Available online at www.sciencedirect.com
ScienceDirect
Applied Soft Computing 7 (2007) 612–625
Applied Soft Computing
www.elsevier.com/locate/asoc

A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms

Brijesh Verma [a,*], Ping Zhang [b]

[a] Central Queensland University, Australia
[b] Bond University, Australia

Received 27 October 2003; received in revised form 21 February 2005; accepted 22 February 2005
Available online 24 February 2006

Abstract

Digital mammography is one of the most suitable methods for early detection of breast cancer. It uses digital mammograms to find suspicious areas containing benign and malignant microcalcifications. However, it is very difficult to distinguish benign and malignant microcalcifications. This is reflected in the high percentage of unnecessary biopsies that are performed and many deaths caused by late detection or misdiagnosis. A computer based feature selection and classification system can provide a second opinion to the radiologists in assessment of microcalcifications. The research in this paper proposes a neural-genetic algorithm for feature selection to classify microcalcification patterns in digital mammograms. It aims to develop a step-wise algorithm to find the best feature set and a suitable neural architecture for microcalcification classification. The obtained results show that the proposed algorithm is able to find an appropriate feature subset, which also produces a high classification rate.

Keywords: Neural networks; Genetic algorithms; Digital mammography; Feature selection

1. Introduction

Breast cancer is a leading cause of cancer death in women between the ages of 40 and 55 [1–6]. Currently, there is no certain way to prevent breast cancer [2]. This is why early detection represents a very important factor in its treatment and consequently the survival rate.

Digital mammography is considered to be the most reliable method of early detection, however, in the early stage, the visual clues are subtle and varied in appearance, making diagnosis difficult, challenging even for specialists. In mammography breast abnormalities are divided into exhibiting microcalcification, circumscribed lesions and spiculated lesions. Microcalcification appears as a small bright spot on the mammogram. Fig. 1 shows the occurrence of microcalcification on digitized mammogram. Most of the minimal breast cancers are detected by the presence of microcalcifications [7]. It is however difficult to distinguish between benign and malignant microcalcifications. To decide whether a suspicious area on a digital mammogram contains benign/malignant microcalcifications, traditionally the tissue has to be removed for examination using breast biopsy techniques. The computer classification system of the microcalcifications can provide a second opinion to the radiologists and reduce the number of unnecessary biopsies. A digital mammogram is created directly by the digital machine or digitized from a conventional mammogram. It brought the possibility of using computer-aided diagnosis system.

Current image processing techniques make microcalcification detection easier, however classification of malignant and benign microcalcifications is still very challenging and a difficult problem [7–65] for researchers. One important factor directly affects the classification result is feature extraction. Researchers spend a lot of time in attempt to find a group of features that will aid them in improving the classification for malignant microcalcifications from benign. In the literature, region-based features [7,24], shape-based features [19,31,32], image structure features [7,20,14,15,33,34], texture based features [35,36], and position related features [35] are described and used for experiments.

One feature taken alone might not be significant for the classification but might be very significant if combined with other features. The whole set of the features may include the redundant or irrelevant information. There can also be estimation

* Corresponding author. Tel.: +61 749 3090 58.
E-mail addresses: b.verma@cqu.edu.au (B. Verma), ping.zhang@bond.edu.au (P. Zhang).

1568-4946/$ – see front matter
doi:10.1016/j.asoc.2005.02.008

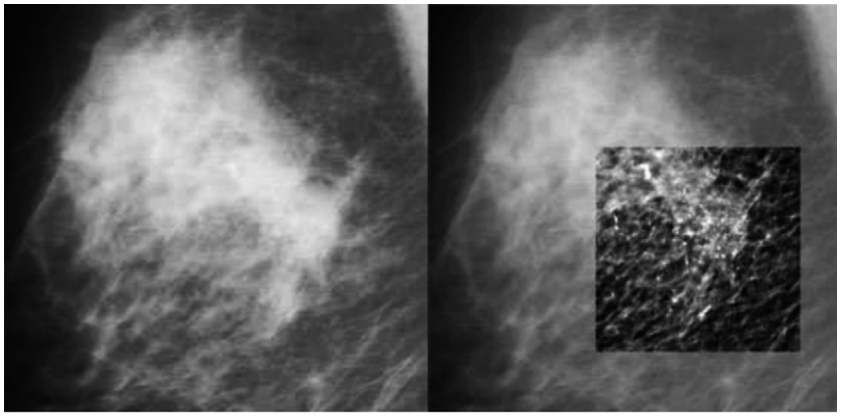

Fig. 1. The occurrence of microcalcifications on mammogram.

errors in the system parameters used to measure the features. Ho [57], combined and constructed multiple classifiers using randomly selected features which can achieve better performance in classification than using the complete set of features.

2. Previous work

Searching for an accurate subset of features is a difficult search problem. The only way to guarantee the selection of an optimal feature vector is an exhaustive search of all possible subset of features. However, search spaces to be explored could be very large. For N features, the number of possible subsets is 2^N. Feature subset selection is defined as a process of selecting a subset of features out of the larger set of features, which maximize the classification performance of a given procedure over all possible subsets. Researchers put lots of effort to find best feature or best combination of features (i.e. feature vector) that gives highest classification rate using appropriate classifier. Search strategies such as *Hill-climbing* and *Best-first search* have been used by Kohavi et al. [60] to find subsets of features with high predictive accuracy. Cost and Salzberg [61] used *feature weighting* technique assigning a real-valued weight to catch feature. The weight associated with a feature, measures its relevance or significance in the classification task. John et al. [62], examined the use of *heuristic* search for feature subset selection. Most of these techniques assume monotonicity of some measure of classification performance and then use branch and bound search. This monotonicity assumption in some form appears to work reasonably well with linear classifiers. However, they can exhibit poor performance with nonlinear classifiers such as neural networks [63].

Racz and Nieniewski [64], employed most discriminative components analysis and a forward/backward selection strategy to reduce the input size from 189 to 46 for his CAD system based on analysis of microcalcifications. Some others [57–59] have explored *randomized* and population based heuristic search techniques such as genetic algorithms to select feature subsets for use with different classifiers. Genetic algorithm (GA) offers a particularly attractive approach to multicriteria optimization, which cannot be handled by most of the other methods. In [30] genetic algorithm was used for feature selection for texture classifier on synthetic aperture radar airborne imagery. They found a few more effective features than the others of image classification. Guerra-Salcedo et al. [58,59] involved genetic strategies for feature selection combined CF/RSC (Common Features/Random Sample Climbing) and Decision Tables dealing with large feature spaces showing a good result.

Overall reviewing the literature, neural networks are particularly effective for fine-tuning solutions once promising regions in the search space have been identified. It is currently used for classification by many researchers. Chitre et al. [7] used a Back Propagation Neural Network for image structure microcalcification classification and compared results with statistical classifiers. Though result is not promising, it is better than the statistical classifiers. Qian et al. [23,55] used the back propagation (BP) algorithm and wavelet transform-based methods with Kalman filtering neural network for mass detection. Verma [15] employed BP with momentum and DSM (Direct Solution Method) based training algorithms to train a feedforward neural network for classification of microcalcification. He achieved the classification rate of 81.25% for benign and malignant. Verma and Zakos [50] developed a computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. They used a fuzzy technique to detect microcalcification patterns and a neural network to classify it. The microcalcification areas from the Nijmegen digital mammographic database were used for their experimentation. Their research achieved a very commendable result with the classification rate 88.9% for classifying the microcalcification as benign or malignant.

Evolutionary algorithms are generally quite effective for rapid global search of large search spaces in multi-modal optimization problems. The use of GA for training neural network (e.g. in [65]) has recently begun to receive a considerable amount of attention. The objective of this paper is to present a neural-genetic algorithm to find the most significant feature or a set of features suitable for classifying abnormalities such as benign and malignant microcalcifications.

The remainder of this paper is organised as follows: Section 3 describes the proposed research methodology followed by the implementation in Section 4. The experimental results are presented in Section 5. Section 6 discusses the obtained results

by the proposed technique. The conclusion and future directions are stated in the final section.

3. Research methodology

The research methodology uses a novel neuro-genetic algorithm proposed in this paper, for finding the most significant feature or a set of features from a number of existing features and classifying the microcalcifications into benign and malignant. An overview of the proposed methodology is shown in Fig. 2.

3.1. Mammographic database

In this research Digital Database for Screening Mammography (DDSM) from University of South Florida is used for experiments. It was downloaded from http://marathon.csee.usf.edu/Mammography/DDSM. The OD (optical density) range of the scanner for the database was 0–3.6. The 12 bits digitizer was calibrated so that the grey values were linearly and inversely proportional to the OD.

In DDSM database, the outlines for the suspicious regions are derived from markings made on the film by at least two experienced radiologists. Each boundary for the abnormality is specified as a chain code, which allows easy feature extraction for each of the suspicious areas in the image files.

3.2. Feature extraction

As mentioned before, feature extraction is a very important part for the classification. To find the best feature or combination of features and get the high classification rate for microcalcification classification is one of the main aims of the proposed research.

The feature extraction technique consists of three parts: (1) area extraction from the marked mammograms; (2) feature extraction from the extracted areas; (3) feature selection for the classification.

3.2.1. Area extraction

Area extraction deals with extracting the grey values from all the suspicious areas in the mammograms marked by the expert radiologists. It accomplished by three steps: (1) according to the chain codes described in the ‘‘.overlay.’’ files of the database, extract the boundary of the suspicious areas, (2) resize the boundary, (3) extract all the grey values in the area and in the boundary area. Fig. 3 shows the whole process of area extraction.

3.2.2. Feature extraction from the extracted area

A set of 14 features is calculated for each suspicious area in this research. Ten of these features are commonly used existing features in the literature, which are histogram, average grey level, energy, entropy, number of pixels, standard deviation, skew, average boundary grey level, difference and contrast. Four of them are modified by us (Verma and Zakos) in our previous research, which are modified energy, modified entropy, modified standard deviation and modified skew.

The formulae for every feature are described below: for each of the formulae: T is the total number of pixels, g an index value of image I, K the total number of grey levels (i.e. 4096), j the grey level value (i.e. 0–4095), $I(g)$ the grey level value of pixel g in image I, $N(j)$ the number of pixels with grey level j in image I, $P(I(g))$ the probability of grey level value $I(g)$ occurring in image I, $P(g) = N(I(g))/T$, and $P(j)$ is the probability of grey level value j occurring in image I, $P(j) = N(j)/T$:

$$\text{histogram} = \frac{1}{K}\sum_{j=0}^{T-1} N(j) \tag{1}$$

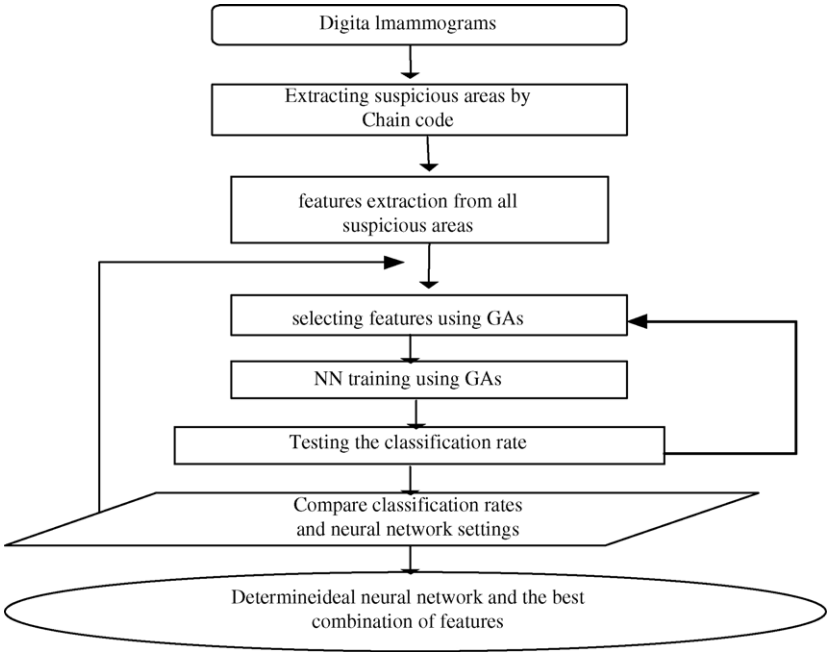

Fig. 2. An overview of the proposed algorithm.

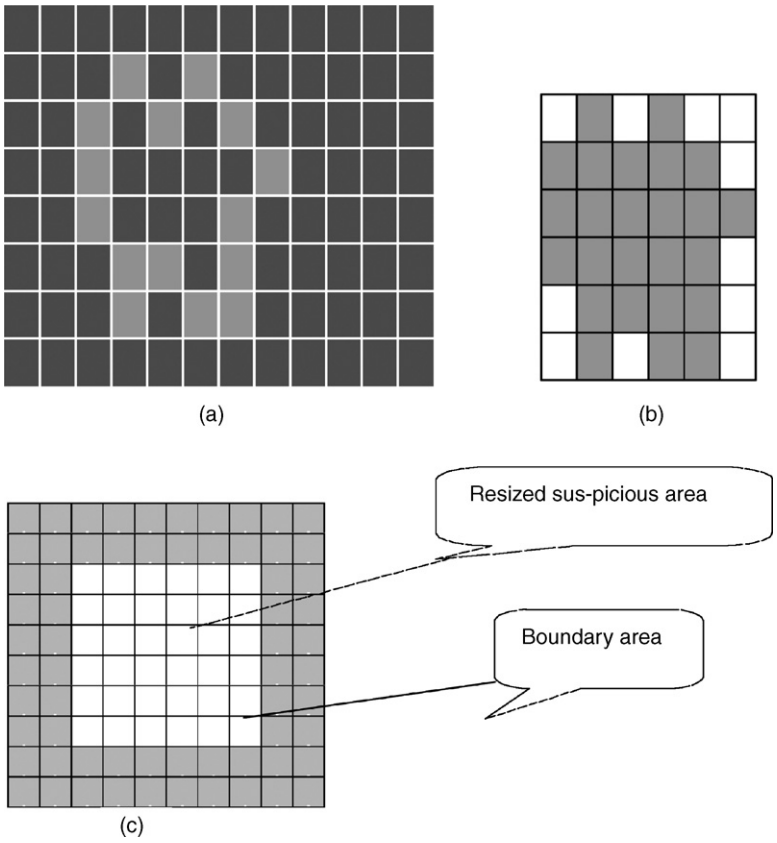

Fig. 3. (a) The boundary of the suspicious area, (b) the area and resized area, and (c) boundary.

$$\text{average grey level} = \frac{1}{T}\sum_{g=0}^{T-1} I(g) \tag{2}$$

$$\text{modified energy} = \sum_{g=0}^{T-1} [P(I(g))]^2 \tag{3}$$

$$\text{modified entropy} = -\sum_{g=0}^{T-1} P(g)\log_2[P(I(g))] \tag{4}$$

$$\text{number of pixels} = \text{count of the pixels in the extracted area} \tag{5}$$

$$\text{modified standard deviation } (\sigma_{\text{m}}) = \sqrt{\sum_{g=0}^{T-1} (I(g) - \text{AvgGrey})^2 P(I(g))} \tag{6}$$

$$\text{modified skew} = \frac{1}{\sigma_g^3}\sum_{g=0}^{T-1} (I(g) - \text{AvgGrey})^3 P(I(g)) \tag{7}$$

$$\text{boundary grey level} = \text{count grey levels at boundary} \tag{8}$$

$$\text{difference} = \text{average grey} - \text{average boundary grey} \tag{9}$$

$$\text{contrast} = \frac{\text{difference}}{\text{AvgGrey} + \text{AvgBoundryGrey}} \tag{10}$$

$$\text{energy} = \sum_{j=0}^{K-1} [P(j)]^2 \tag{11}$$

$$\text{entropy} = -\sum_{j=0}^{K-1} P(j)\log_2[P(j)] \tag{12}$$

$$\text{standard deviation } (\sigma) = \sqrt{\sum_{g=0}^{T-1} (j - \text{AvgGrey})^2 P(j)} \tag{13}$$

$$\text{skew} = \frac{1}{\sigma_j^3}\sum_{j=0}^{K-1} (j - \text{AvgGrey})^3 P(j) \tag{14}$$

3.2.3. Feature selection algorithm

In this research, a neural-genetic algorithm is developed for feature selection based on the neural network pattern classifiers. Each individual in the population represents a candidate solution to the feature subset selection problem. Here, there are 2^{14} possible feature subsets.

In this step, a binary vector of dimension 14 represents the individual in the population. In other words, the chromosome defined contains 14 genes, one gene for each feature, which can take two values. A value of 0 indicates that the corresponding feature is not selected, and a value 1 means that the feature is selected. An initial population of chromosomes is randomly

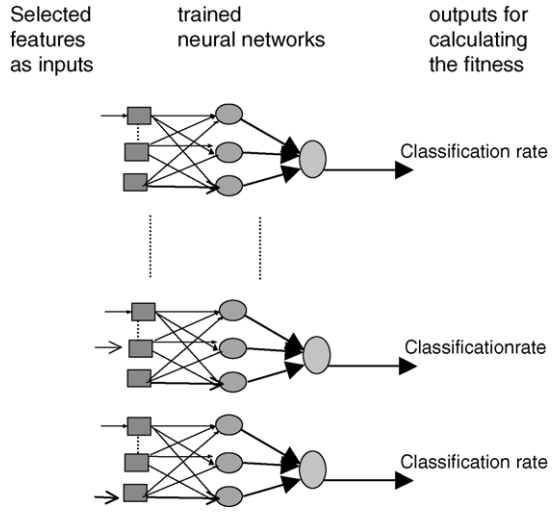

Fig. 4. The NNs trained separately in the process of features selection.

generated. One-point binary crossover and binary mutation are performed. The roulette wheel selection strategy is also used in the algorithm for feature selection. The relevant parameter settings are: population size, 30; number of generation, 200; probability of crossover, 0.8; probability of mutation, 0.2.

The fitness of the chromosome is calculated according to the classification rate of the evolved subset of features, as it is shown in Fig. 4.

3.3. Neural Networks (NNs) for classification

In this research, the selected features are the inputs of the Neural Networks, which are used for classification. The number of the inputs are decided by the automatically selection of GA processing. The values of the inputs are the normalized features that are between 0 and 1. One hidden layer is used in the NN. The nodes of hidden layer were adjusted in an attempt to achieve optimum classification rates. One output of NN is used in the proposed research. The value is also set to be between 0 and 1. The desired output was specified as 0 for benign and 1 for malignant. An output value of an actual NN less than threshold[1] is classified to be benign. That means the relevant input features belong to a benign microcalcification. An output value of more than threshold means that the neural net has classified the input features as belonging to a malignant microcalcification.

The NNs for classification with different selected inputs are trained separately by another genetic algorithm. In the genetic algorithm for feature selection involves many generations. In each generation, evaluation of an individual (a feature subset) involves training neural networks.

A standard genetic algorithm with a roulette wheel selection strategy is used in this research. In the process of NN training, the genes of every individual in the population represent the weights between input and hidden layer and the weights between hidden layer and output of the NN. The results are based on random initialization to the weights of every individual in the population with the following parameters:

- population size, 40;
- number of generation, 50;
- probability of crossover, 0.8;
- probability of mutation, 0.2.

Here crossover is performed by two points real value crossover. Two points (point1 and point2) are selected randomly, where point1 $<$ point2, and point1 $>$ 1, point2 $<$ n, n is the number of genes (here are weights) of one individual NN. For mutation, a small random value between 0.1 and 0.2 is added to every weight of selected member that would perform mutation. After NN is trained the best weights of the NN and the classification rates are saved for the further features selection.

All the programs are implemented using C language on Unix platforms.

4. Implementation

The implementation of the program is divided into four steps: (1) area extraction, (2) feature extraction, (3) feature normalization, and (4) neural-genetic algorithm for feature selection.

4.1. Area extraction program

The area extraction program aimed to extract the microcalcification area from the mammogram. This program reads the information from the ‘‘.ics’’ and ‘‘.overlay’’ files in DDSM database. According to information and the chain code supplied by the ‘‘.overlay’’ files, it extracts suspicious areas from the decompressed image files, and then creates the area and boundary files that include the grey values of all the pixels in the relevant areas. This process is shown in Fig. 5.

4.2. Feature extraction program

Feature extraction program used the area and boundary files as input and extracts the features of every area and creates one file, which includes all the features of every extracted area. Fig. 6 is the block diagram of the program.

4.3. Feature normalization program

Feature normalization program normalizes the features to be real numbers in the range of 0–1. The normalization is accomplished by the following steps: (1) change all the features to be positive by adding the magnitude of the largest minus value of this feature times 1.1,[2] (2) divided by the maximum value of the same feature.

[1] Threshold was initially set to 0.5 as it is the middle value between 0 and 1. However it is not necessarily the most ideal value to achieve the best possible classification rate. In the later experiments it is changed to find the best classification rate.

[2] Here using multiple 1.1 to get rid of 0 values. This step only used when there was the minus value exists.

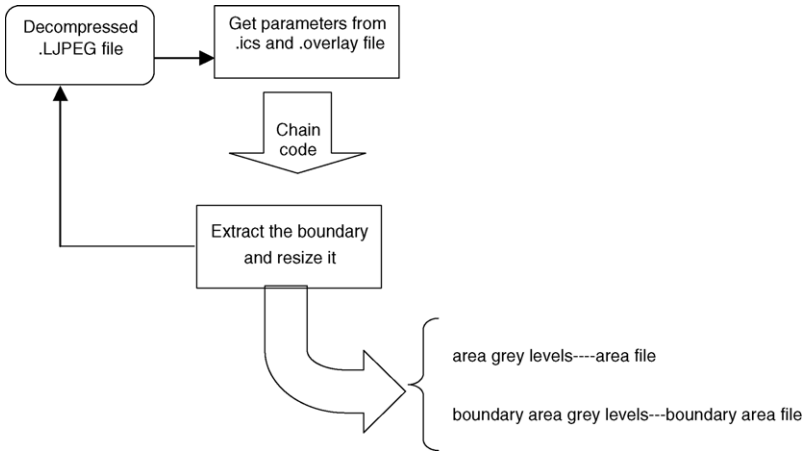

Fig. 5. Area extraction from the mammograms of DDSM database.

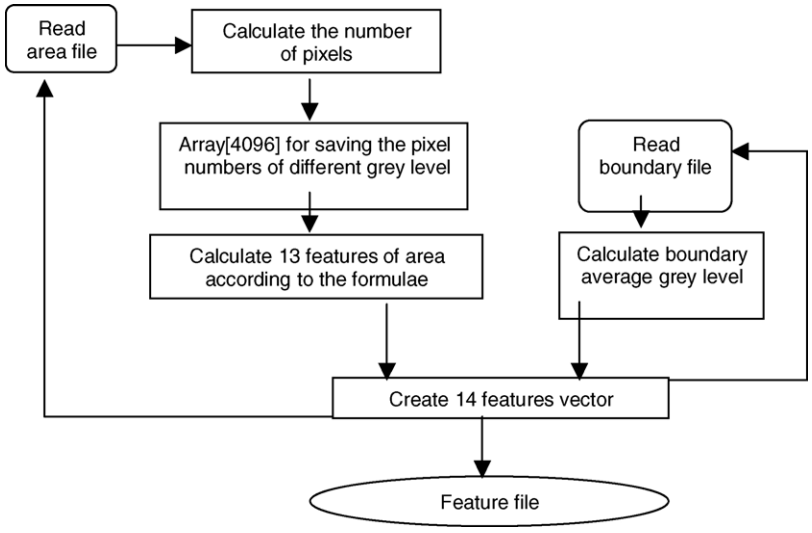

Fig. 6. Diagram of feature extraction program.

For example, there is a set of features as shown in Table 1:

- The normalized the Histogram (Hist) of No. 1 is: 2.420654/2044.85 = 0.001184.
- The normalized skew of No. 1 should be: 1.
- The normalized skew of No. 2 is: $(-0.507747 - 1.1(-1.120872))/(0.22179 - 1.1(-1.120872)) = (-0.507747 + 1.23296)/(0.22179 + 1.23296) = 0.725212/1.45475 = 0.498513$.
- The normalized skew of No. 5 is: $(-1.120872 - 1.1(-1.120872))/(0.22179 - 1.1(-1.120872)) = 0.112087/1.45475 = 0.07705$.

The normalized features are used as the inputs of the proposed neural-genetic approach for feature selection and

Table 1
An example set of features

No.	Hist	AvgGrey	M-entropy	No-pixel	M-S.D.	M-skew	Boundgrey	S.D.	Skew
1	2.420654	3216.655	51.65028	1449	145.7969	0.010777	3033.2036	69.6345	0.22179
2	618.0623	2261.459	557.5256	42217	1331.549	−0.01973	2237.4526	217.2295	−0.507747
3	96.0481	2673.919	288.2817	11177	426.7099	0.020204	2457.2856	91.6372	−0.385337
4	1219.577	2822.901	793.1155	58486	1377.377	0.008712	2501.7749	196.2122	−0.354677
5	1406.039	3191.339	1060.691	44486	1007.179	0.021869	2914.0163	123.9923	−1.120872
6	2572.532	3082.447	1227.898	76777	1560.878	0.001706	2925.9076	179.7458	−0.776106
7	20444.85	2889.106	2961.471	272657	3212.437	0.01601	2272.3514	259.4753	−0.31798
8	650.0071	1409.151	477.5477	56121	1740.77	−0.02035	1174.6264	340.4212	0.052807
9	864.1545	2177.935	665.5077	48409	1373.107	−0.01065	1875.5232	239.8831	−0.972377
10	283.4334	1841.098	336.6776	34067	1411.265	0.025068	1606.6344	282.1944	−0.099654

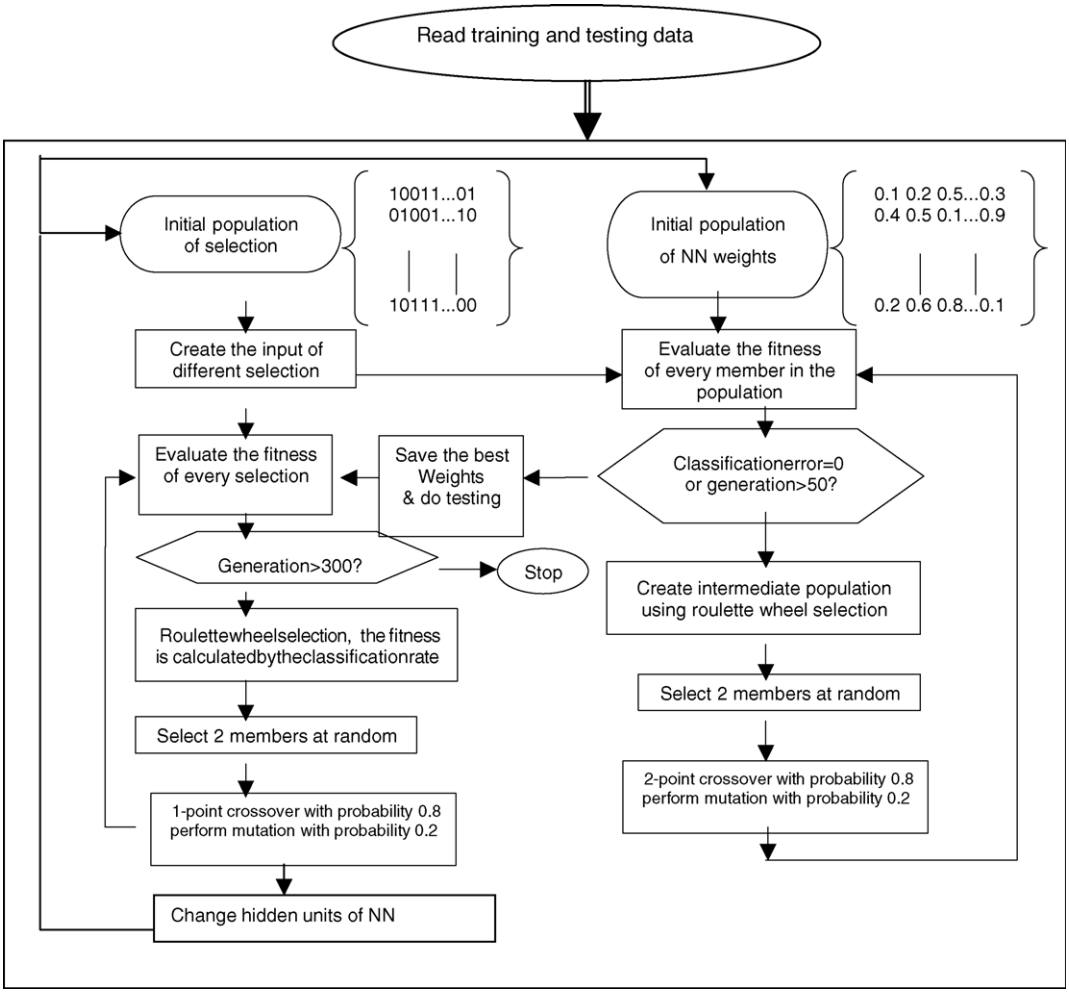

Fig. 7. Diagram of genetic algorithms for feature selection and neural network training.

classification. The program uses the feature file as the input, and creates a file with the normalized features. It can be run with two parameters, one is the feature file name and other is the normalized feature file name to be created.

4.4. Feature selection and classification algorithm

This research involved double genetic algorithms for feature selection and NNs training, respectively. Fig. 7 shows the diagram of this double GAs.

The program reads the parameter settings that are saved in the text file. By changing the settings, more different experiments can be easily done. The program needs the feature file and target output file for training and testing, which are set in the parameter text file.

5. Experimental results

A total of 67 microcalcification areas were extracted from the digital mammograms taken from a Benchmark database for the experiments. The experiments presented here were run using 47 microcalcification areas (24 benign, 23 malignant) for training and 20 microcalcification areas (11 benign, nine cancer) were used for testing.

Many experiments using different parameters were run to find the feature or combination of features that best classifies a microcalcification area into benign and malignant. It was also performed to determine the ideal neural network parameter settings for microcalcification classification with the selected feature set.

The experiments were conducted by the classification rate of testing set to calculate the fitness for reproduction of Genetic feature selection. The number of hidden units and output threshold were adjusted in the experiments to find the combination of the features and NN structures, which can achieve the best classification rate. The results of the experiments are described as following.

In all the tables, the column ‘Features’ was described by using the values 0s and 1s. The 0 means the responded feature is not selected and the 1 means the feature is selected. The sequence of the features is: (1) histogram, (2) average grey level, (3) modified energy, (4) modified entropy, (5) number of pixels, (6) modified standard deviation, (7) modified skew, (8) boundary grey level, (9) difference, (10) contrast (11) energy,

Table 2
The highest classification rate from experiments of different hidden units

Features (1-selected)	Hidden units	Training set ((47 − 24)/23)				Testing set ((20 − 11)/9)			
		B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
10000111110100	2	2	11	13	72.3	1	3	4	80.0
01011111000100	8	6	10	16	66.0	1	3	4	80.0
10100111111010	10	10	6	16	66.0	1	3	4	80.0
00011111010001	12	2	12	14	70.2	1	3	4	80.0
11100111101111	12	5	9	14	70.2	1	3	4	80.0
00111111000011	14	8	11	19	59.6	3	1	4	80.0
00001111100000	16	3	11	14	70.2	1	3	4	80.0
10000110101000	16	3	15	18	61.7	0	4	4	80.0
11001110011100	18	6	9	15	61.7	1	3	4	80.0

Table 3
The feature selection reached classification rate >80% using threshold 0.4

Features (1-selected)	Hidden units	Training set ((47 − 24)/23)				Testing set ((20 − 11)/9)			
		B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
00000110000001	4	8	6	14	70.2	3	1	4	80.0
10011111000111	6	2	15	17	63.8	1	3	4	80.0
10000111000011	8	6	10	16	66.0	0	3	3	85.0
00111011100011	8	11	6	17	63.8	4	0	4	80.0
11000011010111	8	7	9	16	66.0	0	4	4	80.0
11000110110000	12	6	11	17	63.8	1	3	4	80.0
00000110101100	14	3	16	19	59.6	0	4	4	80.0
10001110110110	16	2	13	15	68.1	1	3	4	80.0
01010110110010	18	2	12	14	70.2	1	3	4	80.0

(12) entropy, (13) standard deviation, and (14) skew. The value followed in the column within brackets indicates the number of times the feature subset is selected in all the generations (i.e. 200 generations).

Benign-Error (''B-E'') is used to represent the number of classification errors[3] for benign microcalcifications of the whole training set or testing set. Malignant-Error (''M-E'') represents the number of classification errors for malignant microcalcifications of the whole training set or testing set. Total-Error (''T-E'') refers to the number of classification errors for all the microcalcifications of the whole training set or testing set. ''T-rate'' is the abbreviation of Total-classification Rate. It is calculated by the following formula:

$$\text{''T-rate''} = ((\text{total number of samples}) - \text{''T-E''})/(\text{total number of samples})$$

5.1. Experiments using threshold 0.5 and hidden units 2–18

When analyzing the results of the experiments from two hidden units to 18 hidden units, it is noticed that there are a few features which are more frequently selected than the others are. In every experiment, most of the feature sets that get the highest classification rate are selected many times or are the most frequently selected in all the generations.

Table 2 gathered all the feature combinations, which get the highest classification rate using different number of hidden units. We can see that in every selection, feature numbers 6 and 7 are selected constantly. The eigth feature is almost selected by all of them, too.

Fig. 8 indicates the frequency of every feature occurred in the feature sets which gave the high classification rate.

5.2. Experiments using threshold 0.4 and NN hidden units 2–18

All previous experiments were carried out using 0.5 as the output threshold for classification, because it is the middle value between 0 and 1. It was found from the results that in most of the

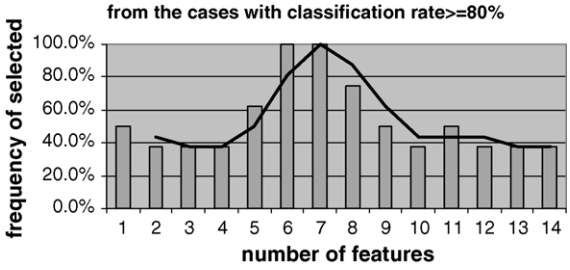

Fig. 8. The selection rate of every feature in the experiments with output threshohold 0.5.

[3] Classification errors mean how many microcalcification areas are not classified correct.

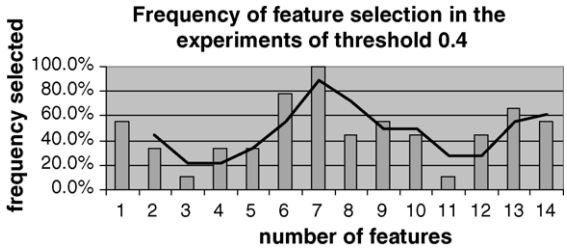

Fig. 9. The selection rate of every feature in the experiments with output threshold 0.4.

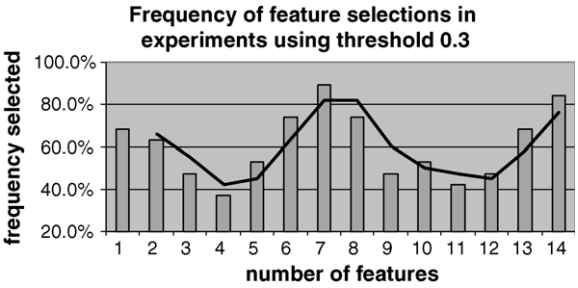

Fig. 10. The selection rate of every feature in the experiments with output threshohold 0.3.

cases the benign classification rates are higher than malignant classification rates. Here change the threshold lower to do the further experiments and find out if it can achieve better results.

Actually, the threshold 0.6 was used to do the experiment too. The result is the same as predicted. It is not better than using threshold 0.5. That is because the benign classification rate is higher than malignant, but not opposite.

The following is the results of the experiments by using output threshold 0.4. Table 3 shows all the feature subsets what achieved the classification rate of testing set not <80%. Obviously, in every feature set include feature number 7. Except two of them, all include number 6. It is interesting that these two selections are the two, which did not reach the highest classification rate in the experiments using eight hidden units.

This result is mainly consistent with that of the experiments using threshold 0.5. Fig. 9 shows the selected frequency of every feature in the high classification rate cases of the experiments using threshold 0.4. The trend line shows the consistent result as it is shown in Fig. 8.

5.3. Experiments using threshold 0.3 and NN hidden units 2–18

Although the feature selection result is consistent in using threshold 0.4 and 0.5, in the experiments using threshold 0.4 achieved the higher classification rate. Further experiments using 0.3 as the output threshold were run in order to get more evidence for the consistent feature selection result, and expected to reach higher classification rate.

Table 4 shows the result of using different NN hidden units with threshold 0.3. The highest classification rate is achieved again in the experiment using eight hidden units, also in using two and six hidden units.

It is not surprising that the feature selection shows the similar result as it appeared in the previous experiments. Number 7, feature 'modified skew' is still most frequently selected, and traditional skew here is selected more than it is in the previous experiments. The frequency of feature selection from the experiments with threshold 0.3 is shown in Fig. 10. We can see the trend line is not changed too much.

So far, a few feature subsets have achieved the highest classification rate 85.0%. More combinations of features classified the testing set with correct rate 80%.

Table 5 shows feature combinations, which reached 85.0% classification rate.

A new character of the results is that all the feature subsets here include traditional standard deviation and skew, beside the most popular feature modified standard skew. Another feature,

Table 4
The feature selections reached classification rate ≥ 80% using output threshold 0.3

Features (1-selected)	Hidden units	Training set ((47 − 24)/23)				Testing set ((20 − 11)/9)			
		B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
11001111011111	2	9	8	17	63.8	2	1	3	85.0
11110011111111	2	1	19	20	57.4	0	4	4	80.0
10001111001011	4	9	6	15	68.1	2	2	4	80.0
11100111011111	6	9	10	19	59.6	2	2	4	80.0
11101111000000	6	9	7	16	66.0	1	3	4	80.0
11111011111011	6	10	6	16	68.1	2	1	3	85.0
00110011010011	8	6	11	17	63.8	1	2	3	85.0
00110101100011	8	7	15	22	53.2	1	3	4	80.0
01010111001101	10	3	15	18	61.7	1	3	4	80.0
01010111001111	10	9	8	17	63.8	1	3	4	80.0
10001110001111	10	11	6	17	63.8	3	1	4	80.0
10001111110001	10	5	13	18	61.7	0	4	4	80.0
11101011100111	10	9	9	18	61.7	3	1	4	80.0
11001110110011	14	10	9	19	59.6	1	3	4	80.0
11010110010011	14	4	13	17	63.8	0	4	4	80.0
11100111010101	14	6	12	18	61.7	1	3	4	80.0
00101111100000	16	2	12	14	70.2	2	2	4	80.0
11001111100111	16	2	12	14	70.2	1	3	4	80.0

Table 5
The feature selections reached classification rate 85% using different number of hidden units and thresholds

Features (1-selected)	Hidden units	Threshold	Training set ((47 − 24)/23)				Testing set ((20 − 11)/9)			
			B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
11001111011111	2	0.3	9	8	17	63.8	2	1	3	85.0
11111011111011	6	0.3	10	6	16	68.1	2	1	3	85.0
00110011010011	8	0.3	6	11	17	63.8	1	2	3	85.0
10000111000011	8	0.4	6	10	16	66.0	0	3	3	85.0

Table 6
Classification rates reached using discriminant classifier with the selected features

Features (1-selected)	Training set				Testing set			
	B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
11001111011111	7	6	13	72.3	1	5	6	70.0
11111011111011	5	6	11	76.6	3	7	10	50.0
00110011010011	9	5	14	70.2	5	5	10	50.0
10000111000011	10	7	17	63.8	2	6	8	60.0

Table 7
Classification rates reached using logistic regression technique with the selected features

Features (1-selected)	Training set				Testing set			
	B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
11001111011111	10	5	15	68.1	3	3	6	70.0
11111011111011	7	3	10	78.7	4	5	9	55.0
00110011010011	9	3	12	74.5	5	5	10	50.0
10000111000011	12	2	14	70.2	5	1	6	70.0

boundary average grey level is also selected by every one of them.

More experiments using 0.2 as threshold were conducted, which achieved the highest classification rate 80.0% with four and eight hidden units. The overall results were not better than in previous experiments.

5.4. Further experiments with statistical classifiers

For further validating the selected features, we had the experiments with discriminant and logistic regression techniques and the results are displayed in Tables 6 and 7. As it can be seen, the results are much worse than neural classifier. We also used the random selection of feature sets and did the experiments using neural network, discriminant analysis and logistic regression methods, respectively. We have never found any other feature subset reached over 80% classification rate with NN. The highest classification rate with the random selected features was 65% and 70% with discriminant classifier and logistic regression, respectively. This somehow proved that GA for feature selection is effective.

6. Discussion and analysis

This section discusses the results presented previously. Furthermore, it analyses the advantages and disadvantages of the proposed approach.

6.1. Discussion on the experimental results

Various results were achieved using various parameters of neural networks and genetic algorithms, as well as the output threshold. We can see every parameter, i.e. the number of hidden units of NN, and the output threshold, can affect the final classification rate.

In this research, the input vector of NN is changing all the time in order to find the best combination of features. One of the criteria of the best feature set is to get the best classification rate. However, the best classification achievement is affected by many factors, as mentioned above. In addition, the algorithms used for NN training and feature selection are also very important.

Look back the experimental results shown in Section 5. When doing the experiments using the classification rate of training set as the fitness function to reproduce the next generation for feature selection, the selection result did not converge to certain set of a few feature subsets that gave the best classification rate. We could not get any conclusion from here. To find the reason, in the process of feature selection, every NN for classification is trained separately. It is explainable that we cannot find the consistency, because the NN, which is trained better does not always give the better testing classification rate.

The better selection convergence and good classification rate happened in the experiments with the following method and parameters:

- Using the classification rate of testing set to be the fitness function for reproducing next generation in the GAs for feature subset selection.
- The number of hidden units is 8–16.
- The output threshold 0.3 is the most suitable threshold for classification, but not the middle value 0.5.
- The generation size for feature selection of this case should not be <200.
- The generation size for NN training can use 50 as usual.

The best classification happened in a few feature subsets. They are:

(1) "11111011111011": Histogram, average grey level, modified energy, modified entropy, pixel numbers, modified skew, boundary average grey level, difference, contrast energy, standard deviation and skew.
(2) "11001111011111": Histogram, average grey level, pixel numbers, modified standard deviation, modified skew, boundary average grey level, contrast, energy, entropy, standard deviation and skew.
(3) "00110011010011": Modified energy, modified entropy, modified skew, boundary average grey level, standard deviation and skew.
(4) "10000111000011": Histogram, modified standard deviation, modified skew, boundary average grey level, standard deviation and skew.

The experiments with the statistical classifiers and the experiments with the random selected features show the effectiveness of GA, although with selected features the highest classification rate is still lower than our previous experiments. In our previous experiments we used the different database with the different cases.

In some researchers' view, the different type of classification errors (benign to malignant/malignant to benign) should be considered differently. The error of classifying the malignant to benign is more serious than classifying the benign to malignant. However, we believe that both errors are serious. Classifying the malignant to benign can cause death of the patient, likewise classifying the benign case to malignant can also destroy the patient's life. It is controversial to say the four subsets mentioned about are a 100% the best the feature sets for the microcalcification classification. However, it is important to note that in all these feature sets we find the following features: modified skew, boundary grey level, standard deviation and skew. It at least tells us that these four features are important for microcalcification classification. Actually, we did the experiments with only these four features for three different classifiers. The classification rates reached with NN was 80% for testing and 61.7% for training. With discriminant classifier, the testing classification rate was 70% with 72.3% for training. Logistic regression reached 60% and 70.2% for testing and training with these four features. These are quite well results comparing with the results of random selected features.

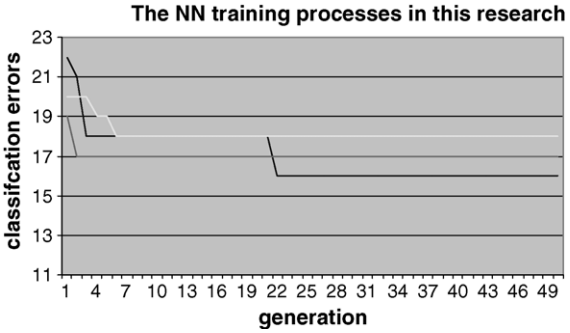

Fig. 11. Three examples of GA process used in NN training.

Reviewing the experiments, there is another feature, which is also frequently selected, so we think it is quite important too, it is modified standard deviation.

6.2. Analysis on double GAs involved in feature selection and NN training

Fig. 11 shows three examples of GA processes used for training the neural networks which were used for the classification.

We can see, in the NN training processes the number of classification errors changed only a few times, but at least it is getting better through all the generations. This indicates that the GA used for NN training is effective. In all the generations, the classification error changed most in the first few generations, it tells us that generation size 50 is appropriate for NN training.

However, the result presented previously showed that the classification rate for training set never reach 90%. For the four feature combinations that reached the highest classification rate for testing, their corresponded classification rates were never more than 70%.

This may have two reasons: one is the database limitation, this can be explained by comparing the number of training set used for training. When using 37 samples for training the classification rate of training set reached 86.5%. When using 47 samples for training the highest training classification rate is only 80.9%.[4] This indicates that it is hard to find a proper NN for the features extracted from the database for classification. Another one may be genetic algorithm itself. Although it is effective, as mentioned, the number of classification errors were not changed too much through all the generations. As Verma and Zakos said: "The good classification rate that the BPNN did achieve meant that there was no need to use another classifier to try and improve the classification rate", the experiments in this research did not show GA was more efficient than BP for NN training.

Fig. 12 traces an entire process of GA used for feature subset selection. It is made according to the record of the experiment using eight hidden units and threshold 0.4. The fitness function for reproduction is the classification rate for testing set. The

[4] The highest classification rate of training set is recorded in the feature selection process using the classification rate of training as the fitness function to reproduce next generation.

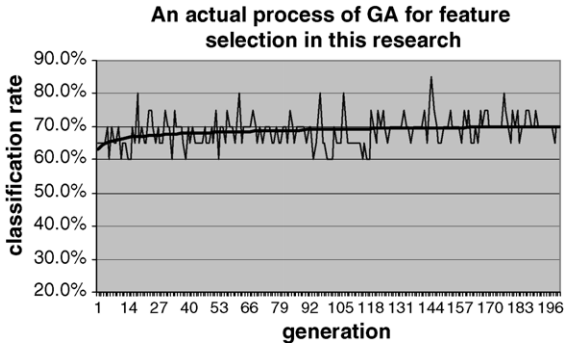

Fig. 12. The entire process for feature selection using testing classification rate as the reproduction fitness function.

main trend line in the figure shows the classification is getting better as the generation number getting bigger.

However in the whole process, the classification rate is not constantly getting better, it goes up and down through all the generations. This is easy to explain. Because in each generation of feature selection, evaluation of an individual involves training the neural network, and every NN is trained separately.

There are a few lucky factors in GA, about initialization, mutation and crossover. This has been proved in the experiments. The experiments have shown that the same feature set can be selected through many generations, but the classification rate is not always the same. That is because every time the NN is trained differently. It can always have some unlucky mutation or crossover happen.

Although, there is some lucky factor affects GA, it is still effective for both NN training and feature selection. The reproduction involved in GA can always keep the best individual to the next generation. By doing more experiments this disadvantage can be weakened.

7. Conclusions and further research

We have proposed and implemented a novel approach for finding the most significant set of features to classify benign and malignant microcalcifications in digital mammograms. After conducting many experiments and analyzing them, the conclusions of this research are drawn below.

7.1. Conclusions

The achieved classification rate and feature sets are promising. The highest classification rate achieved for testing set was 85.0%.

Five features are considered to be the most significant features of a digital mammogram for microcalcification classification. They are modified skew, boundary average grey level, standard deviation, skew and modified standard deviation.

For most experiments, the neural network architecture of eight hidden units is most suitable for classification. In addition 0.3 is proved to be a good output threshold for classification with one output NN.

Our genetic algorithm based feature selection approach is effective. It is also proved by using the best four feature subsets, which are selected by our approach, as the inputs of the neural network trained using backpropagation algorithm. It has got the best classification rate than any other random selected feature combinations.

7.2. Problems and further research

As mentioned in last section, genetic algorithms rely on 'lucky' random crossovers and mutations; so one method for improving the algorithm was to pursue an increase of the probabilities of success of the crossovers and mutations. After all the experiments discussed previously combining the classification rate for testing and training to do feature selection has been considered. Adding some other criteria to calculate the fitness for reproduction is also thought about.

Another big problem has met in this research is that the double GAs training process is quite slow. That is because for every selection individual, the neural network has to be trained separately, and every training process involved many generations and populations. To find a less cost and more effective method to train the neural networks and combine with GA for feature selection can be the next stage of current research.

Although the achieved classification rate for testing is high, it can still be improved, by improving the classification rates of corresponding feature sets for training, which were low. Bigger database will be used and further validation with different split of training and testing set will be in our further experiments.

Acknowledgements

This research work was partially funded by GURD grant. Author would like to thank Griffith University and Central Queensland University for supporting this research and also would like to thank Rinku Panchal for helping in the experiments.

References

[1] Y-me Nation breast, http://www.y-mechicagoland.org/breast_info.htm.
[2] Breast Cancer Facts—2002, http://www.breastcancerfund.org/disease_-facts.htm.
[3] Cancer screening programs—breast cancer, http://www.cancerscreening.nhs.uk/breastscreen/breastcancer.html.
[4] National Breast Cancer Centre, http://www.nbcc.org.au/.
[5] V. Vogel, Breast cancer prevention: a review of current evidence, Ca. Cancer J. Clin. (2000) 156–170.
[6] Breast Cancer Society of Canada, http://www.bcsc.ca/bcsc_html/about/about_historybc.html.
[7] Y. Chitre, A.P. Dhawan, M. Moskowitz, Artificial neural network based classification of mammographic microcalcifications using image structure features, International Journal of Pattern Recognition and Artificial Intelligence, vol. 7, no. 6, World Scientific Publishing, 1993, pp. 1377–1401.
[8] Edward J. Delp, The Analysis of Digital Mammograms: Spiculated Tumor Detection and Normal Mammogram Characterization, http://www.ece.-purdue.edu/∼ace.
[9] Breast Screen Queensland, http://www.health.qld.gov.au/breastscreen/.
[10] Developing Technologies for Early Detection of Breast Cancer: A Public Workshop Summary, IOM Committee for the Early Detection of Breast Cancer: First Workshop, February 9–10, 2000. pp. 1–15.
[11] Lawrence Livermore National Laboratory, http://www.llnl.gov/str/fitchhilite.html.

[12] P. Taylor, J. Fox, A. Todd-Pokropek, Evaluation of a decision aid for the classification of microcalcifications, in: Proceedings of the Fourth International Workshop on Digital Mamography, 1998, pp. 237–245.
[13] F. Rafiee-Rad, H. Soltanian-Zadeh, M. Rahmati, S. Pourabdollah, Microcalcification classification in mammograms using multi-wavelet features, in: Proceedings of the SPIE Annual Meeting, Denver, CO, July, 1999.
[14] B.K. Verma, A neural network based technique to locate and classify microcalcifications in digital mammograms, in: Proceedings of the IEEE World Congress on Computational Intelligence, WCCI'98, Anchorage, USA, (1998), pp. 2163–2168.
[15] B.K. Verma, Comparative evaluation of two neural network based techniques for classification of microcalcifications in digital mammograms, Knowl. Inform. Syst. l (1) (1999) 107–117.
[16] J. Marti, J. Español, E. Golobardes, J. Freixenet, R. Garcia, M. Salamo, Classification of Microcalcifications in Digital Mammograms using Case-Based Reasoning, IWDM2000, Canada, June 11–14, 2000, p. 55, http://www.sunnybrook.on.ca/∼iwdm2000/iwdm2000_abstract_book.pdf.
[17] F. Aghdasi, R. Ward, B. Palcic, Restoration on mammographic imges in the presence of signal-dependent noise, State of the Art in Digital Mammographic Image Analysis, vol. 7, no. 6, World Scientific Publishing, 1994, pp. 42–63.
[18] N.M. Laura, M. Jose, L. Clinton, Automated analysis for microcalcifications in high resolution digital mammograms, http://www-eng.llnl.gov/documents/imaging/jmhspie93.html.
[19] L. Shen, R.M. Rangayyan, J.E.L. Desautels, Detection classification mammography clacifications, in: International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing, 1994, pp. 1403–1416.
[20] J. Zokos, Computer-Aided Diagnosis of digital mammograms using computational intelligence techniques, Honours Thesis, Griffith University, 1998.
[21] N. Karssemeijer, Adaptive noise equalization and recognition of microcalcification clusters in mammograms, State in the Art of Digital Mammographic Image Analysis, vol. 7, no. 6, World Scientific Publish, 1994, pp. 148–166.
[22] M. Karssemeijer, M. Thijssen, J. Hendriks, L. Van Erning (Eds.), Digital Mammography, Kluwer Academic Publishers, Nijmegen, 1998, pp. 237–244.
[23] W. Qian, L.P. Clarke, Adaptive multistage nonlinear filtering and wavelet for medical image enhancement, ICIP (3) (1994) 711–715.
[24] B. Zheng, W. Qian, L. Clarke, Multistage neural network for pattern recognition in mammogram screening, in: Proceedings of the IEEE International Conference on Neural Networks (ICNN), 1994, pp. 3437–3447.
[25] M. Masek, Y. Attikiouzel, C. de Silva, Automatic removal of high intensity labels and noise from mammograms, in digital mammography, IWDM 2000, in: Proceedings of the Fifth International Workshop on Digital Mammography, Toronto, Canada, June 2000, Medical Physics Publishing, Madison, WI, 2001, pp. 576–581.
[26] H. Neiber, T. Mueller, R. Stotzka, Local Contrast Enhancement for the Detection of Microcalcifications, IDWM2000, Canada, p.103, http://www.sunnybrook.on.ca/∼iwdm2000.
[27] Pereira Jr., M. Azevedo, Wavelet Transform Improves Contrast Enhancement of Digital Mammographic Images, IDWM2000, Canada, p. 104, http://www.sunnybrook.on.ca/∼iwdm2000.
[28] J. Yoon, Y. Ro, S. Kim, D. Park, Contrast enhancement of mammography image using homomorphic filter in wavelet domain, IEICE Trans. Inf. Syst. E85-D (1) (2002) 298–303.
[29] G. Lee, M. Bottema, Classification of masses in screening mammograms as benign or malignant, IWDM2000, Canada, June 11–14, 2000, p. 52, http://www.sunnybrook.on.ca/∼iwdm2000/iwdm2000_abstract_-book.pdf.
[30] S. Erich, S. Ferdinand, Computer aided diagnosis of clustered microcalcifications using artificial neural nets, Austria, J. Comp. Inform. Technol. CIT8 2 (2000) 151–160.
[31] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, M.L. Giger, R.A. Schmidt, C.J. Vyborny, K. Doi, Malignant, Benign clustered microcalcifications: automated feature analysis and classification, Radiology 198 (1996) 671–678.
[32] L. Shen, R.M. Rangayyan, J.E.L. Desautels, Application of shape analysis to mammographic calcifications, IEEE Trans. Med. Image 13 (1994) 263–274.
[33] S. Kevin, C. Christopher, W. Kevin, Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography, Int. J. Pattern Recog. Artif. Intel. 7 (6) (1993) 1417–1436.
[34] C. Chris, Y. Tina, Performance Enhanced Genetic Programming, Evolutionary Programming VI, Lecture Notes in Computer Science, 1213, 1997, pp. 87–100.
[35] A. Maria-luiza, Z. Osmar, C. Alexandru, Application of data mining techniques for medical image classification, in: Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in Conjunction with ACM SIGKDD Conference, San Francisco, USA, August 26, (2001), pp. 94–101.
[36] Y. Marcoz, M. Torres-Torriti, Feature Selection using Genetic Algorithms in SAR Airborne Imagery, 2001, http://www.ifrance.com/silkswift/ga/.
[37] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, New York, 1996.
[38] J. Dengler, S. Behrens, J.F. Desaga, Segmentation of microcalcifications in mammograms, IEEE Trans. Med. Image 12 (1993) 634–642.
[39] T. Wang, N.B. Karayiannis, Detection of microcalcifications in digital mammograms using wavelets, IEEE Trans. Med. Image 17 (1998) 98–509.
[40] H. Yoshida, W. Zhang, W. Cai, K. Doi, R. Nishikawa, M. Giger, Optimizing wavelet transform based on supervising learning for detection of microcalcifications in digital mammograms, in: Proceedings of the ICIP-95, 1995, pp. 152–155.
[41] H. Yoshida, K. Doi, R. Nishikawa, Application of the Wavelet Transform to Automated Detection of Clustered Microclacifications in Digital Mammogram, Academic Reports of Tokyo Institute of Polytechnics, 16, 1994, pp. 24–37.
[42] H. Li, Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms, IEEE Trans. Med. Image 16 (6) (1997) 785–798.
[43] N. Karssemeijer, Stochastic model for automated detection of calcifications in digital mammograms, Image Vision Comp. 10 (1992) 369–375.
[44] E. Cernadas, R. Zwiggelaar, W. Veldkamp, T. Parr, S. Astley, C. Taylor, C. Boggis, Detection of mammographic microcalcifications using a statistical model, in: M. Karssemeijer, M. Thijssen, F.M. Hendriks, F.L. van Erning (Eds.), Digital Mammography, Kluwer Academic Publishers, Nijmegen, 1998, pp. 205–208.
[45] R. Nishikawa, M. Giger, K. Doi, C. Vyborny, R. Schmidt, Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms, State in the Art of Digital Mammographic Image Analysis, vol. 7, no. 6, World Scientific Publishing, 1994, pp. 82–102.
[46] M. Salfity, G. Kaufmann, P. Granitto, H. Ceccatto, A Computer-Aided Diagnosis Method for Automated Detection and Classification of Clustered Microcalcifications in Mammograms, http://www.sis.org.ar/sis2000/microcalcifications_mammograms.pdf.
[47] B. Kovalerchuk, N. Conner, J. Ruiz, J. Clayton, Fuzzy logic for formalization of breast imaging lexicon and feature extraction, in: Proceedings of the Fourth Workshop on Digital Mamography, 1998, http://www.cwu.edu/∼borisk/cv/kcrc.pdf..
[48] B. Kovalerchuk, N. Conner, J. Ruiz, J. Clayton, Fuzzy logic in computer-aided breast cancer diagnosis, Anal. Lobul. Artif. Intel. Med. (11) (1997) 75–85.
[49] A. Drijarkara, G. Naghdy, F. Naghdy, Automatic detection of microcalcifications in mammogrtams using a fuzzy classifier, in: Proceedings of the Fifth International Conference on Soft Computing and Information/Intelligent Systems—Methodologies for the Conception, Design and Application of Soft Computing, World Scientific Publishing, USA, 1998, pp. 371–374.

[50] B. Verma, J. Zakos, A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques, IEEE Trans. Inform. Technol. Biomed. 5 (1) (2001) 46–54.
[51] P. Carlos, S. Moshe, A fuzzy-genetic approach to breast cancer diagnosis, Atif. Intel. Med. 17 (1999) 131–155.
[52] H. Barman, G. Granlund, L. Haglund, Feature extraction for computer-aided analysis of mammograms, State in Art of Digital Mammographic Image Analysis, vol. 7, no. 6, World Scientific Publishing, 1994, pp. 128–147.
[53] A. Rooij, L. Jain, R. Johnson, Neural Network Training Using Genetic Algorithms, World Scientific Publishing, 1996.
[54] N. Metin, C. Heang-Ping, S. Berkman, Optimal neural network architecture selection: improvement in computerized detection of microcalcifications, Acad. Radiol. 9 (2002) 420–429.
[55] W. Qian, X.J. Sun, D. Song, R.A. Clarke, Digital mammography: wavelet transform and Kalman-filtering neural network in mass segmentation and detection, Acad. Radiol. 8 (2001) 1074–1082.
[56] J. Hertz, A. Krogh, R. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley Publishing Company, 1991.
[57] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intel. 20 (8) (1998) 832–844.
[58] C. Guerra-Salcedo, D. Whitley, Genetic Approach to Feature Selection for Ensemble Creation, GECCO-99, 1999, http://www.cs.colostate.edu/~genitor/Pubs.html.
[59] C. Guerra-Salcedo, S. Chen, D. Whitley, S. Stmith, Fast and Accurate Feature Selection Using Hybrid Genetic Strategies, GECCO-99, 1999, http://www.cs.colostate.edu/~genitor/Pubs.html.
[60] R. Kohavi, D. Sommerfield, Feature subset selection using the wrapper method: overfitting and dynamic search space topology, in: The First International Conference on Knowledge Discovery and Data Mining, 1995, http://citeseer.nj.nec.com/kohavi95feature.html.
[61] S. Cost, S. Salzberg, A weighted nearest neighbor algorithm for learning with symbolic features, in: Machine Learning, Morgan Kaufmann, 1996, pp. 57–58.
[62] G. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Proceeding of the 11th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, (1994), pp. 121–129.
[63] H. Liu, R. Setiono, Aprobabilistic approach to feature selection—afilter solution, in: Proceeding of the 13th International Conference on Machine Learning, Morgan Kaufmann, (1966), pp. 319–327.
[64] J. Racz, M. Nieniewski, Computer Aided Diagnosis based on Analysis of Microcalcifications, IDWM2000, Canada, http://www.sunnybrook.on.ca/~iwdm 2000.
[65] R. Bevilacqua, N. Lanconelli, Optimization of a distributed genetic algorithm on a cluster of workstations for the detection of microcalcifications, International Journal of Modern Physics C, vol. 12, no. 1, World Scienctific Publishing, 2001, pp. 55–70.