Breast density estimation in DBT

Anirudh Puligandla, Di Meng and Dousai Nayee Muddin Khan

***Abstract*— The aim of our work is to find the possibilities of detecting the breast cancer in women based on various Computer Aided Diagnostics (CAD) systems. This project describes about finding the tissue densities for breast based on Breast Image Reporting and Data System (BIRADS). We are using Local Binary Patterns (LBP) for getting the features of the given dataset. As we have the feature extraction for the given data we will need a classifier which is used for data classification. We are using a basic and simple classifier called KNNC which is based on finding the nearest neighbourhood. By the following features and classifier described we have evaluated the given data and the results for finding the breast cancer are mentioned in the result section.**

I. INTRODUCTION

From recent research and studies, breast cancer is considered as most common cancer for women around the world of about 10-15% considering the second most common cancer around the world. The frequent technique used in medical to find the symptoms of breast cancer are by using mammography. This is basically the X-ray of the breast to find the changes in breast tissue and the density of the breast tissue. A woman's breast is made up of composition of fat tissues, which describes the breast density. There are higher and lower breast density, higher is described as greater amount of breast compared to tissue while lower is described as greater amount of fat compared to tissue. Density of the breast is based on many factors as Age, Weight, Vitamin and calcium intake.

In our work we are using Computer Aided Diagnostics (CAD) systems to known the densities and to evaluate the breast densities. The main goal of this work is to know the breast density from the given dataset of DICOM images by using CAD systems. For our project we have given data for 16 patients with ground truth referred as BIRADS (Breast Image Reporting and Data System) in four different classes. All the given data for 16 patients is divided into four classes based on density of the breast tissues. The given dataset is divided as BIRAD 1- 4 with respective to densities as BIRAD-1 is for below 25%, BIRAD-2 is for 25-50%, BIRAD-3 is for 50-75% and BIRAD-4 is for the images with more than 75%. For training the datasets we are using 15 patients data and for testing we have used one. In the below sections we will discuss about the feature extraction, classification, results and about the challenging day results.

This project was commenced and accomplished under the guidance of Dr. Robert Marti Marly, Department of Computer Architecture and Technology, University of Girona, Spain,robert.marti@udg.edu

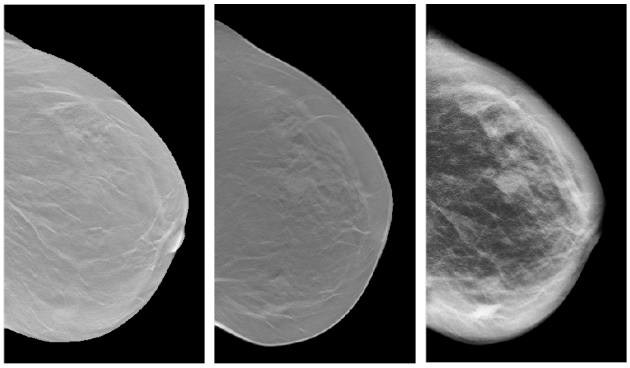
Anirudh, Meng, Nayeem are students of VIBOT and MSCV studying at University of Girona, Spain, pv.anirudh@gmail.com, nayeem.khan@stu.upes.ac.in, ctmengdi@gmail.com

All the code evaluation and plotted images are evaluated by using MATLAB.

II. METHODOLOGY

All the mammograms provided are removed background, labels and pectoral muscle areas. Each Digital breast tomosynthesis(DBT) is three dimensional and each slice with different depth has different information. The basic idea to classify the breast tissue density is to extract features from the DBT images and train the classifier with limited dataset. Testing each given image and comparing the obtained result with ground truth to evaluate the accuracy of our method.

A. Pre-processing



(a) (b) (c)

Fig. 1: (a): DBT slice in depth 10. (b): DBT slice in depth 30. (c): The projection of all the slices after preprocessing.

The DBT is an imaging technique that allows a volumetric reconstruction of the whole breast from a finite number of low-dose two-dimensional projections. Each slice from the DBT has different illumination and contrast. The(a)and (b)from Figure 1 are the slices from the DBT of one patient, we can see that they have different contrast and the dense intensity tissues are lying on different positions. The details are not shown clearly so that we can hardly differentiate the dense tissues which have high risk of cancer from fatty tissues. Thus, it is necessary to apply pre-processing to the original images before extracting features.

The skin of breast which holds high intensity values is useless and affecting the density result. It is removed by extracting the contour of the breast firstly. Usually the skin has a width, in the project we specify the skin width as 10 pixels. Along every pixel on the contour, we take 10 more pixels on the left side of the contour which represent the

enlarged breast skin. The skin can be removed by assigning these pixels with 0 value which are black as background.

After removing the skin, each slice is enhanced by histogram equalization. Considering taking all the information into account and saving computation time, all the slices are added up so that we get 2D projections from 3D reconstructed images. It is noted that the projection ((c) from Figure 1) without skin has clear dense and sparse areas.

B. Feature Extraction

Image Features were considered the key feature for the classification of the provided dataset. As the dense tissue can be clearly seen with higher intensities, texture of each slice can be considered as a unique descriptor for these volumes. It is to be noted that segmentation can also distinguish the dense tissue clearly, but, the computational cost would be much higher for larger databases. Here we propose the use of Local Binary Patterns (LBP) to extract feature descriptors from the volumetric images.

Initially, GLCM, was considered for feature extraction, but, later it was observed that the variation in the co-occurence matrices was random for different images. LBP, on the other hand, are fast to compute and provide feature vectors that are distinguishable for each class of images. LBP works on the principle of generating a regional histogram of binary patterns, based on threshold, for each pixel while considering a neighbourhood window as shown in Figure 3. Figure 2 shows the concatenation of regional histograms. The image is first, divided into few regions. The LBP window is then, applied on each pixel of the image to generate a binary pattern for the pixel, based on the neighbourhood. The binary patterns in each region are plotted as a histogram and, lastly, the histograms of all the regions, thus, computed, are concatenated to provide a single feature descriptor for the image.

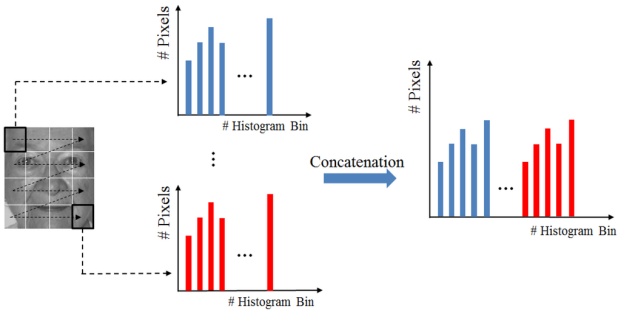


Fig. 2: Regions and histogram concatenation

This process was repeated to generate the features for each the slices, and also, for 2D projections of the whole volume. It was noted that 2D projections have all the key information from all the slices and can increase the computational cost, while providing similar results. But 2D projections may cause overlap in the tissue, sometimes, leading to some mismatch. Although, LBP are fast to compute, they can be costly if all the slices are considered for large databases. There are many advancements to the basic LBP, that are mentioned in various publications. There exists also a MATLAB function that computes LBP with various parameters that can be configured based on the situation and requirement. This specific function was used here, and was tested with different parameters such as, increasing window size, normalization of the histograms. It may be noted here, that the default configuration of the function can accomplish the task in itself.

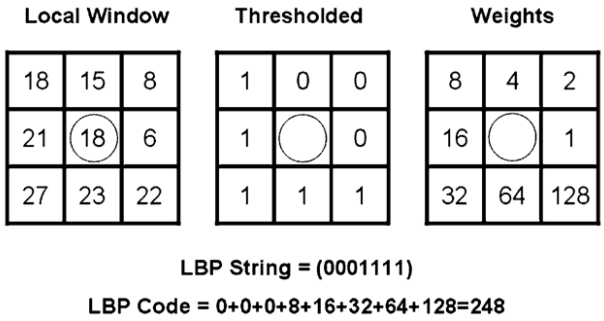
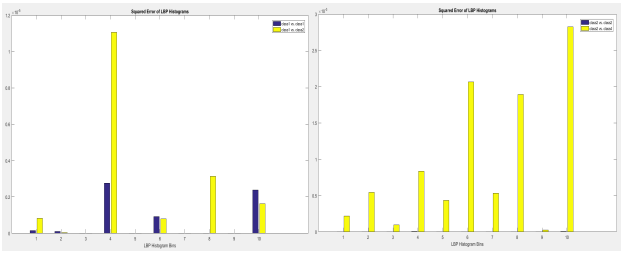


Fig. 3: LBP window

After computing the features, they were verified for uniqueness for different classes by using the squared error metric. It was affirmed that the features for different classes actually have quite distinctive features. Figure 4 shows a couple of examples for the squared error between the features.



(a) Comparison between classes 1 and 2 (b) comparison between classes 2 and 4

Fig. 4: Squared error between Feature sets

C. Classification

Choosing the right classifier from a big pool of available classification technique is a challenging task. One classifier might work well for one database while, it may fail for other databases. Multi-class Support Vector Machine (SVM) and K-Nearest Neighbouring Clustering (Knnc) and Bayes Classifier(BC) were tried with the computed feature vectors. But, Only Knnc provided consistent results for differently computed features and diferent train and test combinations. Since, knnc works on the basis of allocating the feature points into different regions or clusters based on thier distance from the center of the cluster, it is well suitable for feature descriptors like LBP. Since, the feature vectors describe very few parameters, and also considering the fact

that the classification was to be done under only 4 classes, Knnc was preferred over the others.

Knnc is basically a supervised classification scheme. A subset of the entire data set (called the training set), for which the user specifies class assignments, is used as input to classify the remaining members of the data set. The user specifies the number of expected classes, and the training set should contain examples of each class. Figure 5 shows the basic scheme of a 4-class knn classifier. In the figure, X is the new feature that is to be assigned into any of the 4 existing clusters. The default configuration of the MATLAB function for knnc was used to classify the data.

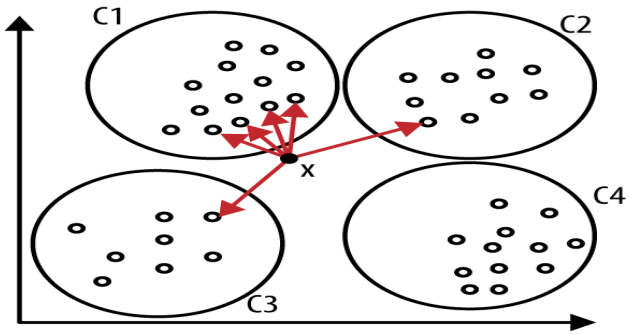


Fig. 5: Working principle of Knnc

Initially, during the testing phase, one image from each class was used for testing against 3 images from each class for training. The labels assigned to the new testing dataset was cross-verified with the given ground truth. Also, one testing image from the 16 volumes database was tested against the rest 15 volumes for training. The latter method provided quite accurate results with only one image classified incorrectly. Initially, the results were not up to the mark as the database provided was very small. But having the benefit of testing over a larger database during the challenge day helped in obtaining better results. The results will be discussed in detail, under the results section.

Although, the classifier was compared with other classifiers such as SVM and BC, the results and comparison will not be covered under this report as they were not consistent over the databse. It may be noted that SVM can be a better classifier for the case of multi-class classification that involves classification into many number of classes. Knnc may fail for larger number of classes. Lastly, selection of a classifier depends wholely, on the problem and there cannot be one global classification technique for any problem.

III. Results

Based on the given 16 images with labels, the training dataset and testing dataset are built. For providing a better classifier with limited dataset, each time 15 images are taken for training and the rest 1 image for testing. Thus, all the 16 images can be classified into a class and be compared with ground truth. The confusion matrix is shown in TABLE II .

TABLE I: Confusion matrix

	B-I	**B-II**	**B-III**	**B-IV**
B-I	3	0	0	0
B-II	1	4	0	0
B-III	0	0	4	0
B-IV	0	0	0	4

From the confusion matrix, we get the accuracy of 93.75%. It is noted that 15 images are perfectly classified to their classes, only one image which is belonging to class B-I is classified into class B-II. From good view, the class B-I and B-II has similar tissue density so that the model is performing properly. The following table shows the comparison metrics, namely, True Positives (TP), False Positives (FP), False negatives (FN), True Negatives (TN), specificity, precision and sensitivity.

TABLE II: Comparison Metrics

	B-I	**B-II**	**B-III**	**B-IV**
TP	3.0	4.0	4.0	4.0
FP	0.0	1.0	0.0	0.0
FN	1.0	0.0	0.0	0.0
TN	12.0	11.0	12.0	12.0
Precision	1.0	0.8	1.0	1.0
Sensitivity	0.75	1.0	1.0	1.0
Specificity	1.0	0.92	1.0	1.0

IV. Discussion

The breast density can be estimated from many aspects such as texture features and intensity ratios. There are many available methods to extract feature vectors for the classification. But which model is the most efficient and accurate for this specific case? That mainly depends on the amount of data. With limited data, the classifier can be either over-learned or under-learned. The classification problem which is actually the machine learning subject can be better solved by big data.

V. Conclusion

A novel breast tissue density classification methodology is proposed which extracts local binary patterns as descriptors and uses k-nearest neighbors method for classifying. With the given 16 datasets, each image has been tested while it is not been training. The result gives 93.75% accuracy and 97.094247 seconds computation time. Nevertheless, the method can be improved by providing larger dataset.