

# Computer-aided diagnosis of digital mammography images using unsupervised clustering and biclustering techniques

Mohamed A. Al-Olfe<sup>a</sup>, Fadhl M. Al-Akwa<sup>a</sup>, Wael A. Mohamed<sup>b</sup>, and Yasser M. Kadah<sup>a</sup>

<sup>a</sup>Biomedical Engineering Department, Cairo University, Giza, Egypt

<sup>b</sup> Department of Electrical Engineering, High Institute of Technology, Benha University, Benha, Egypt

e-mail: {al\_olfe2001, f\_alakwa, waelbhit, ymk}@k-space.org

## ABSTRACT

A new methodology for computer aided diagnosis in digital mammography using unsupervised classification and class-dependent feature selection is presented. This technique considers unlabeled data and provides unsupervised classes that give a better insight into classes and their interrelationships, thus improving the overall effectiveness of the diagnosis. This technique is also extended to utilize biclustering methods, which allow for definition of unsupervised clusters of both pathologies and features. This has potential to provide more flexibility, and hence better diagnostic accuracy, than the commonly used feature selection strategies. The developed methods are applied to diagnose digital mammographic images from the Mammographic Image Analysis Society (MIAS) database and the results confirm the potential for improving the current diagnostic rates.

**Keywords:** computer aided diagnosis, clustering, biclustering

## 1. INTRODUCTION

Breast cancer is a second cause of fatality among all cancers for women. However, the etiologies of breast cancer are unknown and no single dominant cause has emerged. Still, there is no known way of preventing breast cancer but early detection allows treatment before it is spread to other parts of the body. Currently, X-ray mammography is the most effective, low-cost, and highly sensitive technique for detecting small lesions resulting in at least a 30% reduction in breast cancer deaths [1]. However, the sensitivity of mammography is highly challenged by the presence of dense breast parenchyma, which deteriorates both detection and characterization tasks [2]. Since the consequences of errors in detection or classification are costly, and given that mammography alone cannot prove that a suspicious area is normal, benign or malignant, breast biopsy techniques are often used to confirm the diagnosis. Unfortunately, false-positive diagnosis causes many unnecessary biopsy procedures to be performed. It has been estimated that only 15–30% of breast biopsy cases are proved to be cancerous. On the other hand, false-negative diagnosis means that a serious tumor remains undetected. Retrospective studies showed that 10–30% of the visible cancers are undetected [3]. Thus, there is a need to develop methods more reliable for automatic classification of mammograms, as a means of aiding radiologists to improve the effectiveness of screening programs.

Among the various types of breast abnormalities which are visible in mammograms, clustered microcalcifications (or ‘calcifications’) and mass lesions, distortion in breast architecture, and asymmetry between breasts are the most dangerous ones. Masses and clustered microcalcifications often characterize early breast cancer [3] that are detectable in mammograms before a woman or the physician can palp them. Masses appear as dense regions of varying sizes and properties and can be characterized as circumscribed, spiculated, or ill defined. On the other hand, microcalcifications appear as small bright arbitrarily shaped regions on the large variety of breast texture background. Finally, asymmetry, and architectural distortion are also very important and abnormalities are difficult to detect. The great variability of the mass appearance along with the other abnormalities in digital mammograms is the main obstacle of building a unified mass detection method [3].

Several methodologies have been developed in order to improve the interpretation of mammograms. Among those, the incorporation of computer aided diagnosis (CAD) provides a tool that works as a second observer to the radiologist. CAD systems have been demonstrated as effective tools for helping radiologist identify malignancies in mammograms.

Various techniques were developed to detect and classify masses and microcalcifications in digital mammograms. Most of these techniques used supervised classification to derive their decision. In spite of the success of these methods in improving the overall diagnosis, the labeling of training image samples limits the classification to classes that are deemed independent ignoring the relationships between different pathology types and the progression of each. As a result, images that represent stage of transition between different types are often misclassified. Therefore, a technique that would look at the unlabeled data and provide unsupervised classes would provide an insight into that thus improving the overall effectiveness of the diagnosis. Moreover, using unsupervised classification methods can be extended to utilize biclustering methods which allow for definition of unsupervised clusters of both pathologies and features [17,18]. That is, for each pathology type, a particular set of features that can diagnose this particular type are defined. This has potential to provide more flexibility, and hence better diagnostic accuracy, than the commonly used feature selection strategies.

Unsupervised classifiers, such as *k-mean* clustering, fuzzy *c-mean* clustering, and self-organizing maps, can be used in the literature to diagnose breast cancer. Kim *et al.* [4] designed a new type of classifier combining an unsupervised and a supervised model and applied to classification of malignant and benign masses on mammograms. The unsupervised model was based on an Adaptive Resonance Theory (ART2) network that clustered the masses into a number of separate classes. Lee *et al.* [5] used *k-means* clustering for classifying unlabeled MRI data. Howard *et al.* [6] used a SONNET self-organizing to produce a taxonomic organization of the mammography archive in an unsupervised manner. Chen *et al.* [7] used *k-means* classifier to classify breast ultrasound images to benign and malignant. Meyer-Baese *et al.* [8] introduced automatic lesion segmentation and classification system based on unsupervised clustering and ICA techniques for breast MRI images. None of these methods targeted the development of a CAD system and also biclustering was never considered as a tool for feature selection.

In this paper, a new CAD system for digital mammograms using unsupervised classification is presented. The potential of using this technique is demonstrated in improving the accuracy of the overall diagnosis.

## 2. METHODOLOGY

The proposed system consists of four stages: preprocessing, feature extraction, feature matrix visualization, and unsupervised clustering/biclustering.

### 2.1 Preprocessing stage

In the preprocessing, the region of interest (ROI) was selected from the digital mammograms images. The data used in our experiments were obtained from the Mammographic Image Analysis Society (MIAS) database [9]. It consists of 322 images belonging to normal, benign and malignant classes. All images have a resolution of  $1024 \times 1024$  pixels and 8-bit accuracy (gray level). They also include the locations of any abnormalities that may be present. The existing data consists of the location of the abnormality (like the center of a circle surrounding the tumor), its radius, breast position (left or right), type of breast tissues (fatty, fatty-glandular and dense) and tumor type if exists (benign or malign). Using the locations of abnormalities supplied by the MIAS for each image, a ROI of size  $32 \times 32$  pixels is extracted with breast cancer centered in the window. We used 100 images for normal cases, 88 images for masses (circumscribed, spiculated, ill-defined, architectural distortion, and asymmetric) cases (51 benign images and 37 malignant images), and 25 images for microcalcification cases (13 benign images and 12 malignant images).

### 2.2 Feature Extraction

A typical mammogram contains a vast amount of heterogeneous information that depicts different tissues, vessels, ducts, chest skin, breast edge, the film, and the X-ray machine characteristics. In order to build a robust diagnostic system towards correctly classifying normal and abnormal regions of mammograms and then classify between benign and malignant regions, we have to present all the relevant information in mammograms to the diagnostic system so that it can discriminate between different pathologies effectively. However, the use of all the information results to high dimensionality of feature vectors that degrade the diagnostic accuracy of the utilized systems significantly in addition to sharply increasing their computational complexity. Therefore, a reduced set of reliable features should be considered that summarize only the relevant information. In our approach, we examined an initial set of 224 features obtained from the ROI. These features can be divided into five categories: wavelet features (136 features) [10], first order statistics features (18 features) [11], second order statistics features (60 features) [12,13], shape features (8 features) [11,14], and fractal dimension features (2 features) [15]. From all the above features we construct a feature matrix with images as rows and features as columns.

### 2.3 Feature matrix visualization

In order to address the problem of feature set reduction, the feature matrix is displayed as a heatmap, where feature values are represented by color map [16]. This technique is commonly used in the area of gene expression data analysis where the gene expression matrix is visualized this way. Given the different scales of different features, it is necessary to normalize all features to the same range in order for their range of values to appear properly on the map. The advantages of feature matrix visualization include the ease of data interpretation by visualization rather than figures, gain better understanding of how features perform with different cases.

### 2.4 Unsupervised Clustering

Unsupervised clustering techniques are able to discover clusters inherent in the data. Among those, the *k-means* clustering algorithm partitions a given data set into  $k$  mutually exclusive clusters such that the sum of the distances between data and the corresponding cluster centroid is minimized. The above distance measure between two data points is taken as a measure of similarity. A number of distance measures can be used depending on the data.

Mathematically, given a set of data vectors  $X = [x_1, \dots, x_n]$  where  $n$  is the number of observations, the *k-means* clustering algorithm groups the data into  $k$  clusters with the aim at minimizing an objective function, a squared error function. Therefore, the *k-means* clustering algorithm is an iterative algorithm that finds a suitable partition which minimizes the sum squared error. The algorithm begins with the initialization of  $k$  cluster centroids. Different approaches in initialization have been suggested. A simple method is to initialize the problem by randomly select  $k$  data points from the given data. The remaining data points are classified into the  $k$  clusters by distance. The centroids are then updated by computing the centroids in the  $k$  clusters [5].

The limitation of clustering algorithms is that all features are given equal weights in the computation of image similarity. However, some features do not contribute information and instead increase the amount of cluster ambiguity. Moreover, this technique assigns each image to a single cluster, whereas images at different degrees of disease progression may be mid-way between clusters (e.g., normal tissue turning into tumor). An effective means for dealing with such overlaps is through the use of biclustering methods. The difference between clustering and biclustering methods is that clustering methods can be applied to either the rows or the columns of the feature matrix, separately whereas biclustering methods, perform clustering in the two dimensions simultaneously. This means that clustering methods derive a global model while biclustering algorithms produce a more effective local model. As a result, biclustering was used in this work to find a set of the images participating in a common pathology of interest while defining a subset of features that best describe this pathology. The biclustering techniques used in this work were based on the BiCAT toolbox [17] and SAMBA algorithm [18]. After computing the bicluster files, BicOverlapper [19] was used to pictorially illustrate the found biclusters.

## 3. RESULTS & DISCUSSIONS

Normal and abnormal (mass and microcalcification) heat maps are illustrated in Fig. 1 and Fig. 3 respectively. From these figures the power of this visualization method becomes evident as the feature sets that show difference between the two categories can be easily identified visually. Fig. 2 and Fig. 4 show the *k-means* results for the normal vs. mass and normal vs. microcalcification images respectively based on the found discriminative set of features. The sensitivity and specificity for these results are shown in table I, which indicates the significance performance of unsupervised clustering.

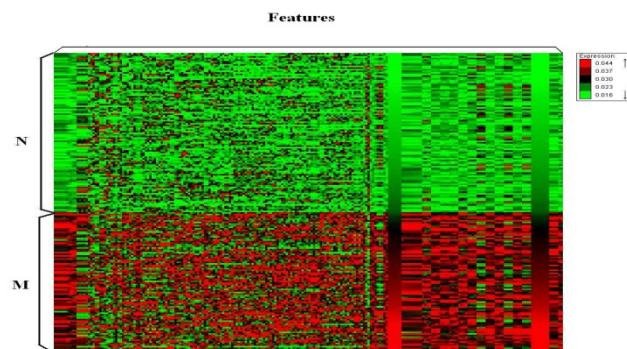


Fig. 1. Feature matrix heatmap for normal(N) and mass(M).

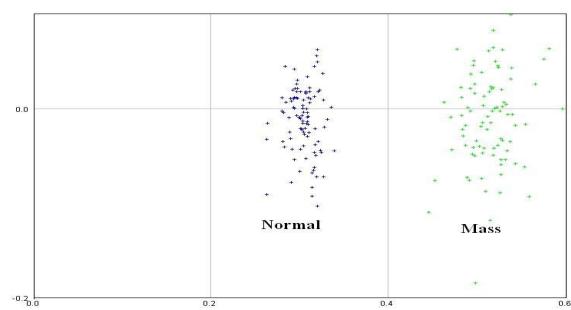


Fig. 2. *k-means* clustering results for normal and mass images.

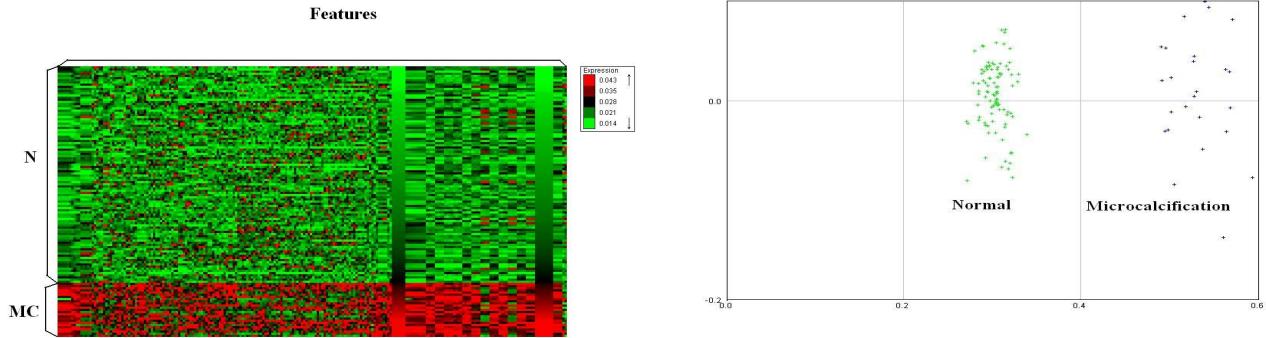


Fig. 3. Feature matrix heatmap for normal(N) and microcalcification(MC).

Fig. 4. *k-means* clustering results for normal and microcalcification.

Table I: *k-means* clustering results for normal, mass, and microcalcification

Types	Sensitivity (Abnormal)	Specificity (Normal)
Normal & Mass	100%	100%
Normal & Microcalcification	100%	100%

The heatmap for feature matrix for mass (37 malignant, and 52 benign) and microcalcification (12 malignant, and 13 benign) images are shown in Fig. 5 and Fig. 6 respectively. Again, from these figures it was clearly visible that the fourteen features (6 from shape feature category and 8 from second order statistical feature category) represent the most significant features that could be used in differentiating between benign and malignant images. Table II shows the sensitivity and specificity after applying *k-means* using all the features compared with using the selected significant features, which indicate that using only these feature has an impact factor in classification between benign and malignant.

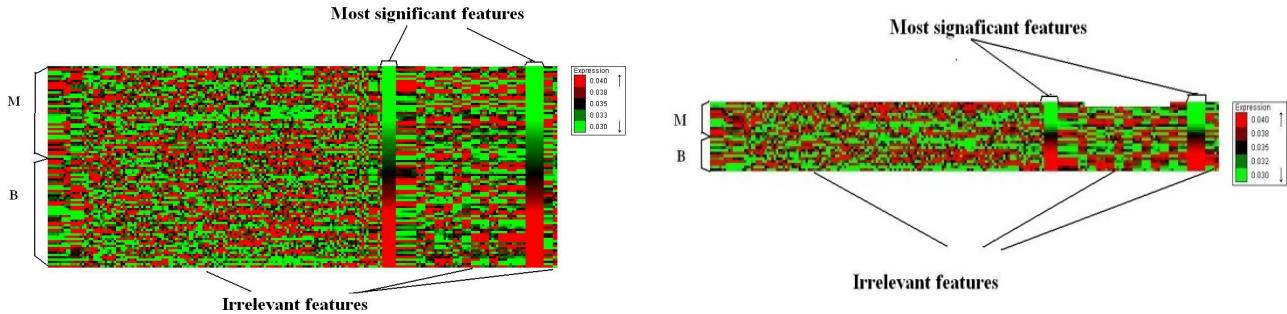


Fig. 5. Feature matrix heatmap for mass images (benign & malignant).

Fig. 6. Feature matrix heatmap for microcalcification images (benign & malignant).

Table II: *k-means* clustering results for benign, and malignant

Types	Clustering using all features		Clustering using significance features	
	Sensitivity (Malignant)	Specificity (Benign)	Sensitivity (Malignant)	Specificity (Benign)
Microcalcification	38.5%	76.9%	91.7%	100%
Mass	48.6%	72.5%	100%	70.6%

Finally, Fig. 7 shows the bicluster results obtained from Bimax [20] algorithm using BioOverlapper program which was applied on the mass feature matrix where images and features are represented by circle and square respectively. The overlapping images and features corresponding to each its pathology are being under our investigation with pathologist

in our group. Also for more interpretation of these results, we try to conduct the database curator to provide us with more pathology details on these images.

To illustrate further the value of using biclustering methods, in Fig. 8 we draw the bicluster number 8 obtained from SAMBA algorithm. All images belong to this bicluster are malignant with overrepresentation of speculated cancer type. Also the feature overrepresented with bicluster are the seven variant moment and difference entropy, which has unknown interpretation in this time.

#### 4. IV. CONCLUSION

In this study, a computer-aided diagnostic system for breast cancer in the digitized mammograms of the breast has been presented. The system provides several features including the visualization of feature matrix by using heat maps, selection of the most significant features to use them in unsupervised clustering, as well as computing the biclustering results of the feature matrix. This system leads to better classification results for the MIAS database data based on the *k-means* method. The system has potential in providing more insight into data and show the value for exploratory data analysis methods.

#### REFERENCES

1. H. Cheng, Y. M. Lui, and R. I. Freimanis, "A Novel Approach to Microcalcification Detection Using Fuzzy Logic Technique," IEEE transactions on medical imaging, vol. 17, no. 3, June 1998.
2. P. M. Sampat, M. K. Markey, and A. C. Bovik, Computer-aided detection and diagnosis in mammography, Handbook of Image and Video Processing, 2nd ed., A. C. Bovik Ed. Academic Press, pp.1195-1217, 2005.
3. I. Christoyianni, A. Koutras, E. Dermatas, and G. Kokkinakis, "Computer aided diagnosis of breast cancer in digitized mammograms," Computerized Medical Imaging and Graphics, vol. 26, pp. 309–319, 2002.
4. J. K. Kim, J. M. Park, S. S. Song, H. W. Park, "Detection of clustered microcalcifications on mammograms using surrounding region dependence method and artificial neural network," J. VLSI Signal Process., pp. 251–262, vol.18, 1998.

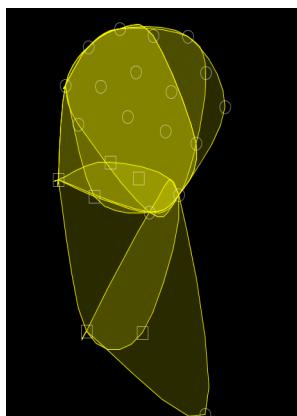


Fig. 7. Biclustering results obtained from Bimax algorithm of mass images where images and features are represented by circle and square respectively.

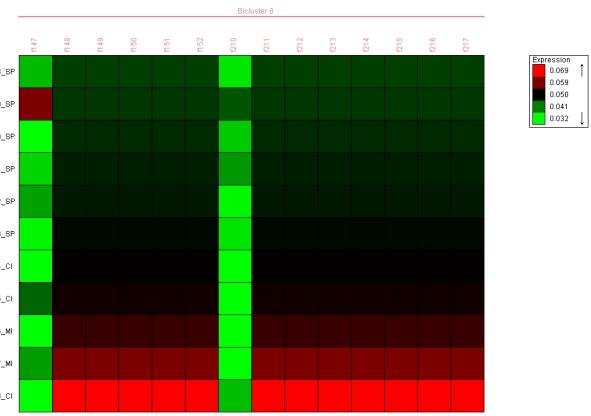


Fig. 8. Bicluster results obtained from SAMBA algorithm of mass images where rows and columns represent images and features respectively.

5. A. Meyer-Baese, O. Lange, T. Schlossbauer and A. Wismuller, "Computer-Aided Diagnosis and Visualization Based on Clustering and Independent Component Analysis for Breast MRI," IEEE International Conference on Image Processing (ICIP 2008), pp. 3000-3003, 2008.
6. J Suckling et al., "The Mammographic Image Analysis Society Digital Mammogram Database," Excerpta Medica, International Congress Series 1069, pp.375-378., 1994. (Available at: <http://peipa.essex.ac.uk/info/mias.html>).
7. R. Mousa, Q. Munib, and A. Moussa, "Breast Cancer Diagnosis System based on Wavelet Analysis and Fuzzy-Neural," Expert Systems with Applications, vol. 28, pp. 713-723, 2005.
8. R. C. Gonzalez, and R. E. Woods, Digital Image processing, Prentice-Hall. Inc., New Jersey, 2002, pp. 76-142.

9. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. System Man. Cybernetics*, vol. SMC-3, pp. 610–621, 1973.
10. R. F. Walker, P. Jackway, and I. D. Longstaff, "Improving co- occurrence matrix feature discrimination, " in Proc. 3rd Conference on Digital Image Computing: Techniques and Applications (DICTA'95), pp. 643-648, Brisbane, Australia, 1995.
11. H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized Detection of Malignant Tumors on Digital Mammograms," *IEEE Transactions on medical imaging*, vol. 18, no. 5, 1999.
12. C. T. Leondes, *Medical Imaging Systems Technology Analysis and Computational Methods*, World Scientific Inc., New Jersey, 2005, pp. 63-85.
13. R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon, "EXPANDER - an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, p. 232, 2005.
14. S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, and E. Zitzler, "BicAT: a biclustering analysis toolbox," *Bioinformatics*, vol. 22, pp. 1282-1283, May 15, 2006 2006.
15. A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, pp. S136-144, July 1, 2002 2002.
16. R. Santamaría, R. Theron, and L. Quintales, "BicOverlapper: A tool for bicluster visualization," *Bioinformatics*, vol. 24, pp. 1212 - 1213, 2008.
17. A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A Systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, pp. 1122 - 1129, 2006.