness features

Four edge-sharpness features have been considered in the following investigations. Acutance (A) is a measure of the sharpness or change in density across a mass margin [14, 20]. In order to obtain the directional derivatives required for the computation of acutance, a line of pixels is defined in the direction normal to the contour at each boundary pixel. A distance of 80 pixels (4 mm) inside and outside the mass contour is used where possible; in the case of narrow spicules or concavities, as many pixels as available are used such that the line of pixels along the normal to the contour does not cross the contour more than once. The average of the differences between equidistant pixels on the inside and outside of the mass contour along the normal to each boundary pixel is computed, and further averaged over the entire contour to obtain a root-mean-squared (rms) measure of edge-sharpness, labeled as Acutance 1 or A in Table 1. A second version of acutance, labeled as Acutance 2 in Table 1, was computed in a similar manner, but by taking the differences between adjacent pixels along the normals. In addition to the above, a measure of contrast (Co) and the coefficient of variation (CV) of the average gradient along the normals to the contour were computed [20].

Table 1 List of features used to represent breast masses in mammograms

Feature type	Feature number	Feature name
Edge-sharpness	1	Co: Contrast
	2	A: Acutance 1
	3	Acutance 2
	4	CV: Coefficient of variation
Shape	5	C: Normalized compactness
	6	FF: Fourier-descriptor-based factor
	7	SI: Spiculation index
	8	F_{cc}: Fractional concavity
Texture	9	F_1: Energy
	10	F_2: Contrast
	11	F_3: Correlation
	12	F_4: Sum of squares
	13	F_5: Inverse difference moment
	14	F_6: Sum_average
	15	F_7: Sum_variance
	16	F_8: Sum_entropy
	17	F_9: Entropy
	18	F_{10}: Difference variance
	19	F_{11}: Difference entropy
	20	F_{12}: Information measure of correlation
	21	F_{13}: Information measure of correlation
	22	F_{14}: Maximal correlation coefficient

2.2 Shape features

The four shape features extracted were normalized compactness (C), a normalized feature based upon Fourier descriptors (FF), spiculation index (SI), and fractional concavity (F_{cc}).

Normalized compactness is a simple measure of shape complexity computed as $C = 1 - \frac{4\pi a}{p^2}$, where a is the area and p is the perimeter of the contour [13]. With this expression, C is equal to 0.0 for a circle, and the parameter increases with the complexity of the contour to a maximum value of 1.0.

Shen et al. [13] proposed a normalized shape factor (FF) based upon the Fourier descriptors of the given contour. The Fourier descriptors were normalized and divided by the corresponding harmonic index. The sum of the scaled coefficients was divided by the sum of the unscaled coefficients, and the result was subtracted from 1 to obtain FF. The shape factor FF has been shown to be useful in classifying calcifications [13] and masses [9, 14].

Spiculation index (SI) represents the degree of spicularity of a mass contour. Rangayyan et al. [8] proposed an algorithm to compute SI based upon a polygonal model of the given contour, and a combination of the segment lengths, base widths, and angles of possible spicules. Invasive carcinomas, due to their nature of infiltration into surrounding tissues, often form narrow, stellate distortions at their contours, and hence yield higher values of SI than benign masses.

Fractional concavity (F_{cc}) is the ratio of the cumulative length of the concave portions of the contour to the total length of the contour. Benign masses, due to their round or oval contours, result in low values of F_{cc}. On the other hand, contours of microlobulated or spiculated malignant tumors may be expected to have several significant concave portions. The estimation of F_{cc} may be achieved by segmenting the contour into concave and convex portions separated by the points of inflexion on the contour [8].

The features described above have been found to be effective in the discrimination of masses, as reported previously by Alto et al. [16] and Rangayyan et al. [8].

2.3 Texture features

Mudigonda et al. [20] showed that texture features computed using the mass margin (a ribbon surrounding the mass) could lead to improved discrimination between benign and malignant masses, as compared to texture measures computed using the entire mass region. Fourteen texture features were computed according to Haralick's definitions [11, 12] for the mass ribbons. For convenience of reference all of the features are numbered and listed in Table 1. Alto et al. [16] gave a fuller explanation of the features and their use in the classification of masses; the same dataset is used in this paper.

3 Genetic programming

Genetic programming [22] is a type of evolutionary learning algorithm and is an extension of the genetic algorithm (GA). The main difference between GP and GA lies in the representation of the solution. GP evolves computer programs as the solution, whereas GA creates a string of numbers or parameters that influence the performance of a fixed solution. GP naturally allows nonlinear mapping and implicitly includes feature selection. GP is a relatively new classification technique, and has been successfully applied to a variety of classification problems [23–26]. GP embraces some concepts present in natural selection. An outline of how GP works is given below. At the beginning, GP has available a set of feature values, a set of functions, and a set of parameters. Initially, a set of individuals, called GP trees, (for an example, see Fig. 2) are created (randomly or otherwise).

The GP tree in Fig. 2 calculates feature 1 + tanh (feature 5). The idea is that one puts in the relevant feature values and gets out a number close to 0 or 1 depending on whether the condition is benign or malignant, respectively. The initial set of solutions (individuals) represents the first generation. The number of solutions in each generation is the population size. GP uses the following steps:

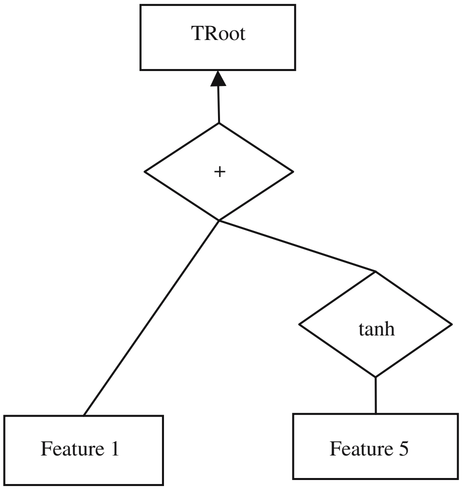

Fig. 2 An example of a GP tree

- Create initial population
- Loop
  - Fitness evaluation of each individual
  - Selection of individuals
  - Modification (by GP operators):
    - Crossover
    - Mutation
    - Replication
- Until some criterion is met.

Three processes occur to create the population for the next generation:

- Crossover—two new individuals are created by randomly selecting nodes from each parent and swapping them; see Fig. 3.
- Mutation—a node is randomly selected, the tree downstream from it is deleted, and a new sub-tree is generated from this node in exactly the same way as the initial population was grown; see Fig. 4.
- Replication—no change is made and an individual is simply copied.

The first generation experiences these processes, each of which happens with a certain probability. The next generation is chosen from the new individuals and from the previous generation according to a fitness function F, given as

$$F = \frac{1}{\sqrt{S+1}} - pN. \quad (1)$$

Here, S is the score of a GP tree, which is the number of incorrect classifications during training in a given generation; p is the node penalty, which is set at a fixed value (like 0.002); and N is the number of nodes in the tree. The purpose of the last term is to limit the size of the trees so that they do not grow out of proportions, in what is often referred to as code bloat, which makes the GP process take too long and the solution too complicated. The process continues until we have completed a certain number of generations. The final solution is the tree with the best score in the last generation obtained.

For the investigations in this paper, some preliminary experiments were carried out to try out a range of parameter values. With the experience gained from these explorations, the parameters were set as follows:

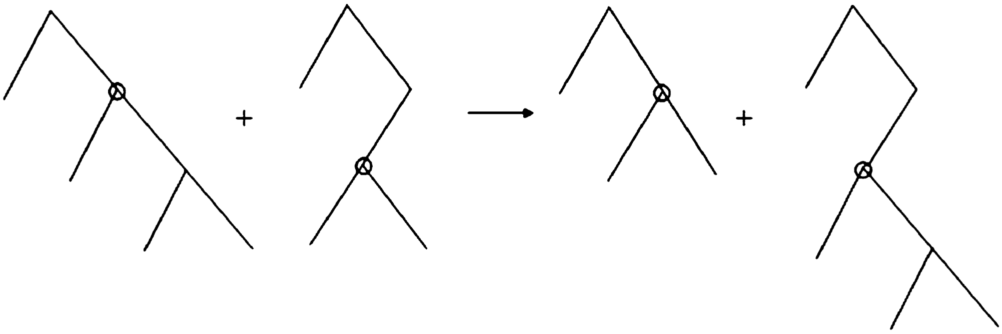

Fig. 3 Crossover operation

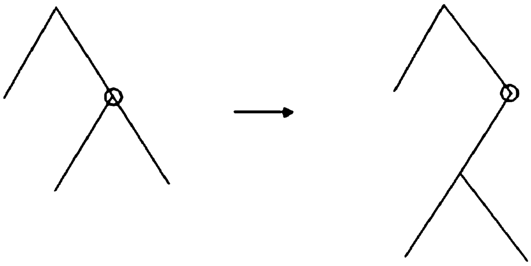

Fig. 4 Mutation operation

population size = 100; number of generations = 300 (for smaller feature set) or 500 (for larger feature set); probability of mutation = 0.4; probability of crossover = 0.4; probability of replication = 0.2; node penalty, p = 0.002.

4 Feature selection

Feature selection is often used in machine learning; it refers to the process where a subset of all the features extracted from the data is selected for use in a machine learning algorithm. Feature selection is necessary because either it is computationally too complex to use all the features available, or there is a problem with a rather limited amount of data and correspondingly large number of features. In the following investigations, we face the latter problem, often referred to as the curse of dimensionality [17, 27].

Although GP has feature selection built into it, it is still helpful to restrict the pool of features in order to make the task easier for the GP classifier. So as not to increase the complexity too much, we only used standalone (i.e., without reference to any classifier) feature-selection algorithms. In the following, we used three statistical tests and two other feature-selection methods in order to restrict the pool of features.

For each feature, there are two distributions: one of the data relating to benign masses and the other of the data relating to malignant tumors. The objective is to find the features where these distributions are separated the most, so one can distinguish between the malignant tumors and benign masses. To do this, one would ideally consider the multidimensional distributions in the feature space. This increases computational complexity and reduces robustness when there are not many feature vectors (data samples) available. To overcome both these shortcomings, we project the multidimensional distributions onto each single feature space, and apply each statistical test in each feature space. The following techniques are used in our attempt to achieve the above objective.

4.1 Sequential forward selection (SFS)

The SFS [17] is performed by initially finding the best single feature in terms of the separation of the two classes. Then, that feature is adopted permanently and put in combination with each of the other features in turn. Out of all of the pairs so created, the best set of two features is taken, again in terms of separation in the feature space. The process continues until the desired number of features is obtained.

For SFS to be optimal, the best set of $(n + 1)$ features must contain the best set of n features for all integers n up to the total number of features. This is not always the case, but even so, the SFS method is useful, because the best set of features is likely to contain the better features, and it is computationally more efficient than checking all possible combinations (an exhaustive search).

4.2 Sequential backward selection (SBS)

The SBS [17] starts with the complete set of features, and initially removes each feature in turn from the set. The feature which, when removed, gives the largest separation between the centroids of the two classes in the remaining feature space, is permanently removed. The process continues in this way, removing features one by one, until we have the desired number of features remaining.

4.3 Student's t test

The conventional statistic for measuring the significance of a difference of means is the Student's t: it is often used to test two sets of experimental results to see whether or not the means are statistically different. The relevant statistic for the unequal variance t test is [28]:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(Var(x)/N_1) + (Var(y)/N_2)}} \,. \tag{2}$$

Here, $Var(.)$ denotes the variance, while x and y are the two variables, which in our case are the values of the data for the benign and malignant categories with respect to one feature. N_1 and N_2 are the numbers of samples in the two categories, and the bar over the variable indicates its mean value. One needs to use the Student's t distribution, with the number of degrees of freedom N given by Eq. 3, to calculate the significance, i.e., the probability that the means differ significantly. In this paper, this probability is used as a measure of how distinguishable the distributions are:

$$N = \frac{(Var(x)/N_1 + Var(y)/N_2)^2}{(Var(x)/N_1)^2/(N_1 - 1) + (Var(y)/N_2)^2/(N_2 - 1)} \,. \tag{3}$$

4.4 Kolmogorov–Smirnov (K–S) test

The K–S test [28] can be used to determine whether two underlying probability distributions differ significantly. It is used frequently for hypothesis testing. The K–S test is performed using the cumulative probability distributions $P(x)$ and $Q(x)$. The K–S statistic D is the maximum difference between P and Q, given by

$$D = \max_{-\infty<x<\infty} |P(x) - Q(x)|. \tag{4}$$

The significance of any nonzero values of D can be calculated. The function f_{KS} used to calculate the significance, as

$$f_{\mathrm{KS}}(\lambda) = 2\sum_{j=1}^{\infty} (-1)^{j-1} \mathrm{e}^{-2j^2\lambda^2} \,. \tag{5}$$

In this paper, there are two distributions of different sizes. The benign distribution contains $N_1 = 37$ points and the malignant distribution contains $N_2 = 20$ points. In Eq. 5, $\lambda \approx \sqrt{\frac{N_1 N_2}{N_1+N_2} D}$.

4.5 Kullback–Leibler divergence (KLD)

The KLD [29] is a measure of the difference between two probability density functions. It is also called the relative entropy, and is important in information theory. It is worth noting that the term ''divergence'' is a misnomer, and also that KLD is not a distance. KLD is not symmetric, and does not satisfy the triangle inequality. The KLD between two probability density functions p and q of a discrete variable is

$$\mathrm{KLD}(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \,. \tag{6}$$

If we are dealing with continuous probability density functions, the sum becomes an integral. We converted the probability density functions into discrete functions by binning the data into categories. Note that KLD is nonnegative, and that it is zero if and only if $p = q$. Therefore, in the following investigations, the higher the KLD value, the more well-separated the probability density functions are, and so the easier it is to distinguish between the underlying variables; hence, the better the feature.

Other researchers have used the statistical tests described above for feature selection; for example, in the work of Nykter [30], the t test is used as a preliminary step to put the features in an order of importance; in the work of Koller and Shami [31], KLD is used in a backward selection procedure to find those features which can be removed with minimal information loss; and, in the work of Levner [32], statistical tests are used to compare class distributions to judge whether they differ significantly. Sahiner et al. [33] investigated a genetic algorithm (GA), different from GP (see Sect. 3), for feature selection for computer-aided classification of breast masses and normal breast tissue. The classifier used for measuring the fitness was a linear discriminant classifier and the fitness measure was based on the area under the receiver operating characteristic curve. They used manually extracted ROIs from 168 mammograms, and from each ROI, they extracted a total of 587 features, containing 572 texture features and 15 morphological features. Using a combination of GA and linear discriminant classifier, they achieved an average classification performance of about 89%; the best result was about 92%.

Table 2 Results of selection of features by various procedures

Feature-selection method or statistical test	Ordered list of features[a]								
KLD	8	5	7	6	2	1	13	15	20
K–S test	8	6	5	7	2	4	1	17	10
t test	8	5	6	7	16	2	9	1	17
SFS	8	5	7	6	2	1	16	9	18
SBS[a]	1	2	5	6	7	8	9	16	18

[a] For SBS, the features are given in numerical order rather than order of importance. See Table 1 in Sect. 2 to see which features the numbers correspond to

4.6 Results of feature selection

All of the above procedures—SFS, SBS, and the three statistical tests—represent stand-alone feature-selection methods, i.e., they perform the selection without reference to any classifier. GP performs classification with built-in feature selection. We use the five procedures described above in an initial step to narrow the pool of features for use with GP. To ensure a significant reduction from the original number of 22 features, nine (approximately 40%) of the top features were selected in all of the procedures. The results of the statistical tests and feature selection are presented in Table 2.

An important point to note here is that all of the methods have resulted in almost the same set of features, with a few minor differences and a slightly different ordering. The top four features resulting from the tests (features 5–8) are those that one would choose by simply looking at the distributions of the data with respect to each feature. In Fig. 5 are presented the values of the top four features (5–8) selected by all tests and the values of feature 13 that is selected by none of these tests. It is clear from the figure that separations between benign and malignant masses are far more pronounced in the cases of features 5, 6, 7, and 8, as compared to the case of feature 13. This demonstrates the efficacy of the tests. In physical terms, the top four features are the shape features used in the study. The relevance of such factors in the diagnosis of breast cancer has already been established by Rangayyan et al. [8, 14]; by Sahiner et al. [9]; and by Alto et al. [16] (who analyzed the same data as in the present study).

In view of the results shown in Table 2, two new feature sets '*Y*' and '*Z*' were created:

Y = {1, 2, 4, 5, 6, 7, 8, 9, 10, 16, 17}, containing 11 features, and *Z* = {2, 5, 6, 7, 8}, containing five features.

See Table 1 in Sect. 2 to see which specific features these numbers correspond to. Although the shape

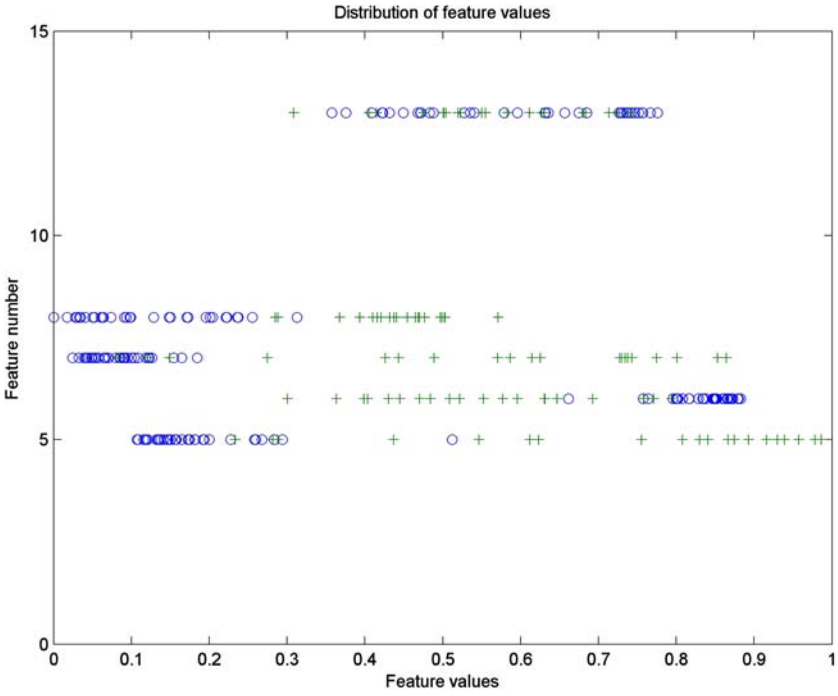

Fig. 5 Distribution of feature values for the 57 masses and tumors in the study. Malignant tumors are labeled by '+'; benign masses are labeled by '*o*'

factors can provide classification accuracies that are far superior to those of the edge-sharpness and texture measures, given the nature of the mammographic images and the radiological definition of the important features that help in distinguishing between benign masses and malignant tumors [34], it is desirable to formulate a set of features that includes measures representing all of the categories of the features mentioned above [16].

It is well to recognize that the search space is greater when the feature set is larger; conversely, the search space is smaller when the feature set is smaller. When running the GP classifier for the feature set Y containing 11 features, the set being larger than the feature set Z, the number of generations was chosen to be 500. For the feature set Z, with about 60% of the features in the feature set Y, the number of generations was chosen to be 300 in the following experiments.

5 Testing

In supervised machine learning, normally, most of the data are used to train the classifier (in the present work, GP), and a small amount of data is set aside for testing. The dataset available for the present study contains 57 ROIs—37 related to benign masses and 20 of malignant tumors, and each ROI is represented by 22 feature values. The number of features is rather large compared with the number of data samples to be classified. In this case of a small, but very well categorized dataset, one cannot afford to set many cases aside for testing. The effects of small sample size in the design of classifiers has been studied by several researchers [15, 35, 36].

One method in common use in situations as above is the leave-one-out procedure [37], where the classifier is trained n times (n = number of data points), each time with a different data point left out for testing. The problem with this method is that leaving out one data point can drastically alter the results of estimation when one has a small amount of data to begin with; furthermore, there is a certain lack of confidence in the robustness in the measured performance.

In this paper, another approach is explored. A well-known statistical approach, bootstrap with resampling [38], is used to obtain statistical measures related to the performance of the classifier. Such a technique has been used in the analysis of breast masses [39]. We performed six experiments as shown in Table 3. Three experiments used the feature set Y and the other three experiments used the feature set Z. With each feature set (Y or Z) and for each condition (benign or malignant), the central m% of the data points in each feature space were selected. Using the sets of data points in each feature space selected as above, a new set of data points was created by performing either the union or the intersection of the sets of data points across all the features.

Table 3 Selection of data samples for classification experiments

Experiment number	Feature set	Union or intersection (U/I)	m
1	Y	U	5
2	Y	I	80
3	Y	I	90
4	Z	U	10
5	Z	I	70
6	Z	I	80

For example, Table 3 indicates that Experiment 1 was based on the feature set Y, and the dataset was obtained from the union of the central 5% of the samples in each feature space. Similarly, Experiment 6 was based on the feature set Z, and the dataset was obtained from the intersection of the central 80% of the samples in each feature space. In each experiment, 100 test sets were created by the well-known statistical procedure called sampling with replacement. Each test set comprised of five malignant tumors and five benign masses.

While the above procedure may not be entirely satisfactory as it does not test the classifier on all the points in a single run, this is exactly the reason for the employment of the bootstrap technique, i.e., sampling with replacement and multiple runs. This is precisely the reason for conducting 100 runs: one can get a good estimate of how the classifier performs overall by examining the training performance as well as the test performance. For the sake of balance, we have also included the results of the leave-one-out procedure for the feature set Z.

6 Experiments and results

There are two aspects to the following results; the performance of the classifier and the features that are found to be important. The results are ordered in terms of how the test set was chosen, i.e., union or intersection, and the value of m; see Sect. 5.

6.1 Experiments

Six experiments were performed, as shown in Table 3.

6.2 Results

The results of classification performance are detailed in Sect. 6.2.1. Analysis of the results of feature selection by GP is recorded in Sect. 6.2.2.

6.2.1 *Classification performance*

The results relating to the performance of the classifier averaged over 100 runs in each experiment are displayed in Table 4.

As mentioned in Sect. 5, it is important to look at both the test performance and the training performance. The last column in Table 4 presents the average overall performance, which is obtained from considering the classification accuracies of the training and test data together. By analyzing the data in Table 4, one can see that the classifier correctly classifies the training data in over 99.5% of the cases and the test data in over 98% of the cases, except for the first experiment where the result is 90.1%.

It should be remarked that the classification accuracies obtained with both feature sets *Y* and *Z* are high. The fact that the classification accuracies obtained with the smaller feature set *Z* is as good or better than the same obtained with the larger feature set *Y* indicates clearly the success of feature selection. These results demonstrate that, although the larger feature set *Y* contains more data than the smaller feature set *Z*, the smaller feature set *Z* contains as much discriminative information as the larger feature set *Y*. Hence, the feature-selection step prior to the GP operation is helpful in that the results with the feature set *Z* appear to be more robust than those with the feature set *Y*.

For the feature set *Z*, a leave-one-out experiment with all the data samples was also performed. The training results are shown in Table 5 and the test results are shown in Table 6. The training performance is essentially perfect, whereas the overall test performance is 96.5%.

Table 4 Results of classification experiments with data sampling with replacement

Experiment description[a]	Average training performance (%)	Average test performance (%)	Average overall performance (%)
1_*Y*_union_5	99.9	90.1	98.2
2_*Y*_intersection_80	99.6	98.4	99.4
3_*Y*_intersection_90	100.0	99.5	99.9
4_*Z*_union_10	100.0	99.6	99.9
5_*Z*_intersection_70	99.9	100.0	99.9
6_*Z*_intersection_80	99.9	100.0	99.9

[a] The notation for the description of each experiment is 'experiment number'_'feature set'_'union or intersection'_*m*

Table 5 Leave-one-out training results for the feature set *Z*

Leave-one-out		Actual	
		Benign (%)	Malignant (%)
GP	Benign	99.9	0.0
Prediction	Malignant	0.1	100.0

Table 6 Leave-one-out test results for the feature set *Z*

Leave-one-out		Actual	
		Benign (%)	Malignant (%)
GP	Benign	97.3	5.0
Prediction	Malignant	2.7	95.0

6.2.2 *Important features selected*

Having demonstrated the high performance of GP, we also explored the selection of important features. To determine which features are important for the purpose of classification using the GP classifier, the percentage of the number of times each feature was selected over all experiments for each feature set was calculated, and the results are presented in Tables 7 and 8. The most commonly selected feature combinations are also recorded for each feature set.

Table 7 shows the percentage of the number of times each feature was selected by the GP classifier in the experiments using the feature set *Y* in Experiments 1, 2, and 3.

Table 8 shows the percentage of the number of times each feature was selected by the GP classifier using the feature set *Z* in Experiments 4, 5, and 6. It is clear that feature 8 (F_{cc}) has been selected almost every single time in each of the experiments with the feature set *Y* or *Z*, and there is no other feature that is as discriminative as this one. Thus, we can make a firm statement that the shape factor fractional concavity is a strong measure for distinguishing benign masses from malignant tumors. This result agrees with the results obtained by Alto et al. [16]; however, the methods used in the two studies are different.

It is also interesting to examine the most common feature combinations for each experiment, as shown in Tables 9 and 10. Table 9 shows the most common feature combinations selected by the GP classifier in experiments using the feature set *Y*, whereas Table 10 shows the most common feature combinations selected by the GP classifier in experiments using the feature set

Table 7 The percentage of selection of each feature in set *Y* in Experiments 1–3

Feature	1	2	4	5	6	7	8	9	10	16	17
Experiment 1	25	1	6	10	16	28	99	29	4	34	7
Experiment 2	14	19	13	22	15	39	90	32	8	19	19
Experiment 3	23	24	19	25	11	19	100	17	8	23	32
Average	20.7	14.7	12.7	19.0	14.0	28.7	96.3	26.0	6.7	25.3	19.3

Table 8 The percentage of selection of each feature in set *Z* in Experiments 4–6

Feature	2	5	6	7	8
Experiment 4	56	45	33	45	99
Experiment 5	54	56	28	53	99
Experiment 6	45	57	35	26	100
Average	51.7	52.7	32.0	41.3	99.3

Table 9 Percentage of occurrence of the most frequent feature combinations (using the feature set *Y*)

Feature combination	8, 16	8, 17	8, 9	2, 5, 8	4, 7, 8
Experiment 1	17	0	9	0	0
Experiment 2	3	2	6	4	4
Experiment 3	10	23	0	8	7
Average	10.0	8.3	4.7	4.0	3.7

Table 10 Percentage of occurrence of the most frequent feature combinations (using the feature set *Z*)

Feature combination	2, 5, 8	6, 8	8	2, 5, 7, 8	5, 8	7, 8
Experiment 4	18	15	8	8	3	7
Experiment 5	18	2	6	12	9	10
Experiment 6	24	13	11	4	11	6
Average	20.0	10.0	8.3	8.0	7.7	7.7

Z. It is worth noting that most of the commonly encountered combinations include features from at least two of the three categories of features used: shape, edge-sharpness, and texture. The results reinforce our observation that it is important to use features representing the different types of radiologically important characteristics of masses. Even though the shape factors perform the best on their own, it would be exceedingly restrictive to rely only upon measures shape to classify breast masses; such an approach would ignore the rich gray-scale and textural information provided by mammograms.

7 Discussion and conclusion

The GP classifier consistently performed well in discriminating between benign masses and malignant tumors using the shape features. The shape measure of fractional concavity was found to be the most important feature: it was selected almost every time by the GP classifier in all experiments. Benign-versus-malignant classification accuracies obtained using various combinations of the shape, edge-sharpness, and texture features varied in the range 90–100%. The results agree with those obtained by Alto et al. [16] using the same dataset but other classifiers such as K-nearest neighbors, Mahalanobis distance, linear discriminant analysis, and logistic regression.

Although the texture features were not favored by the statistical tests or feature-selection methods, two of the texture features in the feature set Y appear in the two most common feature combinations. This shows that the best set of two features is not necessarily made up of the two best features. Also, even if a classifier may not perform well with texture features on their own, it could perform well when using texture features combined with the shape features, a fact observed by Alto et al. [16].

It should be remarked that the estimation of shape factors requires accurate contours, which are not easy to obtain automatically. The contours of the masses employed to derive the features used in the present study were drawn manually by an expert radiologist specialized in mammography; regardless, questions arise regarding the dependence of the results upon the opinion of one expert, as well as the possibility of inter-observer and intra-observer differences. In addition, texture and edge-sharpness are important in radiological diagnosis, and there is a need for defining better features related to these aspects of breast masses in mammograms. It should be noted that the texture and edge-sharpness features used in this work are computed using bands of pixels around the given contour, which reduces the dependence of the features upon the accuracy of the contour to some extent.

The results of the leave-one-out experiment are not as good as those using the other methods of testing used in this paper. Although this is to be expected, the leave-one-out results are adequate to show that the classification of mammographic features related to breast cancer using GP deserves more investigation. One problem in this research work has been the small size of the data set. Although this limitation has been addressed by using the well-known procedure of bootstrap with resampling, the results of analysis with a much larger data set could lead to more confident conclusions. It is worth noting that while GP may require more computation during the training stage than a traditional classifier, such as an artificial neural network, its computational requirement during the testing stage is very competitive.

Acknowledgments This research work was partly funded by the Medical Research Council, UK, through the InterDisciplinary Bridging Awards (IDBA) scheme, and by a grant from the University of Calgary Research Grants Committee. Authors would like to thank Mr. L. Zhang, a research student at the University of Liverpool, for his initial assistance with genetic programming code.

References

1. Page title: Breast Cancer Statistics (2005) Source: UK National Statistics website http://www.statistics.gov.uk/
2. Yaffe MJ (2001) Digital mammography: IWDM 2000, Madison. Medical Physics Publishing, WI
3. Peitgen H–O (2003) Digital mammography: IWDM 2002. Springer, Bremen
4. Rangayyan RM, Ayres FJ, Desautels JEL (2005) Computer-aided diagnosis of breast cancer: toward the detection of early and subtle signs, the 1st world experts' congress on women's health medicine and healthcare. World Academy of Biomedical Technologies, Paris
5. Brzakovic D, Luo XM, Brzakovic P (1990) An approach to automated detection of tumours in mammograms. IEEE Trans Med Imaging 9(3):233–241
6. Kegelmeyer WP, Pruneda Jr JM, Bourland PD, Hillis A, Riggs MW, Nipper ML (1994) Computer-aided mammographic screening for spiculated lesions. Radiology 191(2):331–337
7. Laws KI (1980) Rapid texture identification. In: Proceedings of SPIE, vol 238: Image processing for missile guidance, pp 376–380
8. Rangayyan RM, Mudigonda NR, Desautels JEL (2000) Boundary modeling and shape analysis methods for classification of mammographic masses. Med Biol Eng Comput 38:487–95
9. Sahiner BS, Chan H-P, Petrick N, Helvie MA, Hadjiiski LM (2001) Improvement of mammographic mass characterization using spiculation measures and morphological features. Med Phys 28(7):1455–1465
10. Sahiner BS, Chan H-P, Petrick N, Helvie MA, Goodsitt MM (1998) Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Med Phys 25(4):516–526
11. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybern SMC–3(6):610–621
12. Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67(5):786–804
13. Shen L, Rangayyan RM, Desautels JEL (1993) Detection and classification of mammographic calcifications. Int J Pattern Recognit Artif Intell 7(6):1403–1416
14. Rangayyan RM, El-Faramawy NM, Desautels JEL, Alim OA (1997) Measures of acutance and shape for classification of breast tumors. IEEE Trans Med Imaging 16(6):799–810
15. Sahiner BS, Chan HP, Petrick N, Wagner RF, Hadjiiski L (2000) Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Med Phys 27(7):1509–1522
16. Alto H, Rangayyan RM, Desautels JEL (2005) Content-based retrieval and analysis of mammographic masses. J Electron Imaging 14(2): Article no. 023016, pp 1–17
17. Theodoridis S, Koutroumbas K (2005) Pattern recognition. Academic, New York
18. Pearson K (1901) Principal components analysis. Lond Edinburgh Dublin Philos Mag J Sci 2(2):559
19. Alberta Cancer Board (2004) Screen test: Alberta Program for the early detection of breast cancer, 2001/2003 biennial report, Edmonton, Alberta. http://www.cancerboard.ab.ca/screentest/
20. Mudigonda NR, Rangayyan RM, Desautels JEL (2000) Gradient and texture analysis for the classification of mammographic masses. IEEE Trans Med Imaging 19(10):1032–1043
21. Mudigonda NR, Rangayyan RM, Desautels JEL (2001) Detection of breast masses in mammograms by density slicing and texture flow field analysis. IEEE Trans Med Imaging 20(12):1215–1227
22. Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge, USA
23. Zhang L, Jack LB, Nandi AK (2005) Fault detection using genetic programming. Mech Syst Signal Process 19:271–289
24. Guo H, Jack LB, Nandi AK (2005) Feature generation using genetic programming with application to fault classification. IEEE Trans Syst Man Cybern Part B 35(1):89–99
25. Nordin P, Banzhaf W (1997) Real time control of a khepera robot using genetic programming. Cybern Control 26(3):533–561
26. Kishore JK, Patnaik LM, Mani V, Agrawal VK (2000) Application of genetic programming for multicategory pattern classification. IEEE Trans Evol Comput 4(3):242–258
27. Kudo M, Sklansky J (2000) Comparison of algorithms that select features for pattern classifiers. Pattern Recognit 33(1):25–41
28. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989) Numerical recipes in C. Cambridge University Press, Cambridge, UK
29. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Statist 22(1):79–86
30. Nykter M (2004) Feature selection for Lymphoma outcome prediction. In: Proceedings of the 2nd TICSP workshop on computational systems biology. WCSB'2004, Silja Opera, Helsinki-St. Petersburg 14–16 June, pp 51–52
31. Koller D, Shami M (1996) Toward optimal feature selection. In: Proceedings of the 13th international conference on machine learning. ICML–96, pp 284–292
32. Levner I (2005) Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinf 6:68. doi: 10.1186/1471–2105–6–68
33. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM, Adler DA (1996) Classification of mass and normal breast tissue: feature selection using a genetic algorithm. In: Proceedings of 3rd internatrional workshop on digital mammography, Chicago, pp 379–384
34. American College of Radiology (ACR) (1998) Illustrated breast imaging reporting and data system (BI-RADS), 3rd edn. American College of Radiology, Reston

35. Fukunaga K, Hayes RR (1989) Effects of sample size in classifier design. IEEE Trans Pattern Anal Mach Intell 11(8):873–885
36. Raudys SJ, Jain AK (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 13(3):252–264
37. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
38. Efron B, Tibshirani RJ (1998) An introduction to the bootstrap. CRC Press LLC, Boca Raton
39. Liu Y, Smith MR, Rangayyan RM (2004) The application of Efron's bootstrap methods in validating feature classification using artificial neural networks for the analysis of mammographic masses. In: 26th annual international conference of the IEEE engineering in medicine and biology society, San Francisco. IEEE, CA, pp 1553–1556