Original Research Paper

Volume-9 | Issue-12 | December - 2019 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Engineering

# RESEARCH ON DIFFERENT FEATURE EXTRACTION AND MAMMOGRAM CLASSIFICATION TECHNIQUES

**Anita Kaklotar** Computer Department, Rajiv Gandhi Institute of Technology, Mumbai, India

**ABSTRACT** Breast cancer is the primary and the most common disease found among women. Today, mammography is the most powerful screening technique used for early detection of cancer which increases the chance of successful treatment. In order to correctly detect the mammogram images as being cancerous or malignant, there is a need of a classifier. With this objective, an attempt is made to analyze different feature extraction techniques and classifiers. In the proposed system we first do the preprocessing of the mammogram images, where the unwanted noise and disturbances in the mammograms are removed. Features are then extracted from the mammogram images using Gray Level Co-Occurrences Matrix (GLCM) and Scale Invariant Feature Transform (SIFT). Finally, the features are classified using classifiers like HiCARe (Classifier based on High Confidence Association Rule Agreements), Support Vector Machine (SVM), Naïve Bayes classifier and K-NN Classifier. Further we test the images and classify them as benign or malignant class.

**KEYWORDS :** HiCARe, GLCM, SIFT, Naïve Bayes, K-NN Classifier and Support Vector Machine.

## I.INTRODUCTION

Breast Cancer is the most invasive cancer [1] which is found in females all throughout the world. It comprises of 16% of all the female cancers and accounts for 22.9% of invasive cancer. 18.2% of all cancer deaths are from breast cancer which includes both males and females. Normal cells die as they grow old or get damaged, thereby giving place to new cells. Sometimes, however, this process goes wrong. Body develops new cells when it doesn't require them, and old or damaged cells fail to die. These extra cells often forms a lump (a mass of tissue) or tumor. Cancer that forms in the breast tissues, usually in the tubes that carry milk to the nipple i.e. ducts and in the glands that make milk i.e. lobules, is the breast cancer.

Mammography is used as a diagnostic and screening tool. It is the process of using low-energy X-rays to examine the human breast. Mammography is used for the early detection of breast cancer, usually through detection of characteristic masses or micro calcifications. We have used MIAS (http://peipa.essex.ac.uk/ipa/pix/mias/) database for identifying the benign and malignant tumors. Preprocessing is used to improve the image quality of images by removing external noise or disturbances. After preprocessing the features are detected and extracted. Several techniques are used to analyze [2] detect or to extract features from mammogram images. GLCM is used to extract relevant features. Once the features are extracted, the images are classified into benign or malignant tumor. The severity of lymph's present in the breast [3] are identified by different classification algorithms. Fig. 1 shows the basic process followed for classification of the mammogram images as benign or malignant.

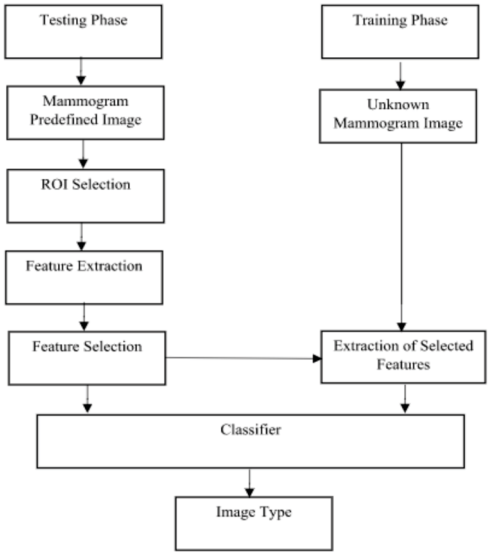


**Fig -1: Process for Classification of Mammogram images**

## II. PREPROCESSING OF IMAGE

Data obtained from the mammographic images is often, noisy, incomplete and inconsistent, hence, pre-processing becomes a must. Preprocessing phase of the images plays an important role in improving the quality of the images and making the feature extraction phase more reliable. Images are usually scanned at different illumination conditions, so there occurs variation in the brightness of images. The first step toward noise removal was pruning the unwanted portions of the image. Thus, we have eliminated almost all the disturbances and most of the noise. The noise from the mammogram images was removed by Median Filter which has the ability to remove noise without reducing the image sharpness. When using median the output pixel's value is determined by the median of the neighborhood pixels. The contrast of the mammogram images was adjusted by histogram equalization [4] which improves the contrast by spreading out the most frequent intensity values.

## III. FEATURE EXTRACTION

The features from ROI are detected and extracted using various feature extraction techniques like GLCM (Gray Level Co-occurrence Matrix) [5] and SIFT (Scale Invariant Feature Transform) for the classification of the mass as benign or malignant. The feature space becomes very large and complex because of wide diversity of the normal tissues and their abnormalities. The performance of any classifier is mainly determined by how well the feature is extracted and selected. Selection of optimum feature set results in higher accuracy of the classifier. Feature extraction techniques analyze objects and images to extract the most prominent features that represents various classes of objects. Features are used as inputs to classifiers that assign them to the class that they belong.

### *A. Gray Level Co-Occurrences Matrix (GLCM)*

The Gray Level Co-Occurrences Matrix (GLCM) size is determined by the number of gray levels in the image. Gray co-matrix uses scaling by default to reduce the number of intensity values in an image to eight, but NumLevels and the Gray Limits parameters can also be used to control the scaling of gray levels. The gray-level co-occurrence matrix has the capability of revealing certain properties about the spatial distribution of the gray levels in the images with texture. For example, if there is a case where most of the entries in the GLCM have more concentration along the diagonal, then the texture is coarse with respect to the specified offset. Several statistical measures can also be derived from the Gray Level Co-Occurrence Matrix. The calculation of first three values in a GLCM is shown in Fig 2. In the Gray Level Co-Occurrences Matrix, the occurrence or instance of each element in the input image having two horizontally adjacent pixels as the value of the elements are displayed. For example, there is only one instance in the input image where two horizontally adjacent pixels have one value as 1 and the other also as 1, so the element (1, 1) contains the value 1. Similarly, there are two instances where two horizontally adjacent pixels have one value as 1 and the other as 2, so the element (1, 2) contains the value 2. There are no instances of two horizontally adjacent pixels with the values 1 and 3, so the element (1, 3) in the GLCM has the value 0. Graycomatrix keeps processing the input image. Other pixel pairs (*i* ,*j*) are also scanned and their sums are recorded in the corresponding elements of the GLCM.

Submitted : 19th July, 2019 | Revised : 27th September, 2019 | Accepted : 02nd October, 2019 | Publication : 01st December, 2019

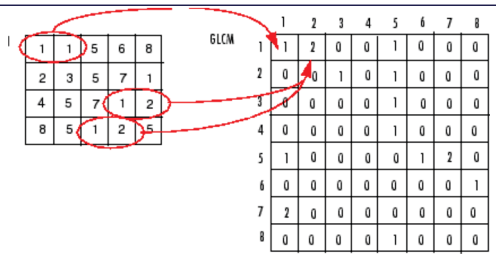


**Fig -2: Process for Creating the GLCM**

### *B.Scale Invariant Feature Transform (SIFT)*

The SIFT approach was proposed by Lowe [6]. It transforms an image into a collection of local feature vectors. The features are scaling, translation and rotation invariant and has been proved to be a robust keypoint descriptor in different image classification, retrieval and matching applications [7]. SIFT provides keypoint detection through the identification of interesting points in the scale space. The images are first convolved with Gaussians at different scales and the Differences of Gaussians (DoG) of the images are generated from the adjacent smoothed images. The detection phase for SIFT features include scale space extrema detection. The points of interest are identified as local extrema of differences of Gaussians i.e. DoG. By the interpolation of nearby data, keypoint localization is done. Orientation assignment involves computing the gradient orientation histogram in the neighborhood of the keypoint. Finally, the keypoint descriptors are generated. Selecting a keypoint orientation, the feature descriptor is computed as a set of orientation histograms on 4 x 4 pixel neighborhoods. The SIFT features were first used for breast density classification by Bosch et al. [8]. The descriptors for SIFT are computed on a regular grid with spacing M pixels over N regular support patches, as in [8]. The features are first extracted from a set of training images and stored in a database. The test image is matched by comparing individually each feature from the test image to the training set database. By searching for stable features across all possible scales using function of scale which is continuous[9]; locations which are scale invariant of the image can be detected.

## IV. CLASSIFICATION

Classification is nothing but identifying to which of a set of class, a new observation belongs, on the basis of a training set of data containing observations (or instances) whose class membership is known. Classifier is used for this purpose. A classifier will train the training set of the mammogram images and further test the unknown samples of mammogram images against the known images. Testing will classify the mammogram images as malignant or benign depending on the features extracted for the trained images.

### *A. HiCARe*

HiCARe (Classifier based on High Confidence Association Rule Agreements) [10] is a new special classifier able to return multiple classes (keywords) when processing a test image. A match occurs when the image features satisfy the body part of a rule. The HiCARe algorithm stores all itemsets (set of keywords) belonging to the head of the rules in a data structure. An itemset *h* is returned in the suggested diagnosis if the condition stated below is satisfied

$$\frac{n\,M(h)}{n\,M(h)+n\,N(h)} \geq \beta \qquad (1)$$

where n M (h) is the number of matches of the itemset h and nN(h) is the number of not-matches. A threshold β (0<=β<=1) is employed to limit the minimal number of matches required to return an itemset in the suggested diagnosis.

### *A. Support Vector Machine (SVM)*

Support Vector Machine (SVM) is based on the concept of decision planes or hyperplanes which is instrumental in defining decision boundaries [11]. The decision plane separates the feature values into benign tumor or malignant tumor. The main task of the hyperplane is to maximize the margin in the training data to classify binary classes, here the classes being malignant tumor and benign tumor. Margin is nothing but the distance between the support vectors and the class boundary. The decision planes defines the decision boundaries. A classification task consists of some data instances for training and testing. Each instance in the training set contains one "target value" i.e. class labels and several "attributes" i.e. features. Here the class labels being 'benign tumor' and 'malignant tumor' and attributes being the feature of every mammographic image. The performance of SVM largely depends on the kernel.

The following stages are followed to measure the accuracy of SVM:

- Preprocessing of the mammographic images in the MIAS database.
- Separation of the database in training and testing sets according to 10-fold cross validation technique.
- Representing the input data as a table which consists of labels of each class and the features of that class.
- Choice of the way of training by selecting:
- Selection of method of training
- Value of the penalty term C (This is the SVM complexity constant which sets the tolerance for misclassification).
- Choice of the kernel.
- Training the features of the mammographic images of the training set.
- Testing the test sets and evaluating the performance of SVM.

### *B. Naive Bayes*

Naive Bayes is used for constructing classifiers or models that assign class labels to instances of problem, represented as feature vector values, where the class labels are drawn from some finite dataset. Given the class variable, it assumes that a particular feature has the value independent of any other feature's value [12].

Naive Bayes Classifier is used for classification, in which some samples of mammographic images are used for training, and the remaining samples are used for testing. The result is displayed in a confusion matrix which describes actual and predicted classes of the mammographic images.

True Positive (TP) – number of all mammographic images which are correctly classified as being malignant.

False Positive (FP) – number of all mammographic images which are incorrectly classified as being malignant, while they are benign.

True Negative (TN) – number of all mammographic images which are correctly classified as being benign.

False Negative (FN) – number of all mammographic images which are incorrectly classified as being benign, while they are malignant The performance evaluation is done using different measures like accuracy, precision, specificity and sensitivity. These measures are calculated from confusion matrix using the following equations. The values of these equations range between -1 (inverse prediction) and +1 (perfect prediction).

Accuracy = (TN+TP)/ (FP+FN+TP+TN) (2)
Precision = TP/ (FP+TP) (3)
Specificity = TN/ (TN+FP) (4)
Sensitivity = TP/ (TN+FP) (5)

### *A.K-NN Classifier*

The k-nearest neighbor algorithm (K-NN) is used for classifying objects depending on closest training images in the feature space of mammogram images [13]. K-NN is an instance-based learning. The function in K-NN is only approximated locally. First the classification of the images is done, only then the further computation is done. The classification of any object is determined by a majority vote of its neighbors. Whichever class is the most common amongst its k nearest neighbors, the object is assigned to that class. k is a positive integer which has typically a small value). If the value of k is 1, then the object is assigned to the class of its nearest neighbor. The neighbors are selected from a set of objects for which the correct classification is already known. This can be considered as the training set for the algorithm, training step is not explicitly required.

**Table- I: Comparative Analysis of various classification techniques**

| Feature Extraction Technique and Classifier used. | Results |
|---|---|
| GLCM +HiCARe | For MIAS dataset the average accuracy computed was 92 % for 85 images in training database and 25 images in the testing database |

| | |
|---|---|
| GLCM + Support Vector Machine(SVM) | Average accuracy of about 93% for MIAS database. |
| SIFT + Support Vector Machine(SVM) | More than 80% accuracy for a set of 104 images from MIAS databases |
| GLCM + Naive Bayes | 80.11% classification accuracy for images from the MIAS database. |
| K-nearest neighbor algorithm (K-NN) | Average accuracy of 80.09 % for MIAS database. |

## II. RESULT AND DISCUSSION

Table – I shows the result of various feature extraction techniques and classifiers when used on MIAS dataset. GLCM along with Support Vector Machine (SVM) shows 93% accuracy for MIAS dataset. GLCM when used with

Naive Bayes gives 80.11% classification accuracy for images from the MIAS database. SIFT and Support Vector Machine (SVM) show more than 80% accuracy. Similar results are achieved when K-nearest neighbor algorithm is used.

## III. CONCLUSION

This Paper presents a brief knowledge about classification of mammogram images using various feature extraction and classification techniques. We have studied and analyzed feature extraction techniques like GLCM and SIFT and classifiers like HiCARe, SVM, Naïve Bayes and K-NN classifier. The images are first trained by the classifier and then tested to classify the test images as benign or malignant cancer. The results show that the accuracy is the highest when feature is extracted using Gray Level Co-Occurrences Matrix and Support Vector Machine classifier as compared to other feature extraction techniques and classifiers for the MIAS database.

## ACKNOWLEDGMENT

We would like to thank our guide, Prof. Jyoti Deskmukh,
Department of Computer Engineering, for the valuable inputs and advice and making our project achievable by constant support and encouragement.

## REFERENCES

[1] "World Cancer Report". International Agency for Research on Cancer. 2008. Retrieved 2011-02-26.(cancer statistics often exclude non-melanoma skin cancers such as basal cell carcinoma which though very common are rarely fatal)
[2] S.M.Salve, V.A.Chakkarwar et .al "Classification of Mammographic images using Gabor Wavelet and Discrete Wavelet Transform" International Journal Of advanced research in ECE ISSN:2278-909X,Vol. 2 pp.573-578, May 2013.
[3] A. Bosch, X. Munoz, A. Oliver, and J. Marti, "Modeling and classifying breast tissue density in mammograms," in Computer Vision and Pattern Recognition, 2006.
[4] ] Dougherty J, Kohavi R, Sahami M. "Supervised and unsupervised discretization of continuous features". In: Proceedings of the 12th international conference on machine learning.San Francisco:Morgan Kaufmann; pp 194–202, 1995.
[5] Aswini kumar mohanty, Sukanta kumar swain ,Pratap kumar champati ,Saroj kumar lenka , "Image Mining for Mammogram Classification by Association Rule Using Statistical and GLCM features" , IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, 2011.
[6] D. Lowe, "Object recognition from local scale-invariant features," in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, 1999, pp. 1150 –1157
[7] J. Caicedo, A. Cruz, and F. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in Artificial Intelligence in Medicine, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, vol. 5651, pp. 126–135.
[8] A. Bosch, X. Munoz, A. Oliver, and J. Marti, "Modeling and classifying breast tissue density in mammograms," in Computer Vision and Pattern Recognition, 2006.
[9] Witkin, A.P."Scale-space filtering". In Proceeding of International Joint Conference on Artificial Intelligence, Karlsruhe, Germany, 1983, pp. 1019-1022.
[10] Deshpande, Deepa S., Archana M. Rajurkar, and Ramchandra R. Manthalkar. "Texture Based Associative Classifier—An Application of Data Mining for Mammogram Classification." Computational Intelligence in Data Mining-Volume 1. Springer India, 2015. 387-400.
[11] Fatima Eddaoudi, Fakhita Regragui, Abdelhak Mahmoudi, Najib Lamouri, 2011 , "Masses Detection Using SVM Classifier Based on Textures Analysis",Applied Mathematical Sciences, Vol. 5, 2011, no. 8, 367 – 379.
[12] S.Krishnaveni1, R.Bhanumathi2, T.Pugazharasan3, "Study of Mammogram Micro-calcification to Aid Tumor Detection Using Naive Bayes Classifier ", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 3, March 2014
[13] S. Mohan Kumar, G. Balakrishnan, "Classification of Microcalcification in Digital Mammogram using Stochastic Neighbor Embedding and KNN Classifier", International Conference on Emerging Technology Trends on Advanced Engineering Research (ICETT'12).