Survey on Methods of Tumor Region Detection and Tumor Classification in Mammographic Image

Neha V. Chauhan[1] Jayna B. Shah[2]
[2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat India

Abstract— At present in Medical science, the reason for the tremendous growth of breast cancer is unknown. Mammography or mammograms play an important role in early detection of breast cancer. Mammography is specialized medical imaging that uses a low-dose x-ray system to see inside the breasts. Digital mammography convert x-rays into mammographic images of the breast. A lump or tumor will show up as a focused white area on a mammogram image. Tumors can be cancerous or benign. Researchers have been emerged many techniques to detect tumor and classify it as cancerous or benign. This survey paper focus on the different techniques on enhancement of mammogram images, detection and classification of breast tumor. In this survey papers, the authors attempt different enhancement techniques such as filtering with morphological operation, Histogram equalization, median filtering, Contrast limited adaptive histogram equalization (CLAHE) to enhance the image. They used various feature extraction methods like GLCM, Gabor filter and DWT to extract features from image. They used various classification methods such as SVM, ANN, BPNN, Bayesian classifier, KNN to classify the image.

Key words: Mammograms, Breast Cancer, Enhancement Classifiers, MIAS

I. INTRODUCTION

Breast cancer is leading cause of death and also one of the most invasive types of cancers among women in worldwide. It happens when cells in the breast start to develop uncontrollably or spread throughout the body. Early detection and effective diagnosis is the only rescue to lessen breast cancer fatality. Accurate classification of breast tumor is an important task in medical diagnosis. The goal of this survey paper is to determine the current state of research in breast cancer and to help extract the key features and problems with existing expert systems. There are many numbers of quantitative models based on support vector machine, neural network, fuzzy logic, hybrid and many others techniques are being operated in medical field to help decision makers in breast cancer detection. The comparison of the various systems is done on the basis of data sets used for diagnosis, the methodology applied and the platform on which the system is implemented.

Although breast cancer disease is still a major cause of death in women, the breast cancer mortality exhibits a decreasing rate with the help of early detection of tumor, appropriate therapy and accurate treatment[1]. The detection and treatment of breast cancer in its earlier phases can reduce the death of patient due to breast cancer [2]. Detection of breast cancer tumor in early stages is most important in low and middle income countries for survival where available medical facilities are very limited so that the number of breast cancer cases will decreases in near future.

II. RELATED WORK

The region of interest is determined from the morphological top hat filtered image by means of thresholding segmentation, in [1]. Various features like first order textural features, Gray Level Co-occurrence Matrix (GLCM) features, Discrete Wavelet Transform (DWT) features, run length features and higher order gradient features are derived for the particular ROI. Support Vector Machine (SVM) classifier is trained with the above mentioned features using MATLAB bioinformatics tool box. Thus the classified results are obtained for the query image based on the trained SVM structure, in [1]. The main advantage of the proposed method is that the number of false positives has been reduced up to 1 for every 100 images.

Grey Level Co-occurrence Matrices (GLCM) technique used in study of remote sensing images. Up till now in breast cancer detection only first and second order GLCM features were mostly used, to the best of our knowledge. In this paper the authors attempted up to 7th order and observed the results by analyzing the effects of higher order features in recognition of malignancy in breast mammograms. They observed that 3rd order GLCM features combined with first and second order significantly improved the classification, in [2].

Combined system based on artificial neural networks (ANN) and complex wavelet transform is proposed in [3]. The study using 322 images of the MIAS database have resulted in classification success rates ranging from 80% to 94.79% for different breast tissue density classes.

The shuqi Cui, Hong Jiang, and Zheng Wang authors used [4] Scale Invariant Feature Transformation (SIFT) algorithm combined with SVM classifier and sliding window to extract the local features and describe ROI precisely in the image. Finally, the extracted feature is used as the input layer of BP neural network in mammary gland X - ray image classification. The experimental results show that the accuracy of neural network classifier based on SIFT is 96.57%, which is 3.44% higher than that of traditional SVM classification accuracy.

In paper [5] three new features are proposed, which can be used to classify breast tissue density into fatty and dense tissue type. The new proposed features are used with gray level co-occurrence matrix features to classify the mammograms through optimal feature selection process. The new features are based on the intensity of the grey level of the image. To corroborate the significance of new features, various standard classifiers are used. The results are able to perceive the feasibility of the proposed method to classify the breast density tissue into fatty and dense. The new proposed method gives 94.5% accuracy.

In paper [6], the author Moustapha Mohamed Saleck[1], Abdelmajid ElMoutaouakkil[2] were introduced a new approach using Fuzzy C-means algorithm, in order to extract the mass from region-of- interested (ROI). The proposed method aims at avoiding problematic of the estimation of the cluster number in FCM by selecting as input data, the set of pixels which are able to provide us the information required to perform the mass segmentation by fixing two clusters only. The Gray Level Occurrence Matrix (GLCM) is used to extract the texture features for getting the optimal threshold, which separate between selected set and the other sets of the pixels that influences on the mass boundary accuracy. The performance of the proposed method is evaluated by specificity, sensitivity and accuracy. The results obtained from experimentations shows a good efficiency at the different measures used, in favor of our method.

III. PREPROCESSING TECHNIQUES

Various preprocessing techniques used in referred papers are thresholding, fuzzy-c-means, median filter, PCA (principal component analysis), contrast enhancement, morphological operation, contrast-limited adaptive histogram equalization.

1) Thresholding

Thresholding is a segmentation technique that is used to classify each pixel into two classes namely vessels and non-vessels. It produces a binary image on the basis of whether the intensity value of the image is greater or lesser than a certain threshold value. The threshold value defines the quality of the segmentation.

2) Fuzzy-c-means

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. Fuzzy – c-mean is used for segmentation.

3) Median filter

Instead of using box filter or weighted average filter, for pixel value at a particular location in the processed image will be the median of the pixel in the neighborhood of the corresponding location on the original image, in that case this filter also reduce the noise but at the same time, it rise to maintain the contrast of the image.

When we used median filter, noise is reduce in the image and it maintain the sharpness of the image but image get blurred.

4) PCA (principal component analysis)

It is a technique to reduce the image dimension by producing a small number of independent images called principal component to represent the variability of the data [9].

5) Morphological operation

This word widely used in the field of biology, it discuss about shapes and structure of different animal, plants and so on. We need to separate of object from background region. *Morphology* is a broad set of image processing operations that process images based on shapes. Morphological operations apply a structuring element to an input image, creating an output image of the same size. In a morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors. By choosing the size and shape of the neighborhood, you can construct a morphological operation that is sensitive to specific shapes in the input image. The number of pixels added or removed from the objects in an image depends on the size and shape of the *structuring element* used to process the image.

6) Contrast-limited adaptive histogram equalization

The contrast limited adaptive histogram equalization techniques, have been applied on mass cluster area to enhance the contrast of small tiles and to combine the neighboring tiles in mass area by using bilinear interpolation, this phase is used to improve the accuracy of texture feature results[7].

Method	Advantages	Limitation
Thresholding	no need of previous information, simplest method	highly dependent on peaks, spatial details are not considered
Fuzzy-c-means	Good accuracy,	Lower speed, longer run time.
Median filter	Works good with sharp/bright features.	Difficult to predict median filter analytically
PCA	Data compression, fast. Accurate	Less accuracy because of data dimension reduction.
Morphological operation	Great segmentation results.	Parameters need to be initialized as per different image properties.
CLAHE	Good equalization results because of adaptive nature	Slow as compared to other histogram equalization techniques.

Table 1: Comparison table of preprocessing techniques

IV. FEATURE EXTRACTION

Image feature is the one piece of information. In the image feature extraction is the process of transfer arbitrary data like image to the relevant numeric data. This numeric data used in the classification process.

A. What is Feature selection (or Variable Selection)?

Problem of selecting some subset of a learning algorithm's input variables upon which it should focus attention, while ignoring the rest. In other words, Dimensionality Reduction. As Humans, we constantly do that! Mathematically speaking,

– Given a set of features $F = \{ f1 , ..., fi , ..., fn \}$ the Feature Selection problem is to find a subset that "maximizes the learner's ability to classify patterns"

Fig. 1: Feature Selection

– Formally *F'* should maximize some scoring function
– This general definition subsumes feature selection (i.e. a feature selection algorithm also performs a mapping but can only map to subsets of the input variables)

- Feature selection can be significantly beneficial for achieving following two goals.
- Especially when dealing with a large number of variables there is a need for Dimensionality Reduction
- Feature Selection can significantly improve a learning algorithm's performance

B. Feature Selection—optimization process.

The goal is to find an optimal feature-subset (one that maximizes the scoring function).

In real world applications this is usually not possible.

- For most problems it is computationally intractable to search the whole space of possible feature subsets
- One usually has to settle for approximations of the optimal subset
- Most of the research in this area is devoted to finding efficient search-heuristics

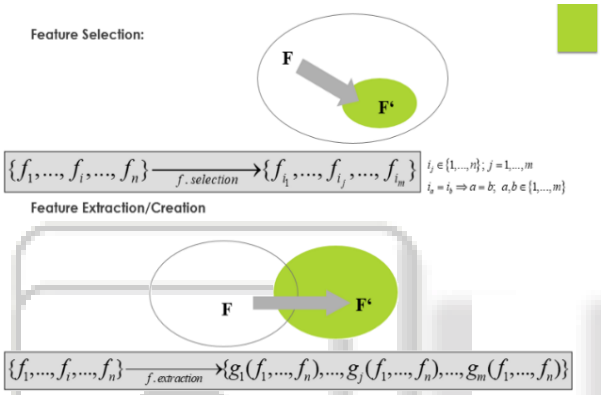

Fig. 2: Difference between feature selection and Feature extraction

C. Optimal Feature Subset:

Often, the definition of optimal feature subset in terms of classifier's performance

The best one can hope for theoretically is the Bayes error rate

D. Relevance of Features

There are several definitions of relevance in literature.

- Relevance of 1 variable, Relevance of a variable given other variables, Relevance given a certain learning algorithm.
- Most definitions are problematic, because there are problems where all features would be declared to be irrelevant
- This can be defined through two degrees of relevance: weak and strong relevance.
- A feature is relevant *iff* it is weakly or strongly relevant and irrelevant (redundant) otherwise.

Various feature extraction techniques used in referred papers are GLCM, DWT, and Gabor filter.

1) GLCM

GLCM texture consider the relation between two neighboring pixels in one offset, as the second order texture. The gray value relationship in a target are transformed into the co-occurrence matrix space by given kernel mask such as 3*3, 5*5, 7*7 and so forth. In the transformation into the co-occurrence matrix space, the neighboring pixels in some of the eight defined direction can be used. Four direction: 0^0, 45^0, 90^0, 135^0. It contain information about the positions of the pixels having similar gray level values.

2) Discrete wavelet transform

Texture feature using wavelet transform is used for generation of features. In wavelet transform the image is represented in terms of the frequency content of local regions over a range of scale. DWT is applied on set of images and statistical features such as mean and energy are extracted from the approximation and detail region of DWT decomposed image. The image is decomposed into four sub-bands LL1, LH1, HL1, and HH1. [2]

3) Gabor filter

A Gabor filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function. The 2D Gabor filter function is obtained by modulating a sinusoidal plane wave at particular frequencies and orientations with a Gaussian envelope [9].

V. CLASSIFICATION

A. SVM

It is a machine learning method which uses a hyperplane that maximizes the margin in the training data to classify binary classes. Support vectors are the training data along the hyperplane. The distance between the support vectors and the class boundary is the margin. The decision planes that define decision boundaries are the basic idea of SVM.

B. BP neural network

It is the supervised method for training artificial neural network. It is multi-layer feed-forward neural network with no backward loop but it back propagates the error so that it is named as "Back propagation neural network". Back propagation calculates the gradient of the error of the network regarding the network's modifiable weights. Each layer in this multi-layer neural network has activation functions associated with it. It is mainly divided into two parts: training algorithm and application algorithm or testing algorithm [10].

C. Bayesian classifier

Naive Bayes (classifier) is a type of generative model that models each possible category based on training samples. We call it as "Naive Bayes" because of the assumption that each attribute has conditional independence. This assumes that each attribute has an independent effect on the eventual classification result.

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)}{P(x)}\prod_{i=1}^{d} P(x_i|c)$$

The next thing to take particular note of is whether or not a probability value of 0 will affect the subsequent estimations. Therefore, we set a very small value for the probability of an attribute that will not appear. This value is not 0. We call it as Laplacian correction.

__> further, we can classify the Naive Bayesian Classification into three steps.

- Stage 1: Preparatory Work Stage
  In this stage, we do the necessary preparation for Naive Bayesian classification. The main task is to determine the characteristics of the attributes according to the attributes' specific conditions and to perform the

appropriate division of each feature attribute. Then, we select a portion of the classification for sampling and use it to form the training sample set. The input for this stage is all the classifiable data, and the output is the characteristic attributes and training samples. This stage is the only stage in the whole of naive Bayesian classification where the quality of the work performed will have a heavy influence on the whole process. The characteristics of the attributes, classification of the characteristic attributes, and the quality of the training samples decide the quality of the classifiers.

- Stage 2: Classifier Training Stage
  In this stage, we generate a classifier. The main task is to calculate the frequency of appearance of each category in the training samples and the conditional probability estimation of each category for each feature attribute and to record the results. The inputs are feature attributes and the training sample; the outputs are classifiers. This stage is a mechanical one; it is based on the formula discussed above, and we can automatically calculate it completely by a program.
- Stage 3: Application Stage
  The task in this stage is to classify the classification items using the classifier. The inputs are the classifier and the classifiable items, and the output is the mapping between the classifiable items and the categories. This stage is also a mechanical and we can complete it through the program.

D. Artificial neural network

Artificial Neural Network is employed for classification process which basically works in two phases. In training phase the GLCM features, extracted from the known Mammogram, are used as inputs to train an ANN based breast cancer detection system. In testing, the trained ANN compares the extracted features with the features of the unknown sample of Mammogram and classifies the new mammogram image into benign and malignant [11].

E. KNN (k - nearest neighbor)

The k-nearest neighbor's algorithm is one of the machine learning algorithms. It is completely supported by the idea that "objects that are near each other will also have the same characteristics. Thus, if you know the characteristic features of one of the objects, you can also predict them for its nearest neighbor." This means that any new instance can be categorized by the 'k' neighbor majority votes, in which k is the positive odd integer [12]

Paper Name	Feature Extraction	Classification	Accuracy	Merits/demerits
Automatic Detection of Tumor Subtype in Mammograms Based On GLCM and DWT Features Using SVM [1]	GLCM and DWT(thresholding)	SVM	95%	False positive has been reduced up to 1 for every 100 images.
Application of Higher Order GLCM Features on Mammograms [2].	GLCM	SVM MLP(multilayer perceptron) KNN	-	Higher order GLCM feature can help improve the classification rate of micro calcification in breast tissue. Difficult to identify owing to its very small size.
A New Combined System Using ANN and Complex Wavelet Transform for Tissue Density Classification in Mammography Images [3].	Complex wavelet transform	ANN	80% to 94.79%	Great for finding tumors by applying density based classification algorithms but again sometimes it creates problems to exactly identify weather the detected tumor is caner or non-cancer tumor.
Application of Neural Network Based on SIFT Local Feature Extraction in Medical Image Classification [4].	Region of interest selection	Back propagation	96.57%	Application of Neural Network based classification with SIFT point features has great detection and classification accuracy but makes the system complex and time consuming.
New Intensity Based Features for Classification of Mammograms [5].	Haralick feature (ROI selection)	Bayesian KNN SGD Random forest	94.5%	Intensity based features detects and extracts bright key point features in the mammographic image, and mainly works good while accuracy is concerned.
Tumor Detection in Mammography Images Using Fuzzy C-means and GLCM Texture Features [6].	GLCM (ROI, median filter, thresholding,)	-	94.6%	Fuzzy C-means and GLCM together serves best for feature extraction and makes ease for post processing and ultimately increases accuracy and reduces classification time.

VI. CONCLUSION

This survey paper concludes that there are several techniques that deals with pre-processing, feature extraction and classification of diagnosing images that gave different accuracies. Most of the works used MIAS database which contain 322 mammographic images.

REFERENCES

[1] Automatic Detection of Tumor Subtype in Mammograms Based On GLCM and DWT Features Using SVM.
[2] Application of Higher Order GLCM Features on
[3] A New Combined System Using ANN and Complex Wavelet Transform for Tissue Density Classification in Mammography Images.
[4] Application of Neural Network Based on SIFT Local Feature Extraction in Medical Image Classification.
[5] New Intensity Based Features for Classification of Mammograms.
[6] Contrast Enhancement and Brightness Preservation using Global-Local Image Enhancement Techniques
[7] Tumor Detection in Mammography Images Using Fuzzy C-means and GLCM Texture Features.
[8] https://in.mathworks.com/help/images/morphological-dilation-and-erosion.html "
[9] Wavelet Feature based SVM and NAIVE BAYES Classification of Glaucomatous Images using PCA and Gabor Filter
[10] Breast Cancer Detection Using Neural Network Models
[11] Mammogram Analysis Using Feed-Forward Back Propagation and Cascade-Forward Back propagation Artificial Neural Network.
[12] A Hybrid Classification Algorithm Approach for Breast Cancer Diagnosis

