A survey on learning and classification approach for the detection of masses and non-masses based on digital mammograms

*** Rajinder kumar[1], **Sumit Chopra[2]**

[1] M-Tech Scholar / Student in CSE, KCCEIT, Nawanshar.

[2] HOD (CSE), K.CCEIT, Nawanshar.

Abstract

Breast cancer is second most dangerous disease in the world after the lung cancer among women. Because of this reason, breast cancer detection is most focused area by many researchers. .Most of the cancercase symptoms are identified at the late stage, when the tumor becomes bigger in size and treatment becomes invasive case. The reduces help in less number of modalities for the treatment if Early detection of the cancer before the development of the symptoms may. The common Screening is the basic procedure for identification of breast cancer at an earliest stage. The mammography is an efficient screening method, in which abnormalities can be detected. It is difficult to identify the tumor in the breast tissue because tumors possess equal intensity in the breast tissue and appears poor in contrast. Then the computer aided detection helps for physicians and radiologist to find abnormality at an earliest in the absence of any symptoms. The proposed segmentation algorithm detects clearly defined region of mass using suitable segmentation technique. The efficiency of the algorithm is measured with many images of Mini-MIAS database. Mammography is a method used for the detection breast cancer.

Many researchers worked in the breast cancer detection using their proposed segmentation methods used in it. So they have no solution given by researchers is best promising. It is a challenging problem to solve for researchers. This study describes the recent advances in image processing and machine learning techniques for breast cancer detection. The study shows that Local Binary Pattern method used for feature extraction and Support vector machine for classification as foremost technique used for breast cancer detection. The comparative study of literature work summarizes the effectiveness of different approach used by researchers for breast cancer detection.It is a challenging problem to solve for researchers.

Keywords: - Breast cancer, Mammography, CAD, image segmentation, feature extraction

.Introduction:-

Cancer begins when healthy cells in the breast change and grow out of control, then creating forming a mass or of cells called a tumor. A tumor can be cancerous or benign. A cancerous tumor is Malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but will not spread. Breast cancer spreads when breast cancer cells move to other parts of the body through the blood vessels and/or lymph vessels. This is called metastasis.

Breast cancer is the most common cause of the death in women, according to a survey conducted by WHO the most of the younger women's are affected by the breast cancer. If eight women live to the age of 85, at least one of them will develop breast cancer in her lifetime. Two thirds of women diagnosed with breast cancer are over the age of 50, and the majority of the remaining women diagnosed with breast cancer are between the ages of 39 and 40. Maximum deaths are registered in India followed by China and USA. In India, Mumbai is one of the leading cities of breast cancer deaths. This implies that one-fourth among all cancer of women is breast cancer. Early detection plays a very important role in the diagnosis of breast cancer. 50% of cases could be solved if the patient undergoes for screening regularly, if failed to diagnoses at screening level, then it might lead to spreading the cause. Hence detecting abnormality without any symptoms and earliest may help to cure cancer.

The breast cancer is most common health issue diagnosed, that leads to cause death among women in both developing and developed countries. It is also the type of cancer that kills the most women. The best known method for preventing breast cancer is early diagnosis, which lowers the mortality rate and enhances treatment efficiency.

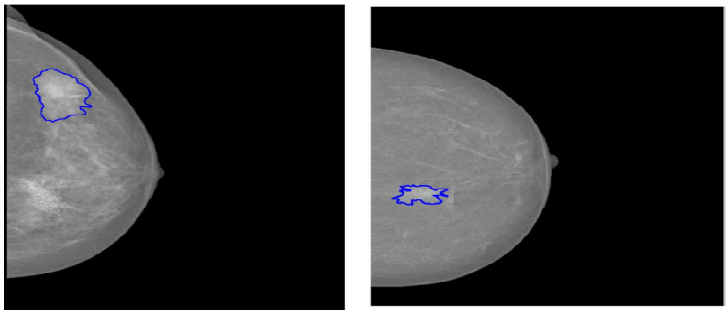
Estimated new Case				estimated deaths		
Year	**Both Sex**	**Male**	**Female**	**Both Sex**	**Male**	**Female**
2005	2,12,930	1,690	2,11,240	40,870	460	40,410
2006	2,14,640	1,720	2,12,920	41,430	460	40,970
2007	1,80,510	2,030	1,78480	40,910	450	40,460
2008	1,84,450	1,990	1,82,460	40,930	450	40,480
2009	1,94,280	1,910	1,92,370	40,610	440	40,170
2010	2,09,060	1,970	2,07,090	40,230	390	39,840
2011	2,32,620	2,140	2,30,480	39,970	450	39,840
2012	2,29,060	2,190	2,26,870	39,920	410	39,510
2013	2,34,580	2,240	2, 32,340	40,030	410	39,620
2014	2,35,030	2,360	2, 32,670	40,430	430	40,000
2015	234,190	2,350	231,840	40,730	440	40,290
2016	249,260	2,600	246,660	40,890	440	40,450
2017	255,180	2,470	252,710	41,070	460	40,610

Table 1. The year wise estimated cases and deaths of breast cancer (according American cancer society)

Furthermore, the medical resource allocation and utilization of particular interest in the case of cancer, it's cost 263.3 billion per year [1]. Breast cancer is the most common cancer in women, with over 1 million new cases diagnosed annually. It is estimated that approximately 500,000 women will die of breast cancer each year, making this the second leading cause of death from cancer in women, with a lifetime risk of the order of 1/10. The molecular events relating to breast cancer biology and pathogenesis had greatly increased over the last decade

Other one of the most effective ways to reduce breast cancer mortality and morbidity is with breast screening program that use mammograms as the main imaging modality. Mean that the mammographic exam... This exam are analyzed by specialists (radiologists).the analysis of breast masses from mammograms represents an important task in the diagnosis of breast cancer , which is mostly a manual process that is susceptible to the subjective assessment of a clinical expert closing operation and image gradient technique to find the region boundary. We highlight the resultant region boundary and detected malignant tissues on the original input image. Table 1 show the estimated new cases and estimated deaths cases

The manual analysis has a sensitivity of 84% and a specificity of 91% in the diagnosis of breast cancer The classification accuracy of this manual interpretation can be improved with the use of a second reading of the mammogram by another clinical expert However, such CAD systems must be robust to false positives and false negatives to be useful in a clinical setting. CAD systems are useful in the detection, segmentation and classification of breast masses, which represent challenging tasks given the low signal-to-noise ratio of the mass visualization, combined with the lack of consistent patterns of shape, size, appearance and location of breast masses. We develop an important and significant method which first detects the cancerous region and then segment the area covered by malignant tissues. We are focusing on to detecting the malignant tissues which represent higher intensity values compared to background information and other regions of the breast. We propose a method including detection followed by segmentation of mammogram images based on simple image processing techniques which provide good results in real time. Such as (1) detection and (2) segmentation. In the detection phase, an averaging filter and thresholding operation is applied on original input image which outputs malignant region area. To find the malignant tissues, we create a rectangular window around the outputted region area and technique. In segmentation phase, a tumor patch is found using morphological



(a) Malignant. Mass **(b) Benign Mass**

Fig. 1

Fig. 1. Two types of breast mass depicted by full field digital mammograms (FFDM) from the IN breast dataset (a) benign (b) malignant

2). Related Studies

In recent decades, many studies have focused on the early detection or diagnosis of breast cancer by means of digital mammograms using image processing and pattern recognition techniques.We review the literature for the problems of mass detection, segmentation and classification in mammograms. We also discuss the current deep learning methods that are relevant to our work.In this section, we provide a brief summary of some works that have a strong connection with the methodology proposed here in.

In [2] Images are the most effective way of revealing information to the world. Images are analyzed and processed by computerized techniques to extract hidden information available in it. Innumerable techniques are available to process the images. In numerous fields, the processed images are used for decision making. In medical field, automated detection and quantitative analysis of theradiological images and other images are processed by Computer Aided Diagnosis (CAD) tool to detect the abnormalities present in images.Segmentation is one of the techniques used in CAD which play a vibrant role in processing the images. It is a process in which regions/ features sharing related characteristics are recognized or grouped together to interpret the images.The segmentation techniques are implemented to analyses mammogram images in future by means of a practical approach.

The work presented in Jasmineetal. [3] Describes a system that automatically classifies breast cancer based on mammographic images .Images from the Mammograms Image Analysis Society (MIAS) was used in this study. The features were extracted using a Non-subsampled Contour let Transform .The work reports a mean maximum accuracy of 98.61% for the classification of regions as normal and abnormal, and 88.05% for classification as malignant and benign, using the support vector machine (SVM).

In [4], we present anmethodology for detecting, segmenting and classifying breast masses from mammograms intervention. This is a problem due to low signal-to- noise ratio in the visualization of breast masses. They are combined with their large variability in terms of shape, size, appearance and location. The problem was break down into three stages: mass detection, mass segmentation, and mass classification.The test in the proposed system on the available in breast dataset and compare the results with the current state-of-the-art methodologies. This evaluation shows that system detects 90% of masses at 1 false positive per image, has a segmentation accuracy of around 0.85 on the correctly detected masses, and overall classifies masses as malignant or benign with sensitivity (Se) of 0.98 and specificity (Sp) of 0.7.

The methodology presented in [5] describes the classification of breast tissue in to mass and non-mass based on regions of interest (ROI) acquired from the DDSM database. The features were extracted by means of Principal Components Analysis (PCA), Gabor wavelet and the efficient coding model based on Independent Component Analysis (ICA) .For classification, the SVM was used, achieving an accuracy of 90.07%.

The methodology developed by [6] in the woman Breast cancer is the most common cancers diagnosed.In Computer-assisted diagnose is systems for breast cancers.The SVM is essentially a local classifier and its performance can be further improved.Experimental and result performedand evaluation from Digital Database for Screening Mammography (DDSM) dataset. In the simple image various types of features, such as curvilinear features,multi-resolution, Gabor features, and texture features fe are extracted. And then select the salient features using the recursive feature elimination algorithm. The structured SVM achieves better detection performance compared with a well-tested SVM classifier in terms of the area under the ROC curve

The work presents in [7]the diagnostic performance and improvesof the breast cancer detection the pectoral muscle detection is an important assignment. An intensity based approach for the pectoral muscle boundary detection in mammograms the intensity based approach is used. The pectoral boundary points from the candidates were detected based on threshold technique. Finally, all the boundary points detected were connected to obtain the boundary of pectoral muscle. The proposed technique has been tested on 320 digitized mammograms form mini-Mammographic Image Analysis Society (MIAS) database of 322 mammograms, with an acceptance rate of 96.56% from expert radiologists. The mean False Positive (FP) and False Negative (FN) rate demonstrate the effectiveness of the proposed method.

The classification of the reigns of mammograms in to mass and non-mass is a critical stage in the development of methodologies for the detection of breast cancer. There lasted works show that methodologies based on texture features provide a good description of the patterns in mammographic images, and statistical measures are widely used.

3). Materials and methods

3.1) Dataset and Methodology

3.1.1) **Dataset**

For measuring the efficiency of the proposed algorithm, we

used the images of Mammographic Image Analysis Society (mini-MIAS) database [23]. The database consists of 322 digitized mammograms which it consist 202 normal and 120 abnormal images. The digitized Images that were 50 micron pixel and represented with an 8-bit word of each pixel and all the images were in 1024×1024 size and padded to a 200 micron pixel.

Table 1. Images Selected for Testing from Mini-MIAS Database

Abnormality type/Tissue type	Fatty	Glandular	Dense
Benign	16	12	6
Malignant	7	7	7

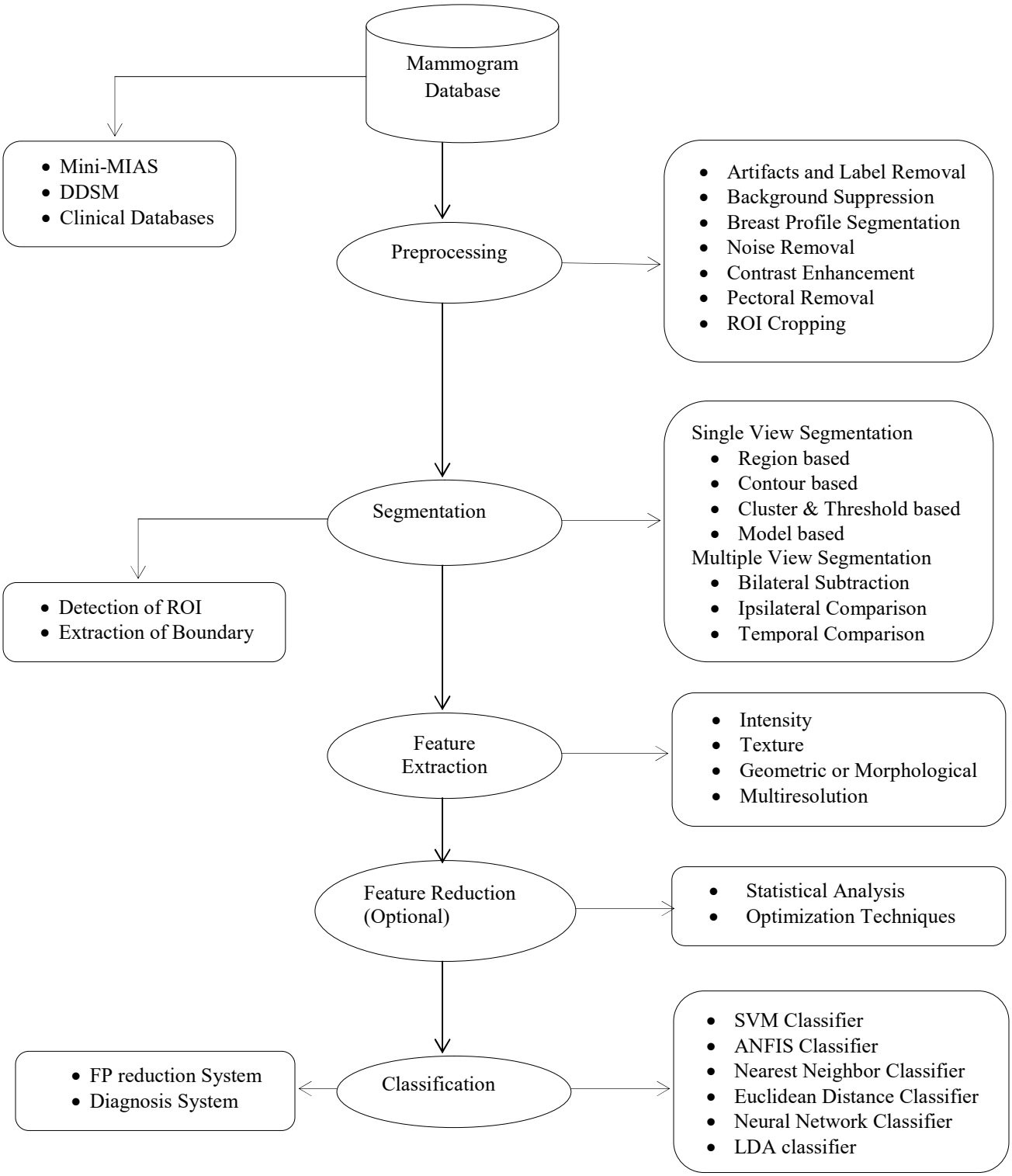


Figure 2 Methodology survey outline

3.2.2) Methodology

The proposed method helps to extract the suspicious region from the breast. Computer aided diagnosis systems on the other hand aim at minimizing interpretation errors. This system is used to classify the suspicious regions in mammogram images. This system makes a decision that the region of interest consists of abnormal or normal tissue and differentiates the abnormalities between benign or malignant type and other classification categories. Though there are different abnormalities present in the mammograms, the most important types of abnormalities are mass and micro calcification.The masses exhibit large variation in size and shape, also it often present poor image contrast. To solve these problems, various techniques for mass detection and diagnosis have been proposed. The existing techniques for preprocessing, enhancement, mass detection/segmentation and classification in digitized mammogram are reviewed. The proposed literature survey outline is shown in Figure 2

3.3) Preprocessing

Preprocessing is the process of simplifying the recognition of abnormalities without leaving the important information. Mammograms may contain noises and low contrast during acquisition. Image enhancement techniques are accomplished through noise removal and contrast enhancement of the images. In mammogram, certain portion of breast regions are superimposed over the background structures which are not necessary for the analysis. The initial preprocessing is done on the input mammogram to remove the noise, artifacts, labels, pectoral muscles, to separate the breast region from the dark background and to enhance contrast in the image. There have been several approaches proposed to segment the breast profile in the mammogram

3.4) Contrast enhancement and noise equalization

Denoising and contrast enhancement of mammogram is very important for CAD system. This noise makes detection of small and subtle structures more difficult. The mammogram images do not provide good contrast between normal and abnormal tissues. Due to the dense breast tissues in younger women, the X-ray attenuation between these two tissues does not differ much. In case of smaller malignancies, it is more difficult for the radiologist to manually outline between normal and abnormal tissues.

The fundamental need of enhancement in mammography is an increase in contrast, especially for dense breasts. Traditional image enhancement techniques are applied to radiography and are often global transformations. So it fails to adopt local information content in the image. Therefore, it is necessary to develop adaptive contrast enhancement techniques, where the transformation is adopted to local context of information. Several techniques for enhancing the mammogram have been reported in literature, such as contrast stretching, histogram Equalizing, filtering, fixed and adaptive neighborhood, morphological operators, unsharp masking and wavelet analysis. Histogram equalization is a contrast enhancement technique that enhances the image by considering the image histogram as a probability distribution. It is an effective and simple technique for contrast enhancement.

3.4.1) Pectoral muscle removal

Pectoral muscle is a dense region close to the chest which may affect the mammogram mass diagnosis process. Due to its higher density than the surrounding tissues, the presence of the same may produce false positive results. it is mandatory to remove the pectoral muscle before mammogram mass detection or segmentation process. The techniques used for pectoral segmentation can be grouped into intensity based segmentation techniques, techniques based on curvature of the edge of the pectoral muscles, wavelet based segmentation techniques, active contour based approaches and model based algorithms

The Intensity based segmentation method depends on the intensity differences between breast tissues and the pectoral muscle. This technique can be highly affected by the fact that the intensity changes between pectoral muscle and breast tissues are generally negligible.

The most commonly used segmentation algorithm for pectoral removal is the seeded region growing algorithm which is simple and efficient.

A method for pectoral muscle extraction based on watershed transformation is proposed. In this technique, watershed transformation is applied on the gradient of a mammogram and watershed regions were extracted. Smoothing followed by merging algorithm is carried out to extract pectoral muscles.

Simple histogram based thresholding technique along with morphological operations has been used to segment the pectoral muscles.

Line detection techniques have been popularly used for pectoral muscle segmentation. This is due to the fact that pectoral muscle boundaries have been assumed to be a straight line.

Dyadic wavelet decomposition has been used for pectoral muscle detection proposed a hybrid method to obtain the delineation of pectoral muscle using Gabor wavelets and pectoral edge using Hough transform. Hybrid method using contour detection and wavelet decomposition to detect the pectoral. A multiple-linked self- organizing neural network approach is proposed to segment

mammogram into four major components including pectoral muscle.

3.5) Mass detection and segmentation

The mammographic detection and segmentation is grouped together according to the computer vision based methodologies. Detection represents identification of potential lesions in the mammogram. It generates a marker at the potential lesion. Segmentation represents detecting precise boundary of the potential lesions. The terms potential and suspicious are interchangeable. There are some algorithms, at the same time, detect and segment masses. there are three possible outcomes for mass detection/segmentation algorithms: detection and/or segmentation of potential lesions, classification of detected lesions as mass or non-mass and diagnosis of a lesion as benign or malignant. The mass detection works in full mammograms, the masssegmentation works in small patch or a given seed point from a mammogram. The detection and segmentation of the mass can also work in full mammograms. Depending on the aim of the approach, the images used are ROIs, single mammograms, pairs of mammograms or full four-mammogram images.

3.5.1) Single view based mass detection

The pixel characteristics within the mass are different from other pixels inside the breast region. They represent either gray level intensity values or texture or morphological measures or distribution of spicules associated with masses. In computer vision, segmentation techniques are mainly divided into unsupervised and supervised methods.

Supervised methods are generally termed as model-based methods. They mainly depend on the prior knowledge about the background and the object to be segmented. Whether a specific region is present within the image are not is determined by this prior information. Unsupervised segmentation divides the image into different homogenous regions based on their specific characteristics such as gray level, texture or color. The approaches to perform the unsupervised segmentation are classified into three important groups.

Region based segmentation divides the image into spatially connected and homogenous regions. Contour based segmentation extracts the boundaries of the regions. Clustering based segmentation groups pixels together those have same properties. This may result in non-connected regions. Thresholding based segmentation is considered as partitioned clustering methods that can be applied to obtain an initial rough representation of suspicious regions. In subsequent step, the result is refined using region or contour based segmentation methods.

a) Region based methods

In region based segmentation, the image is partitioned into connected regions bygrouping neighboring pixels thatare homogenous. This approach is basically divided into two strategies: region growing and split & merges approaches.

Region growing method

Region growing starts from an initial seed point and propagates based on the homogeneity criteria that iteratively increase the size of a region. Several improvements in region growing have been carried out for improving the performance. These improvements can be carried out either before the region growing with the controlled seed selection or the integration of boundary information during region growing. The region growing approach is widely used in mammogram mass segmentation to extract the possible mass region from the background. The two important aspects of region growing algorithm are selection of optimal initial seed point and the homogeneous criteria that control the region growing

1) Watershed method

Watershed segmentation is based on the watershed transform which is a mathematical morphological approach. This method detects the catchment basin that defines the object boundaries. The output of this algorithm is a hierarchy of basins. The selection of most discriminative level of basins is required for each purpose. This methodology is also applied by the researchers in the field of mass detection/segmentation. However, watershed method is sensitive to noise and false edges and also suffers with over-segmentation. Therefore, it needs a necessary pre-processing stage to reduce the over-segmentation.

2) Split & merge method

Split & merge method is also one of the conventional region based segmentation methods. This method recursively split the image until all the regions satisfy the homogenous criteria. In further step, all adjacent regions satisfying second homogenous criterion are merged. In mammographic segmentation, with the region containing the mass to extract its boundary approximately using polygons.

b) **Contour based methods**.

Segmentation based on edge detection is one of the traditional methods in image segmentation. But this segmentation method is far from trivial, since these algorithms do not possess the ability of the human visual system to complete interrupted edges using experience and contextual information. There are only

limited researches on the mass detection using edge based methods, due to the difficulty of extracting the boundary between masses and normal tissue. These methods make use of the filters to find edges, in order to enhance the edges before the detection stage.

c) Model based methods.

In model based methods, at first, the system is trained to detect the specific objects. Subsequently the system has to be able to detect and classify new images based on the presence or absence of similar objects. The training step covers examples with andwithout mass present in the image. The mammograms containing mass have been learned through possible location and the discrepancy in shape and size of the mass. The mammograms without mass have been learned through the features that represent normality. The training phase made the system to learn about the features that can be used to analyze when a new image is presented. The most common used model based segmentation is pattern matching. Pattern matching has been used by many researchers in the field of mass detection in mammograms

3.5.2) Multiple views based mass detection.

Mass detection has also been done by comparing the different mammographic images of same person. The comparison is between either left or right mammograms the algorithms used for the detection of suspicious masses using multiple views of the mammograms. The two important observations have been made when comparing different mammograms of same women. Even though one breast may larger than other, the internal structures have been quite symmetric over broad areas. The overlapping tissues form summation shadows and variations in normal tissues highlight unimportant asymmetries. In order to distinguish masses and asymmetric breast tissues, the characteristic such as size, density, and shape have been considered into account.

3.6) Feature extraction and classification.

In feature extraction, the features that characterize the specific regions are calculated and the important features are selected for the classification of the ROI as normal or abnormal and mass as benign or malignant. The feature space is large and complex due to the wide diversity of normal and abnormal tissues. Some of the features are not significant when observed alone, but combination with other features can be significant. The features are generally categorized into intensity, geometric and texture. The geometric features are also called as morphological or shape features

The segmented mass regions are further classified with step-wise discriminant analysis as benign or malignant disease by computing texture features based on gray level co-occurrence matrix (GLCM) and using the features in a logistic regression method.

The proposed spherical wavelet transform for feature extraction with SVM classifier to classify the detected ROI and reduce the extracted texture features and geometry features from the ROI containing the segmented suspicious regions and the boundaries of the segmentation. The texture features were computed from GLCM and LBP. Finally, the FP reduction was performed by means of SVM, with supervision provided by the radiologist.

4) Conclusion.

Mammography offers high quality images and is the widely accepted imaging method for routine breast cancer screening. The techniques used in CAD systems have major impact on their performance. Although many techniques have been proposed so far, the recent studies show that, the performance of the commercial CAD systems still needs to be improved to meet the requirements of clinic and screening applications. Hence, the improvement on the performance of CAD systems remains a challenging and open problem, particularly in regard to the breast mass detection and diagnosis in mammograms.

We are reviewed different approaches in preprocessing, detection/segmentation and classification of mammographic lesions This described several algorithms, pointing out their specific features. In segmentation, it was clearly shown that few algorithms were contour based, due to the fact that lesions often have not a definite one. Moreover, some region based and clustering algorithms considered shape, gray level or texture information into account to segment lesions. Most of the model based algorithms required the use of a classifier which implies training the system.

The classification stage was mainly used for the FP reduction and the diagnosis of abnormalities into various categories. The feature extraction played an important role in classification stage that affected the classification performance. Various feature extraction and classification techniques were discussed to show the efficiency of the classification system

Over the years there had been an improvement in the detection algorithms, but their performance was still not perfect. Possible reason for such a performance might be the characteristics of abnormalities present in the breast. Masses were sometimes superimposed and hidden in the dense tissue which made the segmentation inaccurate. Another issue was the extraction and selection of features that influenced the results of classifier. The classification of the abnormalities to benign or malignant was also a challenging task even for expert radiologists.

5.) References

[1] Shital Shah, Andrew Kusiak, "Cancer gene search with datamining and genetic algorithms"Computers in Biology andMedicine, Vol. 37, Issue 2, February 2007, pp. 251-261.

[2] M. P. Sukassini1 and T. Velmurugan2. "A Survey on the Analysis of Segmentation Techniques in Mammogram Images" Indian Journal of Science and Technology, Vol 8(22), IPL0259, September 2015.

[3] J.S.L. Jasmine, S. Baskaran ,A. Govardhan, "Automated mass classification system in digital mammograms using contour let transform and support vector machine" Indian Journal of Science and Technology 31(9)(2011)54–61.

[4] NeerajDhungel , Gustavo Carneiro , Andrew P. Bradley ."A deep learning approach for the analysis of masses in mammograms with minimal user intervention" Medical Image Analysis 37 (2017) 114–128.

[5] D.Costa,L .Campos, A. Barros, "A. Classification of breast tissue in mammograms using efficient coding,"BioMed. Eng. On Line10 (1) (2011), ISSN1475-925X.

[6] Defeng Wang, LinShi ,PhengAnnHeng . "Automatic detection of breast cancers in mammograms using structured support vector machines" eurocomputing 72 (2009) 3296–3302.

[7] P. S. Vikhe V. R. Thool "Intensity based Automatic Boundary Identification of Pectoral Muscle in Mammograms" 7th International Conference on Communication, Computing and Virtualization 2016.