World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:9, 2015

Calcification Classification in Mammograms Using Decision Trees

S. Usha, S. Arumugam

***Abstract*—**Cancer affects people globally with breast cancer being a leading killer. Breast cancer is due to the uncontrollable multiplication of cells resulting in a tumour or neoplasm. Tumours are called 'benign' when cancerous cells do not ravage other body tissues and 'malignant' if they do so. As mammography is an effective breast cancer detection tool at an early stage which is the most treatable stage it is the primary imaging modality for screening and diagnosis of this cancer type. This paper presents an automatic mammogram classification technique using wavelet and Gabor filter. Correlation feature selection is used to reduce the feature set and selected features are classified using different decision trees.

***Keywords*—**Breast Cancer, Mammogram, Symlet Wavelets, Gabor Filters, Decision Trees

I. Introduction

BREAST cancer affects women of all ages/ethnic groups. In spite of decades old breast cancer research regarding diagnosis and treatment, prevention continues to be the sole way to lower this disease's human toll which currently affects 1 in 8 women in their lifetime [1]. In the United States in 2012, an estimated 227,000 women and 2,200 men are expected to be diagnosed with this cancer, and around 40,000 women are expected to succumb to it. The term "breast cancer" includes more than one disease being an umbrella term for various cancer subtypes of the human breast. Breast cancer subtypes differ in clinical presentations, and show clear cut gene expression patterns in addition to having different genetic/molecular characteristics [2], [3]. Breast cancer subtypes have some shared, also unique causes, and contributing factors influencing prevention approaches.

Mammography is an effective tool for breast cancer detection in its earliest and most treatable stage and thus is the primary imaging modality for cancer screening and diagnosis [4]. This examination also ensures other pathologies detection suggesting cancer nature as being benign, malignant, or normal. The most important improvement is breast imaging which is possible due to advancement in digital mammography.

Mammography is a type of radiography where radiation levels with specific intervals are used to acquire breast images for the diagnosis of any structures which could indicate the presence of any disease including cancer. Early detection is of prime importance in mammary pathologies. Technological advances in image verification have led to the successful increase in breast cancer detection. Mammography plays an important role in detecting lesions in initial stages for a favorable prognosis.

Ductal carcinoma in situ (DCIS) is a non-invasive, precancerous condition where abnormal cells found in the breast duct lining are detected. Though cancer incidence is possible in all ages, it is more in women who have crossed 50 years [5]. Abnormal cells are yet to spread to other tissues outside the duct. However, 70% to 80% of all cases consist of Invasive (infiltrating) ductal carcinoma, the most common cell type. Tumors are characterized by a solid core, usually hard and firm on palpation. Additionally, ductal carcinoma in-situ is also present usually with comedo necrosis occurring in both invasive and intraductal carcinoma areas. Invasive ductal carcinoma spreads to lymph node regions and has the poorest prognosis among the ductal types. The 2nd most common invasive breast cancer type accounting for 8-14% in breast cancers [6] is invasive lobular carcinoma which is recognized by a greater proportion of multi centricity within the same or opposite breast. Lesions have irregular margins with the only evidence being an occasional thickening or induration.

Pattern recognition based mammogram feature extraction has been widely studied. Recently several mammogram analyses schemes using wavelet was introduced. Liu et al., [7] which demonstrated that mammogram multi resolution analysis improved diagnosis when it is based on wavelet coefficients; a statistical features set with binary tree classifier was used in the diagnosis system. A multi resolution mammogram analysis in multilevel decomposition for extraction of the biggest coefficients ratio to be the corresponding mammogram image's feature vector was used by [8]. Daubechies-4, -8 and Daubechies-16 wavelets with four level decompositions were used. Texture is commonly used in analysis and interpretation of images. Mammography employed textures are distinguished on the basis of three extraction procedures: Statistical methods, Model-based and signal processing methods, according to [9].

Segmentation identifies cancer suspicious regions in digital mammograms. Each region has to be classified as benign, malignant or normal after identification. Image classification's main aim is assigning all image pixels to particular classes/themes. This classification is called pattern recognition.

Some breast cancer segmentation and classification methods in digital mammography were reported. Independent component analysis (ICA) and neural network multilayer perceptron was used by [10] to classify mammograms in 3 classes: normal, benign, and malignant, with 98.7% success.

Usha S. is with Park College of Engineering and Technology, Tamilnadu, India (e-mail: usha.s.cse2011@gmail.com).
Arumugam S. is with Nandha Engineeing College, Tamilnadu, India.

Braz et al. [11] classified regions of interest (ROI) in mammogram screening using spatial statistics leading to 98.24% performance to discriminate between Mass and Non-Mass elements. Enhanced multilevel thresholding segmentation and ROI based on rank mammogram segmentation were used by [12]. This method performed better with 80% sensitivity using ICA, feature extraction and K - means cluster to segment digital mammography according to [13].

Ramani and Vanita [21] proposed a computer aided detection of tumors in mammograms. Mini-MIAS dataset mammograms are used to evaluate the presented method. Only a dataset subset is used. Features extraction is through Symlet wavelets and weighted histogram. Extracted features are reduced through Singular Value Decomposition (SVD) with reduced feature set being classified by Naïve Bayes, Random forest and Neural Network algorithms.

In this paper, an automatic mammogram classification is presented. Symlet wavelets are used to transform the images; Gabor filters are applied to extract features. The feature set is reduced using correlation feature selection. The features selected are classified using decision trees.

II. Methodology

U.K.'s National Breast Screening Programme have been digitized to 50 micron pixel edge with a Joyce-Loebl scanning microdensitometer, a linear device in a 0-3.2 optical density range representing pixels with an 8-bit word [14]. The database includes 322 digitized films and radiologist "truth"-markings on locations of abnormalities detected. The database was reduced to 200 micron pixel edge with padding/clipping to ensure that all images are 1024x1024 pixels at 8 bits per pixel. Erosion followed by dilatation has a similar structuring element, completing the opening function.

The MIAS database [14] though no longer supported, is old used much in literature. MIAS annotations are insufficient for some studies as all circumscribed/speculated lesions are to be manually segmented. Another drawback is its digitized resolution which renders it unsuitable for micro-calcification detection experiments. Regarding calcifications, healthier tissue is found in the ground truth region which is justified through calcifications shape as the latter are small lesions spread over a large area with all this being included in the annotation. Cross validation ensures higher formalism in entry data division considered necessary due to limited images with calcifications available in the database. The database Mini MIAS, prevents excessive network training and so a better system generalization. Fig. 1 shows a dataset image.

In this work, Symlet wavelets are used to transform the mammogram images. Symlet wavelet is part of the daubechies wavelet (dbN) family. Symlets are near symmetric and have the least asymmetry. The wavelet transform works by first decomposing an image into constituent parts in the time-frequency domain on a basis function localized in time and frequency domains. The image is decomposed into four frequencies (approximation, horizontal detail, vertical detail and diagonal detail).

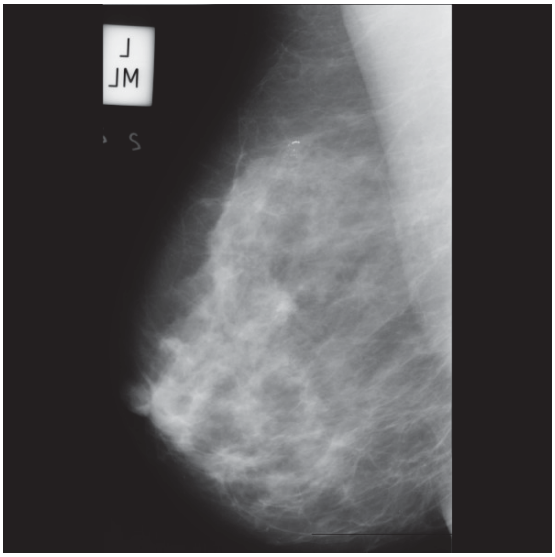

Fig. 1 An image from the mammogram database containing a benign cluster of micro-calcifications

Wavelet transform [15] can be defined as:

$$(W\psi f)(a,b) = \int_{-\infty}^{\infty} f(t)\psi_{a,b}(t)dt \quad (1)$$

where $\psi_{a,b}$ is defined as:

$$\psi_{a,b}(t) = \frac{1}{|a|}\psi\left(\frac{t-b}{a}\right),\ a \neq 0,\ b \in R \quad (2)$$

where $\psi(t)$ is the mother wavelet and $\psi_{a,b}(t)$ are scaled and shifted versions of this wavelet.

As mammography is a proven and reliable breast cancer detection method, hospital produce huge amounts of mammograms and breast screening centers. Content-based image retrieval (CBIR) can retrieve mammograms to help medical procedures. Textural feature extraction is a crucial requirement for development of a content based mammogram retrieval system. But mammograms always contain strong noise being presented in monochrome with low resolutions. This study motivates use of Gabor filters to extract mammographic features due to the following reasons: They provide spatial response profiles similar to a mammalian vision's receptive field [16].

A one-dimensional Gabor filter is a multiplication of a cosine/sine (even/odd) wave with Gaussian windows as,

$$G_E(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-x}{2\sigma^2}} \cos(2\pi w_0 x) \quad (3)$$

$$G_o(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-x}{2\sigma^2}} \sin(2\pi w_0 x) \quad (4)$$

where w_0 defines center frequency (frequency in which the filter yields greatest response) and $\sigma$the spread of the Gaussian window.

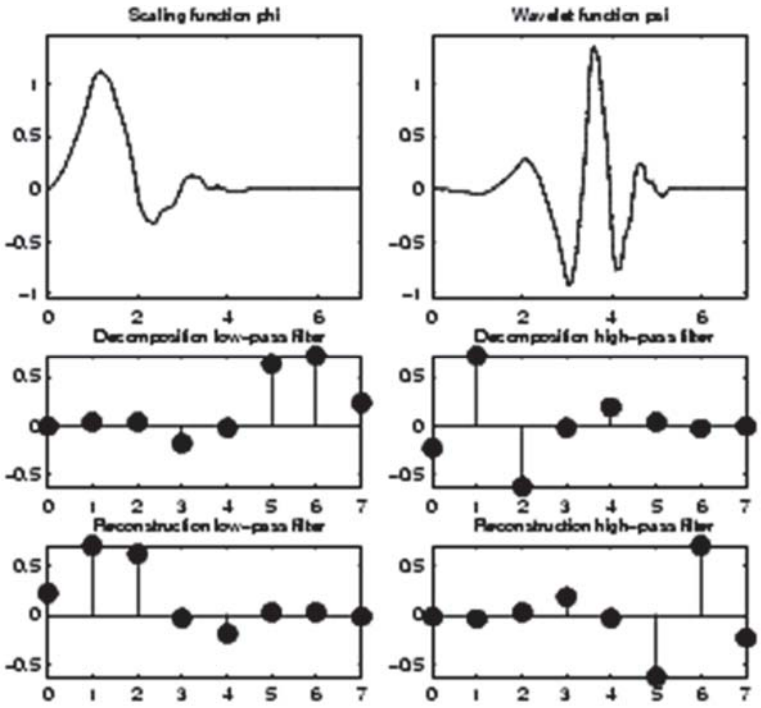

Fig. 2 Symlet wavelet

In space and spatial-frequency [17], Gabor filters exhibit optimal Heisenberg joint resolution. Hence, application of Gabor filters for mammographic textural analysis alleviates low resolution, strong noise, and monochrome problems.

To reduce the feature set obtained, correlation based feature selector (CFS) a filter algorithm is applied. Feature subsets are ranked based on correlation based heuristic evaluation [18] by CFS. It is biased toward subsets with highly correlated features to the class and uncorrelated to others. Irrelevant features due to the low correlation they have with the class are ignored. Redundant features which are highly correlated with one/many features are also screened out. Feature acceptance is based on the extent to which it can predict class where other features have not predicted space. CFS's subset evaluation feature function is given by:

$$M_S = \frac{k\overline{r}_{cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}} \quad (5)$$

where M_S is heuristic merit of a feature subset; S is feature subset; $\overline{r}_{cf}$ is mean feature-class correlation $(f \in S)$; $\overline{r}_{ff}$ is average feature-feature inter correlation;

Following are the classifiers used to classify the mammograms.

A. Decision Stump

Decision trees are widely used machine learning algorithms performing a general to specific feature space search, and adding most informative features as the search proceeds to a tree structure [19]. The aim is selection of a minimal feature set which can efficiently partition feature space into observations classes, and assembles them into a tree. Here, an ambiguous word's manually sense tagged examples are observations in context with partitions corresponding to varied possible senses. A node in the learned decision tree represents each selected feature through a search process. Each node represents a choice point between many different possible values for a feature. Learning continues till all training examples are considered by the decision tree. Thus, such a tree is overly specific to training data and hence will not generalize well to new examples. Therefore, learning precedes a pruning step leading to the elimination of some nodes or even reorganized to ensure a tree that generalizes well to new circumstances. Disambiguated test instances locate a path through a learned decision tree from root to leaf node corresponding to observed features. A decision stump is a one node decision tree [20] whose genesis is due to stoppage of the decision tree learner after addition of a single most informative feature to the tree.

B. J48

A decision tree is a predictive machine-learning model to decide a new sample's target value (dependent variable) dependent on available data's varied attribute values. The internal nodes of a decision tree denote various attributes; inter-nodal branches reveal attribute's possible values in observed samples, while terminal nodes provide information of the dependent variable's final value [21].

Dependent variable is the to-be predicted attribute as its value depends upon/is decided by other attributes values. The latter which aid prediction of dependent variable's value are known as independent variables in datasets.

C. Classification and Regression Tree (CART)

Classification and Regression Tree (CART) performance is based on the target variable type which could be continuous or categorical. For categorical class label, values of predictor variables are used to span through the tree to encounter a leaf node. The value of the leaf node will be the class label assigned. Gini index decides attribute split criteria.

III. Results and Discussion

Mammograms from the Mini-MIAS dataset are used for evaluation of the presented method. Only a subset of the dataset is used. Features are extracted using Symlet wavelets and Gabor filter. The extracted features are then reduced using CFS and the reduced feature set is classified decision stump, J48 and CART. Training set consisted of 60% data and the remaining was used as test set. Experiments were conducted with complete set of features extracted and then with CFS reduced feature set. The results obtained and the classification accuracy is tabulated in Table I and the precision recall obtained is given in Table II. Figs. 3 and 4 show the same respectively.

TABLE I
Experimental Results for Various Techniques

Techniques	Classification accuracy %	RMSE
Decision stump	80.00%	0.3859
J48	70.00%	0.5404
CART	60.00%	0.5178
Decision stump with CFS	80.00%	0.3859
J48 with CFS	80.00%	0.4406
CART with CFS	70.00%	0.4905

Fig. 3 is the graphical representation of the classification accuracy and RMSE. It is observed from Table I that the best accuracy is achieved by decision stump.

It is observed from the graphs that the precision and recall is better for decision stump with or without the feature reduction.

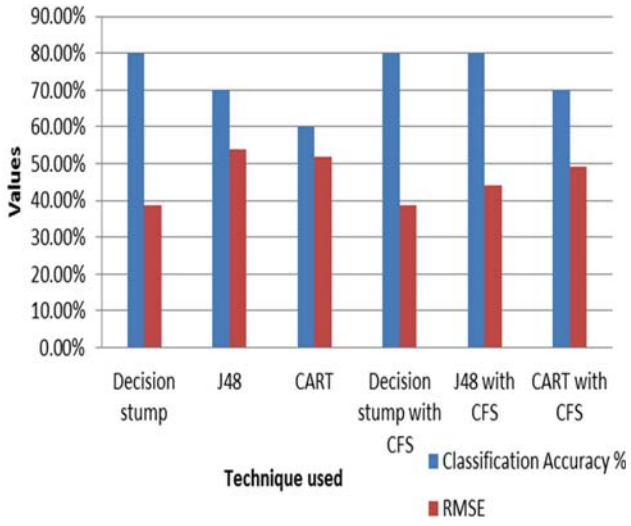

Fig. 3 Classification Accuracy and Root Mean Squared Error

TABLE II
Precision, Recall

Techniques	Precision	Recall
Decision stump	0.857	0.8
J48	0.702	0.7
CART	0.6	0.6
Decision stump with CFS	0.857	0.8
J48 with CFS	0.8	0.8
CART with CFS	0.72	0.7

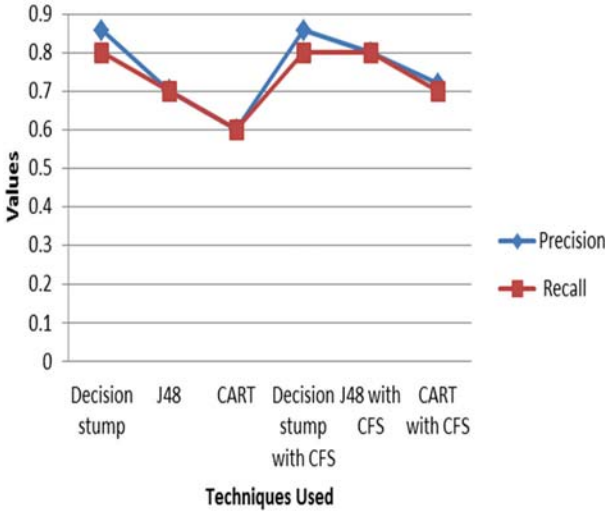

Fig. 4 Precision, Recall

Ganesan et al. [22] proposed an automated diagnosis of mammogram images of breast cancer using Discrete Wavelet Transform and Spherical Wavelet Transform Features. Classification accuracy is achieved 75.67%, 59.41%, 81.73 and 54.05% for QDC, NMC, SVM and ParazenC respectively for DWT and SWT using ten-fold cross validation. Our results is comparable the work in the literature which achieves 80% accuracy. Future work can explore optimizing the classifiers for improving the accuracy.

IV. Conclusion

In this paper, an automatic mammogram classification is presented. Mammograms from the Mini-MIAS dataset are used for evaluation of the presented method. Symlet wavelets are used to transform the images; Gabor filters are applied to extract features. The feature set is reduced using correlation feature selection. The features selected are classified using decision trees. Experimental results indicate that the classification accuracy realized by decision trees is fairly good. Decision stump has better classification accuracy and better precision and recall. Further investigations are required to improve the classification accuracy.

References

[1] Majumder, R., Chaudhuri, B., Ghosh, A., Ledwich, G., & Zare, F. Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, et al. SEER Cancer Statistics Review, 1975-2009. National Cancer Institute. http://seer.cancer.gov/csr/1975_2009_pops09

[2] Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., & Aparicio, S. (2012). The genomic and transcriptomic

architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403), 346-352.
[3] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., & Botstein, D. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747-752.
[4] Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I., & Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. Information Technology in Biomedicine, IEEE Transactions on, 13(2), 236-251.
[5] Virnig, B. A., Tuttle, T. M., Shamliyan, T., & Kane, R. L. (2010). Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. Journal of the National Cancer Institute, 102(3), 170-178.
[6] Wasif, N., Maggard, M. A., Ko, C. Y., & Giuliano, A. E. (2010). Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. Annals of surgical oncology, 17(7), 1862-1869.
[7] S. Liu, C. F. Babbs, E. Delp, "Multiresolution detection of spiculated lesions in digital mammograms", IEEE Transactions on Image Processing 10 (6), 2001,pp. 874–884.
[8] E. A. Rashed, I. A. Ismail, S. I. Zaki., "Multiresolution mammogram analysis in multilevel decomposition". Pattern Recognition Letters 28, 2007, pp.286–292.
[9] A. Oliver, Automatic mass segmentation in mammographic images. PhD thesis, Universitat de Girona (2004)
[10] Campos, L. F. A. ; Barros, A. K. ; Silva, A. C. "Independent Component Analysis and Neural Networks Applied for Classification of Malignant, Benign and Normal Tissue in Digital Mammography", In: Special Issue - Methods of Information in Medicine, v. 46, p. 212-215, 2007.
[11] Braz , G. Jr., E. C. Silva, A. C. Paiva and A. C. Silva, "Breast Tissues Classification Based on the Application of Geostatistical Features and Wavelet Transform", In: International Special Topic Conference on Information Technology Applications in Biomedicine, ITAB 2007, 6th, 227-230, 2007.
[12] Domínguez, A. Rojas. Nandi, A. K. "Detection of masses in mammograms using enhanced multilevel thresholding segmentation and region selection based on rank", In: Proceedings of the fifth conference on Proceeding of the Fifth IASTED International Conference: biomedical engineering. 2007
[13] Campos, L. F. A.; Costa, D. D.; Barros, A. K. "Segmentation on Breast Cancer Using Texture Features and Independent Component Analysis", In: Bioinspired Cognitive Systems, BICS 2008.
[14] Suckling, J. et al. (1994). The mammographic image analysis society digital mammogram database, International Congress Series 1069 pp. 375–378.
[15] Ramanathan, R., Kalaiarasi, K., & Prabha, D. (2013). Improved wavelet based compression with adaptive lifting scheme using Artificial Bee Colony algorithm. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2(4), pp-1549.
[16] J.G. Daugman, "Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression," IEEE Transactions on Acoustics, Speech, Signal Processing, vol. 36, no. 7, pp. 1169-1179, 1988
[17] Wei, C. H., Li, Y., & Li, C. T. (2007, July). Effective extraction of Gabor features for adaptive mammogram retrieval. In Multimedia and Expo, 2007 IEEE International Conference on (pp. 1503-1506). IEEE.
[18] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.
[19] Pedersen, T. (2001, June). A decision tree of bigrams is an accurate predictor of word sense. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-8). Association for Computational Linguistics.
[20] R. Holte. 1993. Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11:63{91.
[21] Zhang, Y., Zhao, Y., A Comparison of BBN, ADTree and MLP in separating Quasars from Large Survey Catalogues, ChJAA 7, 289-296, 2007.
[22] Ganesan, K., Acharya, U. R., Chua, C. K., Min, L. C., & Abraham, T. K. (2014). Automated diagnosis of mammogram images of breast cancer using discrete wavelet transform and spherical wavelet transform features: A comparative Study. Technology in cancer research & treatment, 13(6), 605-615.

Usha S is with Park College of Engineering and Technology. She is currently pursuing her doctorate in India.

Arumugam S is with Park College of Engineering and Technology, India.