

Copyright © 2009 Institute of Electrical and electronics Engineers, Inc.

All Rights reserved.

Personal use of this material, including one hard copy reproduction, is permitted.

Permission to reprint, republish and/or distribute this material in whole or in part for any other purposes must be obtained from the IEEE.

For information on obtaining permission, send an e-mail message to [stds-igr@ieee.org](mailto:stds-igr@ieee.org).

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Individual documents posted on this site may carry slightly different copyright restrictions.

For specific document information, check the copyright notice at the beginning of each document.

# Impact of soft clustering on classification of suspicious areas in Digital Mammograms

#Peter McLeod<sup>1</sup>, Brijesh Verma<sup>2</sup>

School of Computing Sciences, Central Queensland University  
Rockhampton, QLD 4701, Australia <sup>1</sup>peterm@practical.com.au

<sup>2</sup>b.verma@cqu.edu.au

## Abstract

*This paper investigates a soft cluster based approach for determining the impact of soft clustering on the training of a neural network classifier for the classification of suspicious areas in digital mammograms. An approach is proposed that first creates soft clusters for each available class and then uses soft clusters to form subclasses within benign and malignant classes. The incorporation of soft clusters in the classification process is designed to increase the learning abilities and improve the accuracy of the classification system. The experiments using soft clusters based proposed approach and a standard neural network classifier have been conducted on a benchmark database. The results have been analysed and presented in this paper.*

## 1. INTRODUCTION

Screening mammography has long been regarded as the gold standard in detecting breast cancer in women. Mammography has been shown to reduce the mortality rate by as much as 41% according to one South Australian study [1]. However various studies have demonstrated that an estimated 11-25% of breast cancers are missed [2] during screening mammography.

In the United States alone an estimated 240,510 new cases of breast cancer together with an estimated mortality of 40,460 during 2007 will occur [3]. Survival from breast cancer is dependant on the stage at which it is detected and the implementation of appropriate treatment. Early stage detection and treatment results in a 98% survival rate however this plummets to 28% if metastases have spread to distant organs [4].

Different studies have demonstrated that Computer Aided Diagnosis (CAD) of breast cancer can improve the detection rate from 4.7% to 19.5% compared to radiologists [5-9]. Various solutions have been proposed for breast cancer diagnosis including artificial neural network [10-11], bayesian networks [12], case based systems [13] and statistical classifiers.

Manrique et al. [14] utilised a genetic algorithm radial basis function network with masses from a dataset from the Madrid

hospital to obtain 83% classification accuracy (with 83% specificity and 81% sensitivity). Although their accuracy was not high their network converged quickly. Halkiotis et al. [15] used the MIAS database and a Multi Layer Perceptron (MLP) type neural network to obtain a good classification rate of 94.7% with an average of 0.27 false positives per image for microcalcifications. Georgiou et al. [16] utilised morphological features with a support vector machine to obtain 91.54% classification accuracy on masses. Brem [17] used the second look CAD system (version 3.4) to determine the performance of CAD systems on different sized lesions and micro-calcifications to achieve an overall sensitivity of 89%. Brem's investigation was to try and determine if lesion size would adversely affect the performance of a CAD system. Abdalla et al. [18] used textual features with a support vector machine classifier to achieve a classification accuracy of 82.5% on mammograms from the Digital Database of Screening Mammography (DDSM) [19]. Panchal et al. [20] utilised the DDSM [19] together with an auto-associator-MLP based classifier and attained 90.9% accuracy on micro calcifications, while Masotti [21] used a Support Vector Machine and obtained 90% classification accuracy on masses. Archarya [22] obtained a sensitivity of 91.67% using an artificial neural network and 95% using a Gaussian Mixture Model with 93.33% and 96.67% specificity respectively on micro calcifications from the DDSM [19]. In a review of CAD techniques Rangayyan et al. [23] noted that several methods have good sensitivity (> 85%) for the identification of masses but also have a high false positive rate. In general mass segmentation is a more difficult task than microcalcifications because masses are variable in size, shape and density can exhibit poor image contrast and can be strongly intertwined with surrounding tissues making detection and classification difficult [24]. Micro calcifications tend to be of a higher density and are more readily detected. Of these CAD techniques artificial neural networks have demonstrated their capabilities, but despite this, successful commercial CAD systems are not available. The main problem of developing an acceptable CAD system is inconsistent and low classification accuracy.

In order to improve the training process and accuracy, this paper investigates a novel technique that uses clustering to create soft clusters (sub classes) within existing classes

(benign and malignant) and incorporates these soft clustering based new classes within a training process.

This paper is organised as follows: section two describes the proposed approach. Section 3 presents the results obtained. Section 4 presents a comparison of the results obtained with that of other researchers and section 5 presents the conclusions.

## 2. PROPOSED APPROACH

The approach proposed in this research is to utilise a combination of clustering and neural network classifier to determine if the classification accuracy of the system can be improved through soft cluster enhancements to the classifier. An overview of the proposed approach is presented in Figure 1 below, together with a discussion.

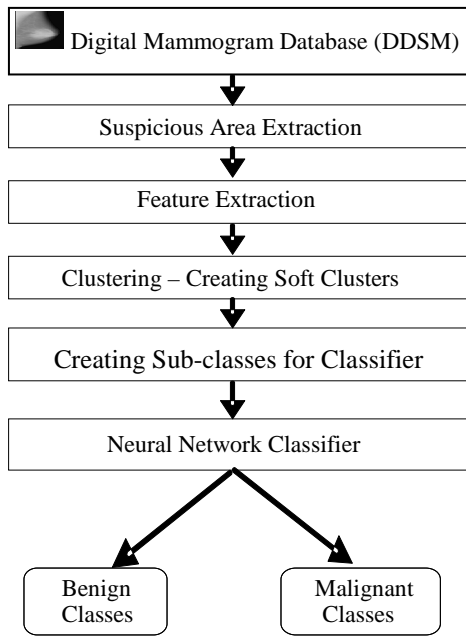


Fig. 1: Overview of the proposed approach.

### A. Stepwise Research Methodology

- Step 1. Extract suspicious areas from digital mammograms
- Step 2. Extract 6 features from suspicious areas
- Step 3. Create inputs and targets for benign class
- Step 4. Create inputs and targets for malignant class
- Step 5. Create  $n$  soft clusters within benign class
- Step 6. Create  $n$  soft clusters within malignant class
- Step 7. Assign inputs to one of  $2n$  ( $n+n$ ) classes based on  $2n$  soft clusters
- Step 8. Train/test classifier with above created inputs and  $2n$  classes for each input.

### B. DDSM Database

The mammograms utilised for this research were obtained from the benchmark Digital Database for Screening Mammography (DDSM) [19]. The database contains approximately 2600 high quality images which can be downloaded from the DDSM website (located at: <http://marathon.csee.usf.edu/Mammography/Database.html>) together with case related information which is available in order to promote research into CAD systems. The use of such a database aids in comparison with other researchers. Both the testing and training set are comprised of an equal number of benign and malignant classes. The benign and malignant classes comprise an equal number of masses with 100 mammograms being selected for training purposes and 100 for testing purposes.

### C. Area Extraction

The method of only utilising a sub-area of a mammogram for classification purposes is called Area Extraction or image segmentation. Images obtained from the DDSM [19] contain a chain code which allows for the extraction of the mammographic abnormality which assists in reducing system resources when performing the classification process. The Region of Interest (ROI) represents both malignant and benign abnormalities.

### D. Feature Extraction

A group of six features have been utilised in this research which represents four BI-RADS descriptor features together with patient age and a subtlety value [19]. All six features are:

- Density
- Mass Shape
- Mass Margin
- Abnormality Assessment Rank
- Patient Age
- Subtlety Value

### E. Clustering- Creating Soft Clusters

Clustering involves the partitioning of a set of data into smaller similar groups to find the natural groups based on similarity of input features. This process is achieved through evaluating the distance between a point and the cluster centroid. The output from the clustering algorithm represents a statistical description of the cluster centroids with the number of components in each cluster. K-means has been incorporated in the proposed methodology. A distance of dissimilarity is calculated. Clustering relies on placing an abnormality in a cluster if their feature expressions are similar as judged by the distance measure employed by the clustering algorithm. K-means uses a minimum “within-class sum of squares from the centres” criterion to select the clusters. The number of clusters that the data is to be partitioned into must be selected in advance. When utilising clustering with breast cancer data for diagnostic purposes it is tempting to try and

specify only two clusters but this is ineffective in that the input features are only weakly correlated with the benign/malignant class pattern. In reality a larger number of clusters are needed in order to accommodate the similar natural feature groups and to separate these different groups. In this work the clustering is used to improve the learning capabilities of the feed forward backward error propagating multi-layer perceptron network.

#### F. K-means Algorithm

K-means is one of the simplest unsupervised learning algorithms [25]. The procedure follows a simple way to classify a given data set into a certain number of clusters which are fixed a priori. It starts by defining  $k$  centroids representing one for each cluster. Following this each point belonging to a dataset is associated with the nearest centroid. When no more points are pending the first step is done and we have our initial grouping. Now we recalculate  $k$  new centroids as the barycentre of the clusters from the preceding step. After this has been done a binding is done between the data set points and the nearest new centroid position. This process iterates through until the centroids do not move. Essentially the algorithm minimises an objective function:

$$J = \sum_{i=1}^k \sum_{j=1}^n \|p_i^{(j)} - c_j\|^2$$

where  $k$  is the number of clusters,  $\|p_i^{(j)} - c_j\|^2$  is a chosen distance measure between the cluster centre  $p_i^{(j)}$  and a data point  $c_j$  which is an indicator of the distance of the  $n$  data points from their respective cluster centres. The general algorithm is composed of the following steps:

1. Place  $k$  points into the space represented by the mammographic abnormalities that are being clustered. These points represent the initial group centroids.
2. Assign each abnormality to a group that has the closest centroid.
3. When all objects have been assigned recalculate the positions of the  $k$  centroids.
4. Repeat steps 2 and 3 until the centroids no longer move.

K-means does have a disadvantage in that there is no general solution to find the optimal number of clusters for any given data set. Normally multiple iterations with different numbers of clusters ( $k$ ) are performed and the best output result is obtained based on the required criterion at the time.

K-means is incorporated in the proposed approach for creating soft clusters within benign and malignant classes.

#### G. Neural Network Classifier

It is through the training process that the input data is mapped to the output vectors and through this process that the network acquires knowledge in the problem domain in question. Once

trained the network is then able to generalise using the acquired knowledge to solve different problems possessing similar characteristics [25]. Thus a neural network maintains knowledge about the problem domain by the weighted interconnections that were used to train the network. Neural networks are able to capture the complex relationship of variables better than many other models because they can capture the non-linear relationship of the training data [26].

Neural networks have been demonstrated to be suitable to the dilemma of classifying medical conditions such as breast cancer [10-11, 27-28]. In fact neural networks have demonstrated their superiority in these fields as their capacity to classify abnormalities as either benign or malignant based on input features that are only weakly correlated with the diagnosis is better than other techniques [27-28]. However the literature also demonstrates that neural networks can suffer from various problems which restrict their efficacy [14].

Traditionally MLP style networks would be utilised in a classification dilemma such as this where a backpropagation of an error component (such as Least Mean Square) is passed back in order to adjust the network weights. Once the error is reduced to a threshold value the network is trained. However it has been noted that a reduction in a component such as Least Mean Square doesn't always lead to an improvement in the classification accuracy of the network [25].

The neural network classifier that has been employed in this research is a single layer multi-layer perceptron style classifier. It utilises six input nodes to represent each input feature. There is only a single hidden layer the number of neurons being determined experimentally to ascertain the optimal configuration. In the proposed approach, the neural network has a variable number of output neurons to represent the natural tendency of the output to group into more than a straight benign or malignant class. The weights of the neural network define the relationship between the input features and the benign or malignant classification. However due to the weak correlation between the input features and the class patterns this relationship is oversimplified by having only two output classes. In its simplest form it could be thought that a set of input features could represent a group such as spiculated malignant or lobulated malignant masses (the expectation would be that the groupings would be more complicated than this) and incorporating two such separate clusters (2 sub-classes within a class) may improve network training and accuracy.

### 3. EXPERIMENTAL RESULTS

The proposed approach has been implemented in C++ on the Windows platform. A number of experiments were performed utilising the standard neural network classifier (MLP-Multilayer Perceptron) and the proposed approach. The experimental results are presented in Table 1.

TABLE 1: CLASSIFICATION ACCURACY OBTAINED

Technique	Hidden Units	Performance	
		Accuracy on Training Set [%]	Accuracy on Test Set [%]
Standard MLP	10	86	94
	16	92	93
	27	96	93
Proposed Approach	16	88	95
	24	94	96
	27	93	96

#### 4. COMPARATIVE RESULTS

The proposed approach has been compared with standard MLP and other existing approaches. In many instances a comparison between one CAD system and another is not an easy task since many factors can affect the classification accuracy of the system. Many neural networks provide a decision threshold for determining the difference between a benign and malignant class and this variable is typically going to be different for different classifiers. Some networks are tested on masses while others on microcalcifications yet others still have representatives from both classes. As well as this the training sample itself can lead to performance differences and even using the same benchmark database (if one is utilised) can still lead to different results due to different training samples being taken from the database. Table 2 details the classification accuracies obtained by different researchers employing different techniques for the diagnosis of breast cancer. As shown in Tables 1 and 2, the results obtained (96% on test set) from this research are comparable with that obtained by other researchers, especially when it is noted that some of the results in Table 2 are only for microcalcifications which are traditionally easier to classify than masses or a combination of masses and microcalcifications. Overall, the proposed approach has outperformed the standard MLP and other existing techniques.

TABLE 2: CLASSIFICATION ACCURACY FROM OTHER RESEARCH TECHNIQUES

Author	Technique	Dataset	Accuracy
Manrique [14]	Genetic Algorithm, Radial Basis Function	Masses	83%
Halkiotis [25]	MLP	Micro.	94.7% (0.27FP per image)
Georgiou [16]	SVM	Masses	91.54%
Panchal [20]	Auto-associator MLP	Micro.	90.90%
Masotti [21]	SVM	Masses	90.00%

#### 5. CONCLUSIONS

This paper has investigated a soft cluster based approach for the classification of suspicious areas in digital mammograms. The results of investigation show that the soft clustering has a significant impact on improving overall classification accuracy. The results presented were obtained with 3 soft

clusters (3 sub classes) for each class (Benign/Malignant). Further research is needed to investigate the appropriate size of clusters and sub classes.

#### ACKNOWLEDGEMENT

This work is supported by an ISSNIP Research collaboration award.

#### REFERENCES

- [1] Roder, D., Houssami, N., Farshid, G., Gill G., Luke., Downey, P., Beckmann, K., Iosifidis, P., Grieve, L. and Williamson, L., "Population screening and intensity of screening are associated with reduced breast cancer mortality: evidence of efficacy of mammography screening in Australia", *Breast Cancer Research and Treatment*, vol. 108, no. 3, pp. 409-416, 2008.
- [2] Goergen, S., Evans, J., Cohen, G. and Macmillan, J., "Characteristics of breast carcinomas missed by screening radiologists", *Radiology*, vol. 204, no. 11, pp. 131-135, 1997.
- [3] American Cancer Society, *Breast Cancer facts and Figures 2007-2008*, American Cancer Society, Inc, Atlanta, 2007.
- [4] American Cancer Society, *Global Cancer facts and Figures 2007*, American Cancer Society, Inc, Atlanta, 2007.
- [5] Brem, R., "Clinical versus research approach to Breast Cancer detection with CAD: Where Are We Now?", *American Journal of Roentology*, vol. 188, pp. 234-235, 2007.
- [6] Freer, T. and Ulissey, M., "Screening Mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center", *Radiology*, vol. 220, pp. 781-786, 2001.
- [7] Dean, J. and Ilvento, V., "Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers", *American Journal of Roentology*, vol. 187, pp. 20-28, 2006.
- [8] Birdwell, R., Bandodkar, P. and Ikeda, D., "Computer-aided detection with screening mammography in a University hospital setting", *Radiology*, vol. 236, pp. 451-457, 2005.
- [9] Morton, M., Whaley, D., Brandt, K. and Amrami, K., "Screening mammograms: interpretation with computer-aided detection – prospective evaluation", *Radiology*, vol. 239, pp. 204-212, 2006.
- [10] Cheng, H., Cai, X., Chen, X., Hu, L. and Lou, X., "Computer-aided Detection and Classification of microcalcifications in Mammograms: a survey", *Pattern Recognition*, vol. 36, pp. 2967-2991, 2003.
- [11] Cheng, H., Shi, X., Min, R., Ju, L., Cai, X. and Du, H., "Approaches for automated detection and classification of masses in mammograms", *Pattern Recognition*, vol. 39, no. 4, pp. 464-668, 2006.

- [12] Ramirez, N., Acosta-Mesa, H., Carillo-Calvert, H., Nava-Fernandez, L. and Barrientos-Martinez, R., "Diagnosis of Breast Cancer using Bayesian networks: A case study", *Computers in Biology and Medicine*, vol. 37, pp. 1553-1564, 2007.
- [13] Tourassi, G., Haarawood, B., Sing, S., Lo, J. and Floyd, C., "Evaluation of information-theoretic similarity measures for content based retrieval and detection of masses in mammograms", *Medical Physics*, vol. 34, pp. 140-150, 2007.
- [14] Manrique, D., Rios, J. and Rodriguez-Paton, A., "Evolutionary system for automatically constructing and adapting radial basis function networks", *Neurocomputing*, vol. 69, pp. 2268-2283, October 2006.
- [15] Halkiotis, S., Botsis, T. and Rangoussi, M., "Automatic detection of clustered microcalcifications in digital mammograms using mathematical morphology and neural networks", *Signal Processing*, vol. 87, pp. 1559-1568, 2007.
- [16] Georgiou, H., Mavrofarakis, M., Dimitropoulos, N., Cavouras, D. and Theodoridis, S., "Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes", *Artificial Intelligence in Medicine*, vol. 41, pp. 39-55, 2007.
- [17] Brem, R., Hoffmeister, J., Zisman, G., Simio, M. And Rogers, S., "A computer aided detection system for the evaluation of breast cancer by mammographic appearance and lesion size", *American Journal of Roentology*, vol. 184, pp. 893-896, 2004.
- [18] Abdalla, A., Deris, S. And Zaki, N., "Breast Cancer detection based on statistical features and support vector machine", *4<sup>th</sup> International Conference on, Innovation in Information Technology*, pp. 728-730, 2007.
- [19] Heath, M., Bowyer, K., Kopans, D., Moore, R. and Kegelmeyer, P., "The Digital Database for Screening Mammography", *IWDM-2000*, Medical Physics Publishing, 2001.
- [20] Panchal, R. and Verma, B., "Neural association of microcalcification patterns for their reliable classification in digital mammography", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 20, pp. 971-983, 2006.
- [21] Massotti, M., "A ranklet-based image representation for mass classification in digital mammograms", *Medical Physics*, vol. 33, no. 10, pp. 3951-3961, 2006.
- [22] Acharya, R., Ng, U., Chang, Y., Yang, J. and Kaw, G., "Computer based identification of Breast Cancer using digitized mammograms", *Journal of Medical Systems*, , 2008, doi:10.1007/s10916-008-9156-6.
- [23] Rangayyan, R., Ayres, F. and Desautels, L., "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs", *Journal of the Franklin Institute, Special Issue: Medical Applications of Signal Processing*, Part I, Vol. 344, pp. 312-348, 2007.
- [24] Delogu, P., Fantacci, M., Kasae, P. and Retico, A., "Characterization of mammographic masses using a gradient based segmentation algorithm and a neural classifier", *Computers in Biology and Medicine*, vol. 37, pp. 1479-1491, 2007.
- [25] Jones, M., *Artificial Intelligence: A systems approach*, Infinity Science Press, Canada, 2008.
- [26] Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 2<sup>nd</sup> Ed, Pearson Education, USA, 2003.
- [27] Paliwal, M. and Kumar, U., heng, H., Cai, X., Chen, X., Hu, L. and Lou, X., "Neural networks and statistical techniques: A review of applications", *Expert Systems with Applications*, 2008, doi:10.1016/j.eswa.2007.10.005.
- [28] Mazurowski, M., Habas, P., Zurada, J., Lo, J., Baker, J. and Tourassi, G., "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", *Neural Networks*, vol. 21, pp. 427-436, 2008.