Computer Vision and Image Understanding 114 (2010) 33–43

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Automated assessment of breast tissue density in digital mammograms

T.S. Subashini *, V. Ramalingam, S. Palanivel

Department of Computer Science and Engineering, Annamalai University, Annamalainagar 608 002, India

ARTICLE INFO

Article history:
Received 9 September 2008
Accepted 11 September 2009
Available online 24 September 2009

Keywords:
Mammograms
Breast tissue density
Segmentation
Pectoral muscles
Artifact removal
Statistical features
Support vector machines

ABSTRACT

Mammographic density is known to be an important indicator of breast cancer risk. Classification of mammographic density based on statistical features has been investigated previously. However, in those approaches the entire breast including the pectoral muscle has been processed to extract features. In this approach the region of interest is restricted to the breast tissue alone eliminating the artifacts, background and the pectoral muscle. The mammogram images used in this study are from the Mini-MIAS digital database. Here, we describe the development of an automatic breast tissue classification methodology, which can be summarized in a number of distinct steps: (1) preprocessing, (2) feature extraction, and (3) classification. Gray level thresholding and connected component labeling is used to eliminate the artifacts and pectoral muscles from the region of interest. Statistical features are extracted from this region which signify the important texture features of breast tissue. These features are fed to the support vector machine (SVM) classifier to classify it into any of the three classes namely fatty, glandular and dense tissue.The classifier accuracy obtained is 95.44%.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Breast density, a measure of the extent of radio dense fibro glandular tissue in the breast, has the potential to be used as a predictor of breast cancer risk, it is a measure of how well tissue can be seen on mammogram [1]. Some tissue, such as the milk gland, is dense and appears white on a mammogram. This density makes it hard for doctors to see tumors, which also appear white. Fatty tissue is less dense and appears clear on the mammogram, allowing better tumor detection. In 1976, Wolfe published an article that demonstrated a relationship between breast density and breast cancer risk. He showed women with dense breasts have been shown to have a four- to six-fold increased risk of developing breast cancer [2]. Dense tissue in more than 50% of the breast could account for approximately one-third of breast cancers [3]. Mammographic density has also been associated with breast cancer tumor characteristics, including tumor size, lymph node status, and lymphatic or vascular invasion in screen-detected cancers [4]. A threefold increased risk of second breast cancers has also been observed in women diagnosed with ductal carcinoma in situ who have highly dense breasts [5,6].

Mammography has major problems due to high breast density which obscures the mammographic image. The main drawback of mammography today is that it is hard to differentiate between normal, dense tissue and cancerous tissue when looking for small tumors surrounded by glandular tissue. A woman's breasts are naturally denser, or more glandular when young, which makes it difficult for the radiologist to analyze the mammogram image. Technology to detect breast cancer is changing rapidly, with recent entrants to the field like digital mammography and computer-aided detection. Enhancing the image by manipulation of fine differences in intensity by means of image processing algorithms forms the basis of any computer aided detection system. In this work breast tissue is classified based on the intensity level of histogram of a mammogram using SVM. Statistical features of a mammogram are extracted using simple image processing techniques. This technique uses texture models to capture the mammographic appearance within the breast.

1.1. Paper outline

Section 2 provides background information about breast cancer and Section 3 describes mammography and computer assisted screening mammography. A literature review of related computer methods applied in mammography is provided in Section 4. Section 5 gives the description of the proposed method. Artifact removal is discussed in Section 6 and pectoral muscle extraction technique is presented in Section 7. Section 8 covers statistical feature extraction and Section 9 describes the support vector machine (SVM) classifier. Breast tissue classification using SVM is illustrated in Sections 10 and 12 concludes the paper.

2. Breast cancer overview

Breast cancer is the one of the commonest malignancies afflicting women. Despite all medical and technological advances, breast

* Corresponding author. Fax: +91 4144 238275.
E-mail addresses: rtramsuba@yahoo.com (T.S. Subashini), aucsevr@yahoo.com (V. Ramalingam), spal_yughu@yahoo.com (S. Palanivel).

1077-3142/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2009.09.009

cancer cases as been on the rise in the last 50 years or so. It is alarming but it is true that there is more women affected by breast cancer now than ever before. It is currently estimated that one in 14 of all female children born will develop breast cancer in their lifetime. Globally, every 3 min a woman is diagnosed with breast cancer in the world, amounting to 1 million cases annually. The incidence could go up by 50% to 1.5 million by 2020, says the World Cancer Report. The incidence of breast cancer is rising in every country of the world especially in developing countries such as India. Breast cancer is the second leading cause of cancer deaths in women today (after lung cancer) and the most common cancer among women, excluding nonmelanoma skin cancers. According to the WHO, an estimated 1.2 million people worldwide were diagnosed with breast cancer in 2004. Estimates indicated that in another 46,400 women, ductal DCIS, a noninvasive breast cancer, were diagnosed. The incidence of breast cancer increased by approximately 4% during the 1980s but leveled off to 100.6 cases per 100,000 women in the 1990s. According to a study by International Agency for Research on Cancer (IARC), there will be approximately 250,000 new cases of breast cancer in India by 2015 [7–9]. The report shows that around 1.7 million breast cancers were diagnosed worldwide in 2007 and 465,000 (approx.) women died due to breast cancer in 2007. What is more alarming, is that 75–80% of patients are in advanced stages of the disease at the time of diagnosis [10].

If the relationship between tissue density and breast cancer risk is to be studied, a more accurate and objective method of assessing tissue density is needed. In this work an attempt has been made to automate classification of breast density using statistical features.

3. Digital mammograms and computer aided diagnosis (CAD)

In standard mammography, images are recorded on film using an X-ray cassette. The film is viewed by the radiologist using a "light box" and then stored in a jacket in the facility's archives. With digital mammography, the breast image is captured using a special electronic X-ray detector, which converts the image into a digital picture for review on a computer monitor. The digital mammogram is then stored on a computer. With digital mammography, the magnification, orientation, brightness, and contrast of the image may be altered after the exam is completed to help the radiologist more clearly see certain areas. A landmark trial, conducted by the American College of Radiology Imaging Network (ACRIN) in conjunction with the Center for Statistical Sciences at Brown Medical School, showed no difference between digital and film mammography in detecting breast cancer for the general population of women [11]. However the trial shows that digital mammography detected more cancers that is up to 28% more than screen film mammography in women under 50 years of age and women with dense breasts.

Benefits of digital mammograms are quicker mammograms, since there is no need to wait for film images to be developed, images can be viewed instantly by the technologist and radiologist, images can be easily transferred electronically with no loss of image quality and digital mammography allows radiologists to use computer software to manipulate the images in order to optimize their ability to evaluate the breast tissue that might be missed on traditional film mammograms.

Computer aided detection (CAD) increases the detection of early breast cancers, especially those in women with dense breast tissue. With the sharp surge in digital mammography system implementations, CAD is riding the curve, too. According to many radiologists this "second set of eyes" that helps to confirm radiological findings in mammograms, makes sense from both a cost perspective and the potential it offers to detect more breast lesions. CAD is easy to integrate into that process because it is so accessible and available on demand either at the location where the images are taken or at remote offices and home offices that have the necessary high-speed internet connections and high-resolution monitors as well as other equipments.

4. Previous work

Radiologists mainly estimate breast density by visual judgement of the imaged breast. Thus automatic tissue classification methods try to imitate such visual judgment, learning from the radiologist experience. In the literature different approaches for classifying breast tissue based only on the use of histogram information have been proposed [12]. BPNN and histogram features were used in [13] for classifying breast densities. A first approach to qualify mammograms according to the radiographic density was the Wolf scheme [14–16] aimed at finding a correlation between density and cancer risk, but the technique lacked objectivity due to intra and inter observer variations. Recently, researchers have studied intensity-histogram features and applied threshold techniques and fractal characteristics to analyze radiographic density in digital images [17–19].

[20] shows that texture information described by multi-scale histogram based on multi-class DAG-SVM classifier is useful in classifying breast densities.

A semiautomatic computer measure based on interactive thresholding and the percentage of the segmented dense tissue over the segmented breast area has been proposed in [21]. In [22] measures were based on skewness and fractal dimension. Texture-based discrimination between fatty and dense breast types applying granulometric techniques and Laws texture masks has been investigated in [23]. Another method based on fractal dimension is proposed in [22]. Spatial gray level dependency matrices were constructed and features we estimated based on these matrices in[24] to classify breast tissue. In [25] SFS + kNN (sequential forward selection) classifier and morphological and textural features were used for classification. Based on co-occurrence matrices [26,27] segmented mammograms into density regions. Textons are used in [28] to classify breast tissue. However in many of the above approaches the entire breast including the pectoral muscle has been proposed to extract features. The inclusion of the pectoral muscle can affect the results of intensity based image processing methods in the detection of breast densities. In our approach an attempt has been made to restrict the region of interest the breast tissue alone by eliminating the artifacts, background and the pectoral muscle. Statistical features are extracted from this well focussed region which signifies the important texture features of the breast tissue. Numerous techniques have been proposed for breast density pattern classification. Bayesian classifier was used in classifying the breast tissue in [22,21,29]. kNN was used in [26,27,30,31]. In [32,33], rule-based classifiers were used to classify breast tissue density. In the proposed method statistical feature that were extracted are fed to the support vector machine classifier to classify it into any of the three classes namely fatty, glandular and dense tissue.

When mammograms are analyzed by computer, the pectoral muscle should preferably be excluded from processing intended for the breast tissue. In the literature different approaches for automatic pectoral muscle segmentation have been proposed. Segmentation of the breast, and the pectoral muscle are often prerequisites for automatic assessment of breast density. In the work of others, the pectoral boundary has usually been approximated by a straight line as the first step [34,30,35,36]. The straight-line approximation has been determined by a number of techniques, including region growing [34], the Hough transform [30,35,37] and local adaptive thresholding followed by line fitting [36]. Once obtained, the straight-line approximation can be refined to follow the slightly curved outline of the pectoral muscle [36,37] as a further step.

5. Proposed method

The various phases of the proposed method is shown in Fig. 1.

5.1. Preprocessing

First, preprocessing is done on the original mammogram. Radio opaque artifacts, if any, in the original mammogram are removed since the artifacts can contribute to misclassification. Next the pectoral muscle is suppressed. Since mammographic parenchyma and the pectoral muscle have similar texture characteristics, removal of the pectoral muscle is necessary. In this paper histogram based thresholding is used where pixels with gray values greater than 40 are retained while all others are set to zero.

5.2. Feature extraction

Feature extraction is an important part of supervised classification. The number of features selected for breast cancer detection reported in literature varies with the CAD approach employed. It is desirable to use an optimum number of features since a large number of features would increase computational needs, making it difficult to define accurate decision boundaries in a large dimensional space. Statistical features such as mean, standard deviation, smoothness, third moment, uniformity, entropy are extracted from the suspicious regions.

5.3. Classification

This stage makes the final decision regarding the segmented regions. Support vector machine is used to classify the suspicious regions.

5.4. Database description

The proposed work was done using Mini-MIAS database. The Mammography Image Analysis Society (MIAS), which is an organization of UK research groups interested in the understanding of mammograms, has produced a digital mammography database. The X-ray films in the database have been carefully selected from the United Kingdom National Breast Screening Programme and digitized with a Joyce-Lobel scanning microdensitometer to a resolution of 50 m $\times$ 50 m, 8 bits represent each pixel. It has been reduced to 200 μm pixel edge and clipped/padded so that every image is 1024 $\times$ 1024 pixels.

6. Segmentation of artifacts

An artifact is defined as any variation in mammographic density not caused by true attenuation differences in the breast [38]. Factors that create artifacts may be related to the processor (e.g., damp film), the technologist (e.g., improper film handling and loading, improper use of the mammography unit and related equipment), the mammography unit (e.g., failure of the collimation mirror to rotate, grid inhomogeneity, defective compression paddle), or the patient (e.g., motion, jewelry, body parts, clothing, implanted medical devices) and these artifacts can be avoided by following proper quality assurance measures [39]. However for the identification of the patient and the type of X-ray view taken mammograms often contain high-intensity artifacts in the form of identification labels, and wedges. These artifacts can contribute to difficulties in mammogram segmentation and enhancement algorithms. The simple noise removal algorithm does not make any attempt to differentiate the regions of noise from parts of the breast that have been separated from the main breast region by the threshold. By limiting the focus to the removal of high intensity noise it is easier to preserve regions belonging to the breast. The trade off is that some of the medium intensity noise, such as tapes and lower intensity labels, will remain. High intensity noise regions include the bright rectangular labels opaque markers and wedges as indicated in Fig. 2.

It is useful to detect these regions as their spatial location near the skin–air interface may affect neighborhood based operators. The high intensity regions have been the source of error in image orientation algorithms based on the difference of intensities between the breast tissue and image background [40,35]. This section presents an algorithm based on connected component labeling to remove such artifacts. The algorithm is explained in six steps:

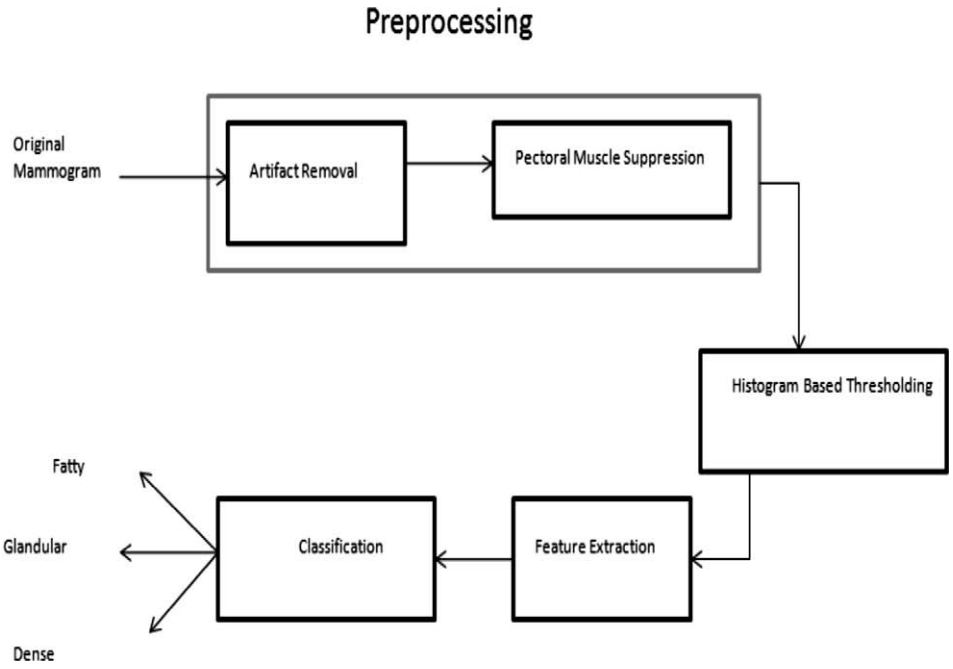

Fig. 1. Various phases of the proposed method.

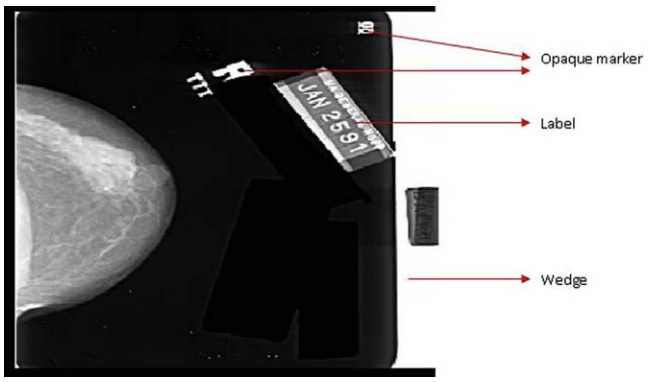

Fig. 2. Artifacts in digital mammogram.

(1) Construction of the intensities histogram. The histogram of a complete mammographic image has the behavior shown in Fig. 3.
   - In the left (lower intensities values) there is a large peak corresponding to the background pixels.
   - In the middle (gray values) there are the pixels corresponding to the breast itself.
   - In the right (brightness pixels) there is another peak corresponding to the pectoral muscle and annotations.

(2) A threshold is used to extract the image from the background. The value of this threshold is determined using the minimum value between the first two most important peaks, which are the peaks of the background and the breast tissue.
(3) A connected component labeling algorithm is used in order to recover the largest region, which will be both the breast and pectoral muscle. Connected components labeling scans an image and groups its pixels into components based on pixel connectivity, i.e., all pixels in a connected component share similar pixel intensity values and are in some way connected with each other. Once all groups have been determined, each pixel is labeled with a gray level or a color (color labeling) according to the component it was assigned to. Connected component labeling works by scanning an image, pixel-by-pixel (from top to bottom and left to right) in order to identify connected pixel regions, i.e., regions of adjacent pixels which share the same set of intensity values.

Fig. 4 shows the experimental results.

7. Pectoral muscle extraction

Previous works on breast tissue identification and abnormalities detection notice that the feature extraction process is affected if the region processed is not well focused. The pectoral muscle represents a predominant density region in mammograms, its inclusion can affect the results of intensity based image processing methods in the detection of breast densities. Thereby, it is important to split the mammogram into targeted regions to achieve optimal breast parenchyma measurements, breast registration or to put into focus a technique when we search for abnormalities. In this paper, the mammogram image is segmented to eliminate the non breast areas such as pectoral muscle, artifacts and background so as to restrict the region of interest to the breast area alone for the assessment of breast tissue density. Both mammographic parenchyma and the pectoral region may have similar texture characteristics, causing a high number of false positives when detecting suspicious masses. In other words, the pectoral muscle could interfere with automated detection of cancers. Also the area overlying the pectoral muscle is a common area for cancers to develop and is particularly checked by radiologists to reduce false negatives. It is, therefore, necessary to segment out the pectoral muscle before lesion detection. Exclusion of the pectoral muscle is required for automatic breast tissue density quantification and classification. This method involves normalization of mammograms to extract the pectoral muscle. Fig. 5 shows the various steps involved in pectoral muscle extraction.

(1) Median filtering is very powerful in removing noise from two-dimensional signals without blurring edges. To apply median filtering to a mammogram, the low-frequency image was generated by replacement of the pixel value with a mean pixel value computed over a square area of 11×11 pixels centered at the pixel location.
(2) Mammogram images are corrected to avoid differences in brightness between the left and right mammograms caused by the recording procedure. In order to reduce the variation, and achieve computational consistency, the images are normalized, by mapping all mammograms into a fixed intensities range $r1$ and $r2$ $(0 \leqslant r1 < r2 \leqslant 255)$.

$$g_k(x,y) = \{r1 + (g_i(x,y) - minG_i) \times (r_2 - r_1)\}/\{maxG_i - minG_i\} \quad (1)$$

Assume an image $g_i(x,y)$ whose maximum gray level is $maxG_i$ and minimum gray level is $minG_i$. The mammogram is nor-

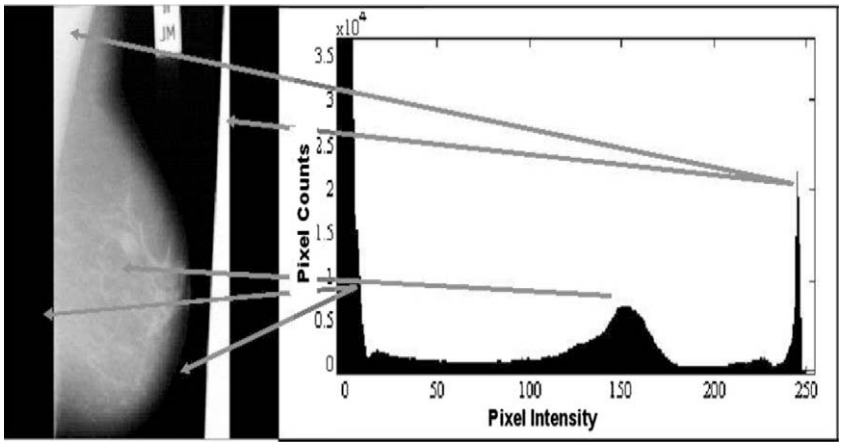

Fig. 3. Typical histogram of a mammogram. Clearly, there are three different zones: background in lowest intensities, breast tissue in medium intensities, and annotations and pectoral muscle in the highest intensities.

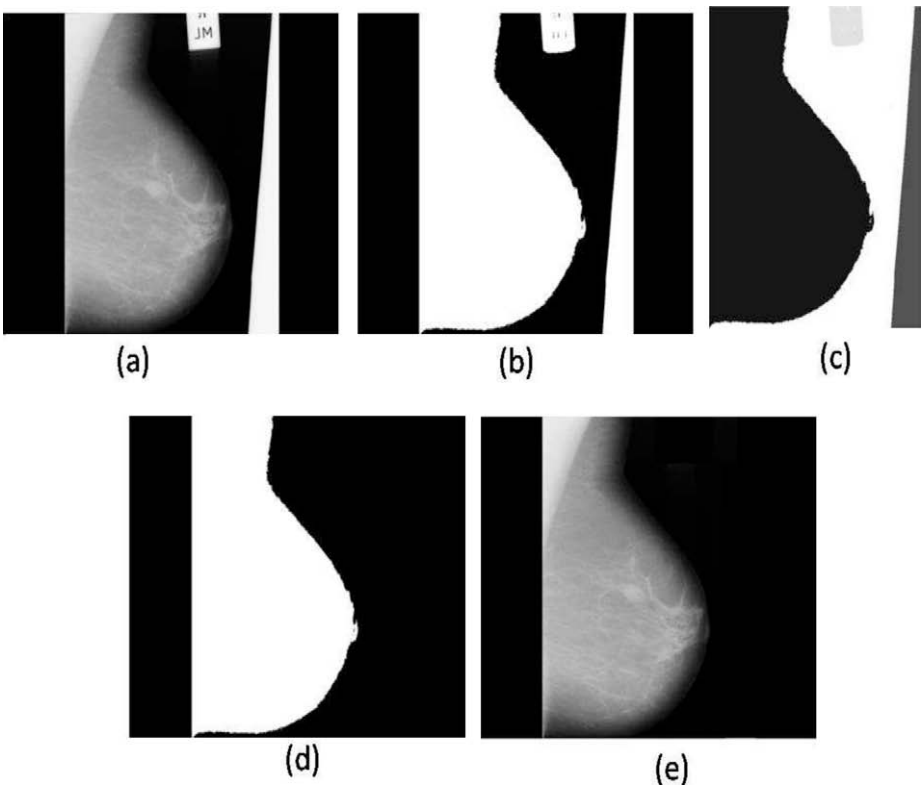

Fig. 4. Segmentation of artifacts using CCL. (a) Original image, (b) image after thresholding, (c) connected component labeled image, (d) binary image without artifacts, (e) image without artifacts.

malized by transforming $g_i(x,y)$ into $g_k(x,y)$ using Eq. (1). The pectoral muscle ia a opaque region which is neither too black nor too white. Based on this assumption the values of $r1$ and $r2$ were found out experimentally as 60 and 210, respectively.

(3) The reliability of boundary matching may be increased by extracting the pectoral muscle from the breast region. The pectoral muscle appears as a bright triangular region in the image corner towards the chest wall and the bottom of the breast region. A histogram-based thresholding technique is

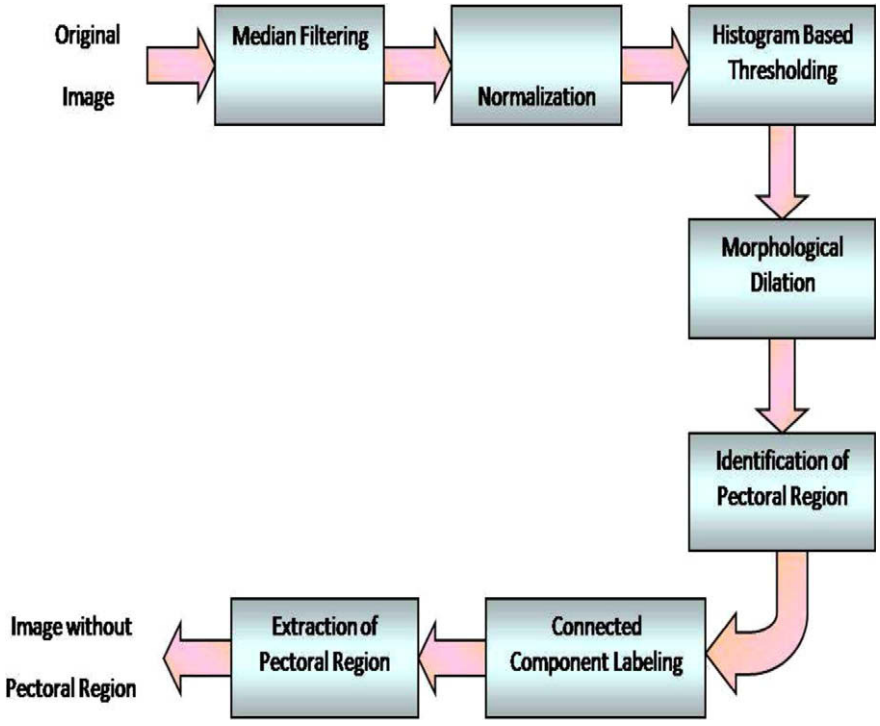

Fig. 5. Various steps involved in pectoral muscle extraction.

used to separate the pectoral muscle region. The global optimum in the histogram is selected as the threshold value. The intensity values smaller than this threshold are changed to black (0), and the gray values greater than the threshold are changed to white 255.

(4) Morphological dilation is applied to the binary image for bridging gaps.
(5) Pectoral muscles lie on the left or right top corner depending on the view of the image. The position of the pectoral muscle (left top corner or right top corner) has to be detected before removing it. For this, a search for non-zero pixels are done simultaneously from both left top corner and right top corner. The position of occurrence of first non-zero pixel is noted. If it is closest to the left top corner, then the pectoral muscle lies in the left top corner or otherwise, it is in the right top corner.
(6) Connected component labeling is applied to the dilated binary image using 8-pixel connectivity.
(7) Using the position of occurrence of first non-zero pixel in the binary image, the gray value of that position in the labeled image is obtained. All the pixels in the image having that same gray value are retained and other pixels are made white, thereby restoring the pectoral region.
(8) Now the original mammogram without pectoral is reproduced. Fig. 6 shows the experimental results.

8. Feature extraction

In general, any image processing and analysis applications would require a particular feature for classification. Feature extraction is an important part of supervised learning and classification. The number of features selected for breast density and cancer detection reported in literature varies with the CAD approach employed. In the literature region based features [3,4], shape-based features [5–7], image structure features [3,11,12], texture based features [13,14] and position related features [13] are described and used. However mainly texture features and statistical features have more significance in pattern recognition area. From the literature it is learnt that in other medical image analysis such as brain tumor classification, breast mass classification the statistical features are found to be more useful in making the classification. Further since the entire breast region is considered for feature extraction statistical feature give more information since it represents the gray level intensity measures. Other features indicated above can be used only in cases where the region of interest is confined to a small region. A frequently used approach for texture analysis is based on statistical properties of intensity histogram. After the preprocessing step the breast region alone is considered for extracting the features.

The statistical features extracted are mean, standard deviation, smoothness, third moment, uniformity, entropy and others. These features shown in Table 1 signify the important texture features of breast tissue and the values obtained are shown in Fig. 7. In Fig. 7 mdbXXX is the identification label where, XXX is the three digit identification number of the mammogram image. From the values in Fig. 7 it can be seen that the features are not redundant. Based on the values of these features an attempt has been made to classify the breast tissue into three basic categories like fatty, glandular and dense.

9. Support vector machine

Statistical machine learning is a research domain mostly related to statistical inference, artificial intelligence, and optimization. Its

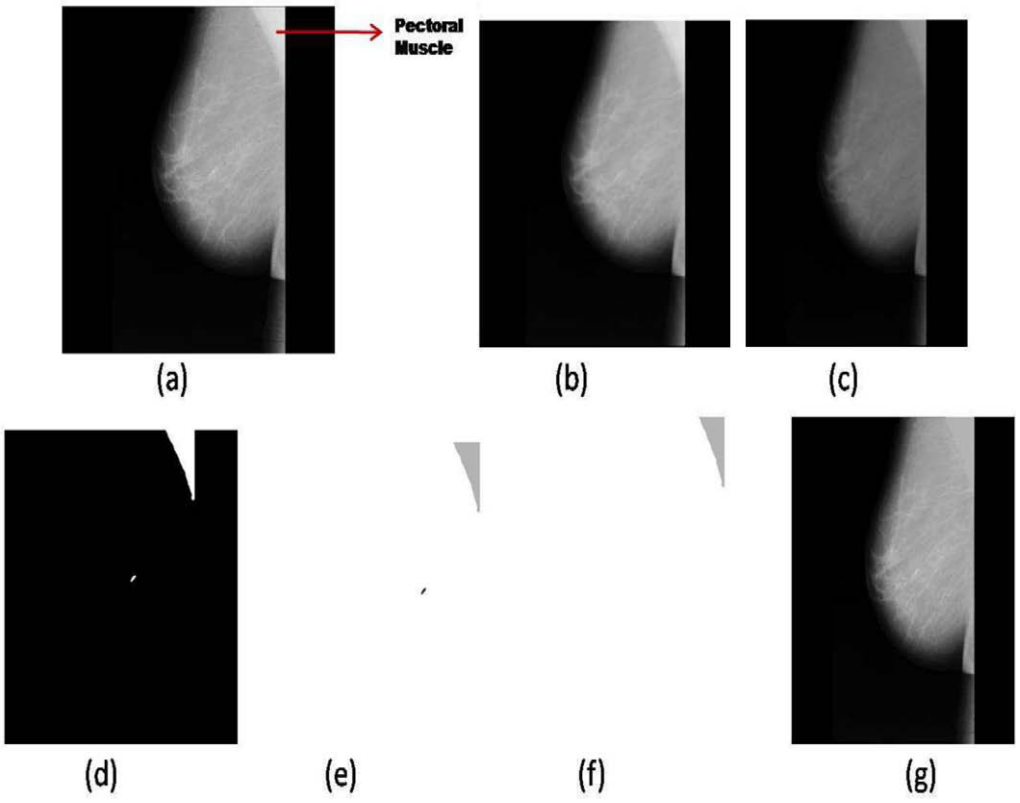

Fig. 6. Pectoral muscle extraction. (a) Original image, (b) median filtered image, (c) normalized image, (d) image after thresholding, (e) labeled image, (f) pectoral region, (g) original mammogram with pectoral suppressed.

Table 1
Summary of statistical features.

Feature	Expression	Measure of texture
Mean	$\mu = \frac{\sum_{i,j} x_{i,j}}{N}$	Mean pixel intensity
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i,j}(X_{i,j}-\mu)^2}{N}}$	The standard deviation of pixel intensity in the region of interest
Smoothness	$R = 1 - \frac{1}{(1+\sigma^2)}$	Measures the relative smoothness of intensity in a region
Skewness	$\frac{\sum_{i,j}(X_{i,j}-\mu)^3}{N\sigma^3}$	A measure of the asymmetry of the pixel values around the image mean
Uniformity	$\sum_{i=0}^{L-1} p(i)^2$	Measures the uniformity of intensity in the histogram
Kurtosis	$\frac{\sum_{i,j}(X_{i,j}-\mu)^4}{(N-1)\sigma^4}$	A measure of whether an image's intensity distribution is peaked or flat relative to the normal distribution
Average histogram	$AH_g = \frac{1}{L}\sum_{i=0}^{L-1} N(i)$	Estimation of the probability of occurrence of a gray level
Modified standard deviation	$\sigma_m = \sqrt{\sum_{i,j}(X_{i,j}-\mu)^2 P(x_{i,j})}$	A measure of average contrast
Modified skew	$\mathrm{MSK} = \frac{1}{\sigma^3}\sum_{i,j}(X_{i,j}-\mu)^3 P(x_{i,j})$	A measure of the asymmetry of the pixel values around the image mean

aim is to construct systems able to learn to solve tasks given a set of examples that were drawn from an unknown probability distribution, and given some a priori knowledge of the task. Another important goal of Statistical Machine Learning is to measure the expected performance of these systems on new examples drawn from the same probability distribution. In the 1990s, a new type

Mammogram	Type of Tissue	Mean	Standard Deviation	Smoothness	Skewness	Kurtosis	Uniformity	Avg. Histogram	Modified Skew	Modified Standard Deviation
mdb006	F	132.5997	26.075720	0.998531	-1.820093	5.741886	0.024604	1797.230469	-697.967285	1473.836304
mdb007	G	134.220505	34.429485	0.999157	-1.012992	3.291501	0.011171	1262.531250	-516.237549	1467.690674
mdb003	D	157.892853	51.558216	0.999624	-0.435822	2.220483	0.007003	1088.242188	491.390991	2124.955078
mdb009	F	135.498962	32.975235	0.999081	-1.295989	3.776298	0.014165	1278.542969	-521.108887	1534.888794
mdb008	G	134.118332	33.188015	0.999093	-1.120895	3.640288	0.012457	1455.574219	-757.429382	1486.399414
mdb033	D	149.465424	43.169956	0.999464	-1.060286	3.163037	0.009228	540.964844	-299.406799	1200.077393
mdb293	F	122.207130	32.161152	0.999034	-0.908600	3.140567	0.013996	1508.976563	-1178.33679	1583.095337
mdb276	G	134.111481	35.011757	0.999185	-0.696104	3.178676	0.009981	1256.671875	-307.758759	1330.472534
mdb035	D	127.860840	49.324020	0.999589	-0.328758	1.697225	0.007292	533.296875	-305.848450	1593.584229

F→ Fatty, G → Glandular and D→ Dense

Fig. 7. Extracted statistical features.

of learning algorithm was developed, based on results from statistical learning theory: the support vector machine. This gave rise to a new class of theoretically elegant learning machines that use a central concept of support vectors and kernels for a number of learning tasks. Kernel machines provide a modular framework that can be adapted to different tasks and domains by the choice of the kernel function and the base algorithm. They are replacing neural networks in a variety of fields, including engineering, information retrieval, and bio-informatics.

Support vector machines are a relatively new learning process influenced highly by advances in statistical learning theory and a sufficient increase in computer processing power in recent years. In the last 10 years SVMs have led to a growing number of applications in image classification and handwriting recognition, to name just a few. Before the discovery of SVMs, machines were not very successful in learning and generalization tasks, with many problems being impossible to solve.

Much like the human brain, SVMs learn by example. Each example consists of a m number of data points $(x_1, x_2 \ldots_m)$ followed by a label (or target), which in the two class classification will be +1 or −1. −1 representing one state and 1 representing another. These states could be cancerous or non-cancerous cells for example. The simplest classification is binary classification. This classification divides two separate classes, which are generated from training examples. The overall aim is to generalize well to test data. This is obtained by introducing a separating hyperplane, which must maximize the margin between the two classes, this is known as the optimum separating hyperplane.

The basic idea of applying SVM to pattern classification can be stated briefly as follows. First, the input vectors are mapped into a feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant with the selection of the kernel function. Then, within the feature space, a hyperplane is constructed which separates two classes (this can be extended to multi-class). Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. An assumption is made that, the larger the margin or distance between these parallel hyperplanes the better will be the generalization error of the classifier. Margin is a distance between optimal hyperplane and a vector which lies closest to it. SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book [41]. Fig. 8 gives a simple view of SVM.

The two classes are then separated by an optimum hyperplane, illustrated in Fig. 9 minimizing the distance between the closest +1 and −1 points, which are known as support vectors. The right-hand side of the separating hyperplane represents the +1 class and the left-hand side represents the −1 class.

9.1. Linear classifier

Given a training set of instance-label pairs

$$\{(\boldsymbol{x_1}, c_1), (\boldsymbol{x_2}, c_2), \ldots\ldots (\boldsymbol{x_n}, c_n)\}$$

where the c_i is either 1 or −1, a constant denoting the class to which the point $\boldsymbol{x_i}$ belongs. Each $\boldsymbol{x_i}$ is a p-dimensional real vector, of normalized $[-1,1]$ values. The scaling is important to guard against variables (attributes) with larger variance that might otherwise dominate the classification. This is the training data, which denotes the correct classification that the SVM will eventually distinguish, by means of the dividing (or separating) hyperplane, which takes the form

$$\boldsymbol{x}\ \boldsymbol{w} - b = 0 \tag{2}$$

The vector **w** points perpendicular to the separating hyperplane. Adding the offset parameter **b** helps to increase the margin. In its absence, the hyperplane is forced to pass through the origin, restricting the solution.

To achieve maximum margin, support vectors and the parallel hyperplanes (to the optimal hyperplane) closest to these support vectors in either class has to be found out as shown in Fig. 9. It can be shown that these parallel hyperplanes can be described by the following equations

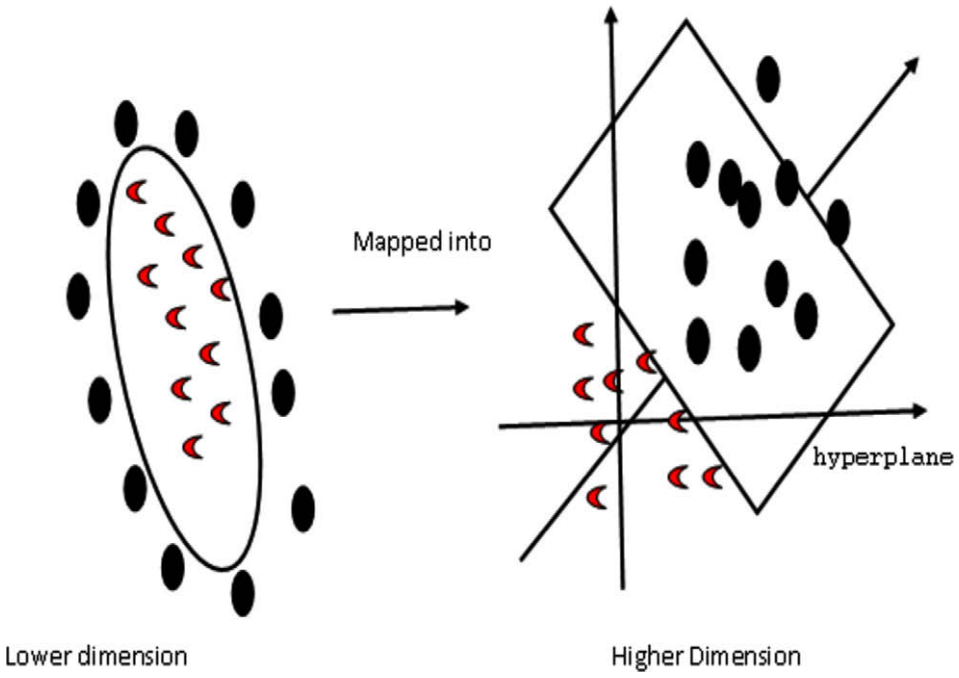

Fig. 8. Simple view of support vector machines.

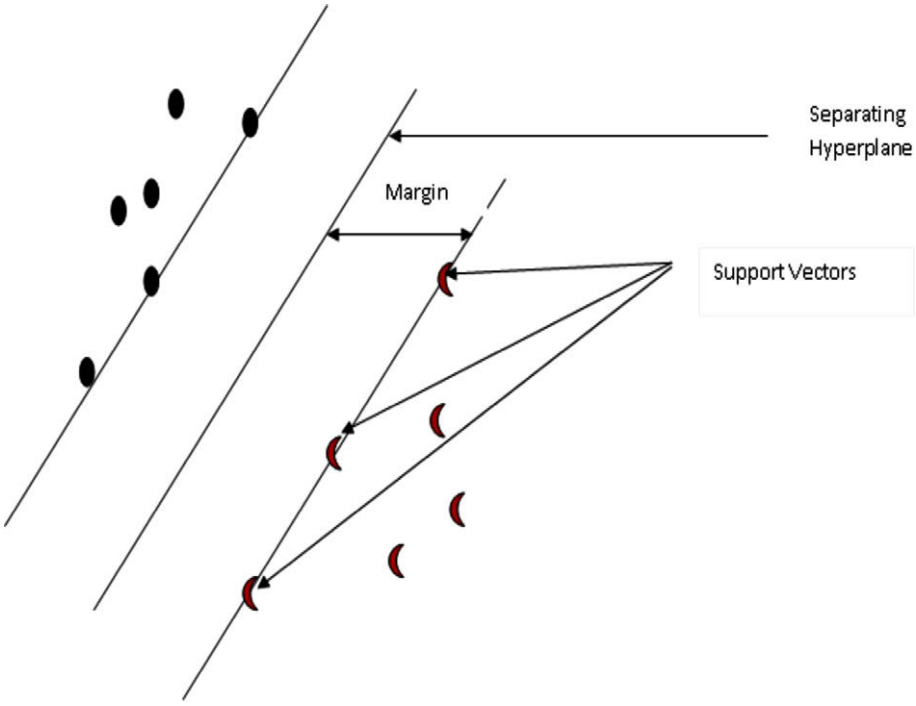

Fig. 9. Maximum margin hyperplane.

$$\boldsymbol{x}\,\boldsymbol{w} - b = 1 \tag{3}$$

$$\boldsymbol{x}\,\boldsymbol{w} - b = -1 \tag{4}$$

If the training data are linearly separable, these hyperplanes can be selected so that there are no points between them and then their distance is maximized to the possible extent. By using geometry, the distance between the hyperplanes is $2/|w|$, which implies that $|w|$ has to be minimized. To exclude data points, care must be taken to ensure that for all i either

$$\boldsymbol{x}\,\boldsymbol{w} - b \geqslant 1 \tag{5}$$

or

$$\boldsymbol{x}\,\boldsymbol{w} - b \leqslant -1 \tag{6}$$

This can be rewritten as

$$c_i(\boldsymbol{x}\,\boldsymbol{w} - b) \geqslant 1, \quad 1 \leqslant i \leqslant n \tag{7}$$

9.2. Non-linear classifiers

The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. However, in 1992, Boser et al. [42] suggested a way to create non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in the transformed feature space [42]. The transformation may be non-linear and the transformed space high dimensional, thus though the classifier is a hyperplane in the high-dimensional feature space it may be non-linear in the original input space. Some common kernels include:

(1) Polynomial (homogeneous): $\kappa(X, X') = (X \cdot X')^d$
(2) Polynomial (inhomogeneous): $\kappa(X, X') = (X, X' + 1)^d$
(3) Gaussian radial basis function: $\kappa(X, X') = \exp\left(-\frac{\|X-X\|^2}{2\sigma^2}\right)$
(4) Sigmoid: $\kappa(X, X') = \tanh(\kappa X \cdot X' + c)$, for some $\kappa > 0$ and $c < 0$.

10. Training the testing the classifier

The statistical features are extracted from 43 normal mammograms in the MIAS database. Of these 14 mammograms have fatty tissue, 14 have glandular tissue and the remaining 15 have dense tissue. Here each pattern consists of nine statistical features. Each of these data points belongs to any of the three classes namely fatty, glandular and dense. All the attributes were first normalized between −1 and +1 in order for the classifier to have a common range to work with. The training data includes the class attribute so a total of (9 features + 1 class attribute) 10 attributes were fed to the classifier while the test data had only nine attributes excluding the class attribute.

SVM Torch a freely available C++ based object-oriented machine learning library was used for training and testing the model [43]. The radial basis function kernel of the SVM is used. The commonly used kernel functions for instance, Gaussian, polynomial and sigmoidal tends to project data onto a very high dimensional space very often, an overfitting risk. SVMs tend to limit this effect by controlling the capacity of the machine through maximization of the margin of the hyperplane selection. While the choice of the kernel still remains a research issue for a given problem, a RBF kernel has been used since good results in many practical problems have been reported so far. The SVM was trained in multi-class mode using two-third of the data randomly chosen and tested with the remaining one-third of the data for evaluating the classifier's effectiveness. In order to evaluate the result threefold cross-validation was used. A program was written to divide the patterns randomly into three different sets for training and testing the classifier. Threefold cross-validation was carried out by training and testing the model with three different sets of train and test data values. This was done to avoid bias in classification.

The SVM is trained to provide a value of 0 for fatty tissue mammograms, 1 for glandular and 2 for dense tissue. Tables 2–4 show the classification results obtained using SVM.

From the tables the overall classification rate of the classifier is calculated as 95.44%.

Table 2
Results for breast tissue classification using SVM-cross-validation I.

Type	Fatty	Glandular	Dense
Class	14	14	15
Recognizes results	14	14	13
Classification rate	100%	100%	86.67%

11. Comparison and discussion

In our work initially 14 features were extracted. They are number of pixels, average gray level, average histogram, energy, modified energy, entropy, modified entropy, standard deviation, modified standard deviation, skew, modified skew, difference, contrast and average boundary level [44] and we trained the classifier using different combination of feature vectors. But we found that the classifier gave good performance of 93.02% with just six features namely mean standard deviation, smoothness, skew, kurtosis and uniformity. It was also found that there was a increase in the performance when three more features namely average histogram, modified standard deviation and modified skew were added. The performance obtained using these nine features was 95.44%. The performance did not improve but reduced when other features were added. Thus we arrived at an optimal set of nine features. And these features are not redundant. The only misclassified images were that of glandular and dense and that is justified because there is only a slight intensity variation between glandular and dense breast tissue and usually here is where the radiologist themselves find it difficult to differentiate by visual interpretation.

Table 5 gives a comparison of our work with that of other works that uses different features and classification models. In all the cases the MIAS dataset was used to test the performance.

In [13] features derived from the histogram included the lowest intensity value of the image, the ratio between the lowest intensity value and the highest intensity value, the ratio of the distance between the initial and the peak values to the total range of the distance, and the ratio of the number of pixels falling between the peak and the highest intensity values to the total number of pixels. A back propagation neural network was established to classify tissue composition with the features derived from histograms and the overall accuracy of the neural network obtained was 71%. In [20] the combination of texture and intensities information for breast density classification using multi-resolution histogram technique was investigated and this approach gave an accuracy of 77.57% when a DAG-SVM classifier was used.

[25] used a set of morphological and texture features. As morphological features, the relative area and the first four histogram moments were calculated. A set of features derived from co-occurrence matrices were used as texture features. The (sequential forward selection) SFS + kNN classifier was used to classify the breast density and the accuracy obtained was 91%. Here, the authors have used the sequential forward selection (SFS) algorithm, which is a widely known technique that selects a local optimum solution in a computationally attractive way. [23]shows that texture analysis forms a good basis for automatically classifying breast tissue using bayesian classifier and the accuracy reported is 80%.

Table 3
Results for breast tissue classification using SVM-cross-validation II.

Type	Fatty	Glandular	Dense
Class	14	14	15
Recognizes results	14	14	14
Classification rate	100%	100%	93.3%

Table 4
Results for breast tissue classification using SVM-cross-validation III.

Type	Fatty	Glandular	Dense
Class	14	14	15
Recognizes results	13	13	14
Classification rate	92.85%	92.85%	93.33%

Table 5
Comparison summary of the proposed work with other works.

Classifier used	Features Used	Accuracy	Reference
BPNN	Histogram features	71%	Wang et al. [13]
DAG-SVM	Multi-resolution histogram feature	77.57%	Zwiggelaar [20]
SFS + kNN	Morphological and textural features	91%	Oliver et al. [25]
Bayesian	Laws texture features	80%	Miller and Astley [23]
SVM	Statistical features	95.44%	Proposed work

It can be seen form the Table 5 that SVM using statistical features is more accurate in classifying the breast tissue according to its density namely fatty, glandular or dense.

12. Conclusion

In this paper, SVM based pattern separation model is applied to statistical features extracted from the breast parenchyma for classification of breast tissue density. Features were extracted from the breast region after preprocessing and segmenting the mammogram. Artifacts were removed by applying global thresholding and connected component labeling and further pectoral region of the breast was removed using histogram based thresholding and morphological operations. Features were extracted using image processing techniques. The mammogram images were processed using Vc++ and Intel's Image Processing library. The radial basis function kernel of the SVM was used to classify feature vectors describing the tissue into any of the three classes namely fatty, glandular and dense tissue. The classifier accuracy obtained is 95.44%.

Acknowledgment

The authors thank Dr. M.K. Sivakkolunthu, Professor of Radiology, Raja Muthiah Medical College Hospital, Annamalai Nagar for his valuable help and comments in carrying out this work.

References

[1] V.A. McCormack, I. dos Santos Silva, Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis, Cancer Epidemiology Biomarkers and Prevention (2006) 1159–1169.
[2] J.N. Wolfe, Breast patterns as an index of risk for developing breast cancer, Journal of Roentgenology 26 (1976) 1130–1139.
[3] N.F. Boyd, J.M. Rommens, K. Vogt, et al., Mammographic breast density as an intermediate phenotype for breast cancer, Lancet Oncology 6 (2005) 798–808.
[4] E.J. Aiello, D.S. Buist, E. White, et al., Association between mammographic breast density and breast cancer tumor characteristics, Cancer Epidemiology Biomarkers and Prevention 14 (2005) 662–668.
[5] L.A. Habel, J.J. Dignam, S.R. Land, Mammographic density and breast cancer after ductal carcinoma in situ, Journal of the National Cancer Institute (2004) 1467–1472.
[6] M.L. Irwin, E.J. Aiello, A. McTiernan, L. Bernstein, F.D. Gilliland, R.N. Baumgartner, K.B. Baumgartner, R. Ballard-Barbash, Physical activity, body mass index, and mammographic density in postmenopausal breast cancer survivors, Journal of Clinical Oncology 25 (9) (2007) 1061–1066.
[7] ICMR, National Cancer Registry Programme, Consolidated report of the population based cancer registries, 1990–1996, Indian Council of Medical Research, New Delhi, 2001.
[8] ICMR, National Cancer Registry Programme, Consolidated report of the hospital based cancer registries, 1984–1993, Indian Council of Medical Research, New Delhi, 2001.

[9] ICMR, National Cancer Registry Programme, 1981–2001, An Overview, Indian Council of Medical Research, New Delhi, 2002.
[10] M. Chatterjee, Breast cancer will become epidemic in India, April 2008.
[11] Digital mammography detects more breast cancers than screen film mammography, September 2005.
[12] C. Zhou, H.P. Chan, N. Petrick, M.A. Helvie, M.M. Goodsitt, B. Sahiner, L. Hadjiiski, Computerized image analysis: estimation of breast density on mammograms, Medical Physics 28 (6) (2001) 1056–1069.
[13] X.H. Wang, W.F. Good, B.E. Chapman, Y.-H. Chang, W.R. Poller, T.S. Chang, L.A. Hardesty, Automated assessment of the composition of breast tissue revealed on tissue-thickness-corrected mammography, American Journal of Roentgenology 180 (2003) 227–262.
[14] J. Wolfe, Breast patterns as an index of risk of developing breastcancer, Journal of Roentgenology 126 (1976) 1130–1139.
[15] J. Wolfe, Risk for breast cancer development determined by mammographic parenchymal pattern, Cancer 37 (1976) 2486–2492.
[16] J. Byng, M. Yaffe, G. Lockwood, L. Little, D. Tritchler, N. Boyd, Automated analysis of mammographic densities and breast carcinoma risk, American Cancer Society 80 (1) (1997) 66–74.
[17] J. Byng, M. Yaffe, R. Jong, R. Shumak, G. Lockwood, D. Tritchler, N. Boyd, Analysis of mammographic density and breast cancer risk from digitized mammograms, InfoRad 18 (6) (1998) 1587–1598.
[18] N. Boyd, G. Lockwood, L. Martin, J. Byng, M. Yaffe, D.L. Tritchler, Mammographic density as a marker of susceptibility to breast cancer: a hypothesis, International Agency for Research on Cancer Scientific Publications 154 (2001) 163–169.
[19] G. Corkidi, L. Vega, J. Mrquez, E. Rojas, P. Ostrovsky-Wegman, A roughness feature of metaphase chromosome spreads and nuclei for automated cell proliferation analysis, Medical and Biological Engineering and Computing 36 (6) (1998) 679–685.
[20] I. Muhimmah, R. Zwiggelaar, Mammographic density classification using multiresolution histogram information, in: Proceedings of the International Special Topic Conference on Information Technology in Biomedicine (ITAB '06) Ioannina, Greece, <http://medlab.cs.uoi.gr./itab2006/proceedings/Mammography/59.pdf>, 2006.
[21] N.F. Boyd, J.W. Byng, R.A. Jong, E.K. Fishell, L.E. Little, A.B. Miller, G.A. Lockwood, D.L. Tritchler, M.J. Yaffe, Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian national breast screening study, Journal of the National Cancer Institute 87 (9) (1995) 670–675.
[22] J.W. Byng, N.F. Boyd, E. Fishell, R.A. Jong, M.J. Yaffe, Automated analysis of mammographic densities, Physics in Medicine and Biology (1996).
[23] P. Miller, S. Astley, Classification of breast tissue by texture and analysis, Image and Vision Computing 10 (1992) 277–282.
[24] K. Bovis, S. Singh, Classification of mammographic breast density using a combined classifier paradigm, in: Proceedings of Medical Image Understanding and Analysis, 2002, pp. 177–180.
[25] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E.R.E. Denton, R. Zwiggelaar, A novel breast tissue density classification methodology, IEEE Transactions on Information Technology in Biomedicine 12 (1) (2008) 55–65.
[26] R. Zwiggelaar, S.M. Astley, C.J. Taylor, C.R.M. Boggis, Linear structures in mammographic images: detection and classification, IEEE Transaction on Medical Imaging 23 (9) (2004) 1077–1086.
[27] R. Zwiggelaar, L. Blot, D. Raba, E.R.E. Denton, Set-permutation occurrence matrix based texture segmentation, IEEE Transaction on Medical Imaging 18 (8) (1999) 712–721.
[28] S. Petroudi, T. Kadir, M. Brady, Automatic classification of mammographic parenchymal patterns: a statistical approach, in: Proceedings of International Conference IEEE Engineering in Medicine and Biology Society, vol. 1, 2003, pp. 798–801.
[29] X. Munoz, J. Freixenet, X. Cufi, J. Marti, Strategies for image segmentation combining region and boundary information, Physical Review Letters 24 (2003) 375–392.
[30] N. Karssemeijer, Automated classification of parenchymal patterns in mammograms, Physics in Medicine and Biology 43 (1998) 365–378.
[31] L. Blot, R. Zwiggelaar, Background texture extraction for the classification of mammographic parenchymal patterns, in: Proceedings of Medical Image Understanding and Analysis, 2001, pp. 145–148.
[32] Y.C. Gong, M. Brady, S. Petroudi, Texture based mammogram classification and segmentation (2006).
[33] K.E. Martin, M.A. Helvie, C. Zhou, M.A. Roubidoux, J.E. Bailey, C. Paramagul, C.E. Blane, K.A. Klein, S.S. Sonnad, H.P. Chan, Mammographic density measured with quantitative computer-aided method: comparison with radiologists estimates and bi-rads categories (2006).
[34] F. Georgsson, A transformation of mammograms based on anatomical features: in digital mammography, in: Proceedings of the 5th International Workshop on Digital Mammography, Madison, WI, Medical Physics Publishing, 2001, pp. 721–726.
[35] R. Chandrasekhar, Systematic segmentation of mammograms, Ph.D. thesis, The University of Western Australia, 1996.
[36] S.M. Kwok, R. Chandrasekhar, Y. Attikiouzel, Automatic pectoral muscle segmentation on mammograms by straight line estimation and cliff detection, in: Proceedings of Information Systems Conference, Perth, Australia, 2000, pp. 67–82.
[37] M. Yam, M. Brady, R. Highnam, C. Behrenbruch, R. English, Y. Kita, Three-dimensional reconstruction of microcalcification clusters from two mammographic views, IEEE Transactions on Medical Imaging 20 (6) (2001) 479–489.
[38] B. LW, Quality determinants of mammography: clinical image evaluation, Radiological Society of North America (1995) 57–67.
[39] J.P. Hogge, C.H. P, et al., Quality assurance in mammography: artifact analysis, Radiographics 19 (1999) 503–522.
[40] S.L. Lou, H.D. Lin, K.P. Lin, D. Hoogstrate, Automatic breast region extraction from digital mammograms for pacs and tele mammography applications, Computerised Medical Imaging and Graphics 24 (2000) 205–220.
[41] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[42] B.E. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth ACM Workshop on Computational Learning Theory, ACM Press, New York, 1992, pp. 144–152.
[43] R. Collobert, S. Bengio, SVM Torch: support vector machines for large-scale regression problems, Journal of Machine Learning Research 1 (1) (2001) 143–160.
[44] P. Zhang, B. Verma, K. Kumar, Neural vs. statistical classifier in conjunction with genetic algorithm feature selection in digital mammography, IEEE Congress on Evolutionary Computation 2, 1206–1213.