Mammogram Classification

Project Report

Submitted To: Mam Mehreen Sirshar
Submitted By: Rida Fatima (2017-BSE-058)
Faiza Mushtaq(2017-BSE-065)
Semester: VI B

Department of Software Engineering
Fatima Jinnah Women University, Rawalpindi

Contents

Mammogram Classification 2

Introduction 2

- How to take maximum clinical advantage from a digital mammogram 3
- Breast Cancer in Pakistan 3

Objective 3

Design 4

Implementation 4

Data Set 4

Steps in Implementation 5

1. Pre- Processing 6

Importance of Pre- Processing for Mammogram Images 6

Adaptive Mean Filtering 6

Working 6

2. Segmentation 7

Challenges in Mammogram image segmentation 7

A Gaussian based HMRF-EM approach towards Mammogram Images Segmentation 8

Working 9

3. Feature Extraction 10

What kind of features helpful in Mammogram Images? 10

Texture Features 10

Gray Level Co-Occurrence Matrix (GLCM) 11

4. Classification 12

Probabilistic Neural Network 12

Why PNN for Mammogram Classification? 13

Working 13

Implementation: Training and Testing 13

Conclusion 15

References 16

Mammogram Classification

Mammography is one of the most popular tools for early detection of breast cancer. Early diagnosis of this illness plays a key role in decreasing its mortality and improves its prognosis. Currently, mammography is considered as the standard examination for detection of breast cancer. However, the identification of breast abnormalities and the classification of masses on mammographic images are not trivial tasks for dense breasts.

Introduction

A Mammogram is an x-ray of the breast tissue which is designed to identify abnormalities. A mammogram can often find or detect breast cancer early, when it's small and even before a lump can be felt. This is when it's easiest to treat. Mammograms can be used for two purposes:

- A **screening mammogram** is used to look for signs of breast cancer in women who don't have any breast symptoms or problems. X-ray pictures of each breast are taken, typically from 2 different angles.
- Mammograms can also be used to look at a woman's breast if she has breast symptoms or if a change is seen on a screening mammogram. When used in this way, they are called **diagnostic mammograms**. They may include extra views (images) of the breast that aren't part of screening mammograms. Sometimes diagnostic mammograms are used to screen women who were treated for breast cancer in the past.

What Information can be extracted from these mammograms. The radiologist looks for different types of breast changes, such as small white spots called ***calcifications***, larger abnormal areas called ***masses***, and other suspicious areas that could be signs of cancer.

Calcification

Breasts calcification are small calcium deposits that develop in a woman's breast tissue. They are very common and are usually benign (noncancerous). In some instances, certain types of breast calcifications may suggest early breast cancer. There are two types of breast calcifications: macrocalcifications and microcalcifications.

Masses

A mass is an area of dense breast tissue with a shape and edges that make it look different than the rest of the breast tissue. With or without calcifications, it's another important change seen on a mammogram. Masses can be many things, including cysts (non-cancerous, fluid-filled sacs) and non-cancerous solid tumors (such as fibroadenomas), but they may also be a sign of cancer.

Cysts are fluid-filled sacs. Simple cysts (fluid-filled sacs with thin walls) are not cancer and don't need to be checked with a biopsy. If a mass is not a simple cyst, it's of more concern, so a biopsy might be needed to be sure it isn't cancer.

Solid masses can be more concerning, but most breast masses are not cancer.

A cyst and a solid mass can feel the same. They can also look the same on a mammogram. The size, shape, and margins (edges) of the mass can help to decide how likely it is to be cancer.

- **Challenges in finding Cancers on Mammograms**

Breast density is based on how fibrous and glandular tissues are distributed in your breast, compared to how much of your breast is made up of fatty tissue. Dense breasts are not abnormal, but they are linked to a higher risk of breast cancer. Dense breast tissue can also make it harder to find cancers on a mammogram.

- **Tumor**

A **tumor** is an abnormal lump or growth of cells. When the cells in the tumor are normal, it is benign. Something just went wrong, and they overgrew and produced a lump. When the cells are abnormal and can grow uncontrollably, they are cancerous cells, and the tumor is **malignant**. When they are harmless it is **Benign**.

- **How to take maximum clinical advantage from a digital mammogram**

It would be valuable to develop a computer aided method for mass/tumor classification based on extracted features from the Region of Interest (ROI) in mammograms. ROI must be segmented from the digital mammogram using the Segmentation techniques. Pattern recognition in image processing requires the extraction of features from regions of the image, and the processing of these features with a pattern recognition algorithm.

- **Breast Cancer in Pakistan**

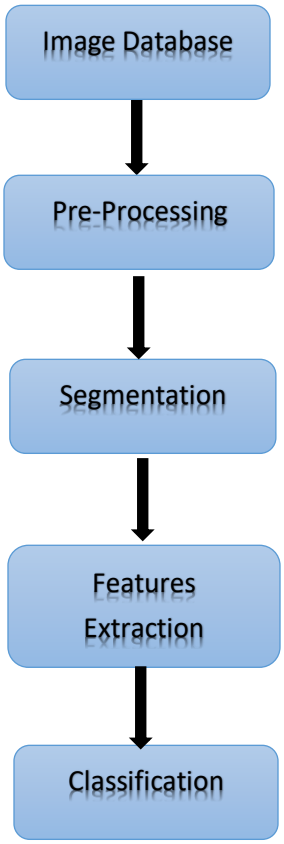
Pakistan alone has the highest rate of Breast Cancer than any other Asian country as approximately 90000 new cases are diagnosed every year out of which 40000 die. According to a research conducted approximately 1 out of every 9 women are likely to suffer from this disease at any point in their lives and about 77% of invasive breast cancer occurred in women above 50 years.Mortality in Breast cancer can be prevented in 1/3rd of women if routine mammography is done in women over 50 years, hence the longer a woman lives the lower is her risk of breast cancer therefore a 50 year old woman who has not had breast cancer has 11% chance of having it, whereas a 70 year old woman who has not had breast cancer has 7% chance of having it

Objective

The classification of benign and malignant patterns in digital mammograms is one of most important and significant processes during the diagnosis of breast cancer as it helps detecting the disease at its early stage which saves many lives. The main aim of this project is to propose a solution for tumor classification based on developed learning models. The solution can classify cancerous cells either **Normal, Malignant** or **Benign**.

Design

Our proposed design deal with four main sections **Preprocessing**, **Segmentation of ROI**, **Features Extraction** and **Classification**. By this classification process it can be found if the cancer cells will spread or not. For training and testing purpose we took dataset of mammogram images from MIAS dataset. With our proposed system when an a is given as input, it will go through the pre-processing, segmentation and then will be classified as its given class. The mammogram image can be seen and analyzed through every stage. The basic design of our system shows diagrammatically below:



Implementation

We have implemented the system in MATLAB 2015. All detailed description of every step is given below:

Data Set

Our dataset consists of distinct mammograms pictures taken from the Mammographic Image Analysis Society (MIAS) is an association of UK investigation organizations engaged with the acknowledgment of mammograms additionally has made a database of computerized mammograms. The database holds 322 digitized mammograms also been decreased to a 200-micron pixel frame and the images are 1024x1024. A community that consists of a sum of 321

mammographic images, 206 signify normal images, 51 malignant including 62 benign images. The models in the MIAS database are collected in the **Portable Gray Map(.pgm)** setup. Each image is 8-bit grey level range images among 256 various grey levels (0– 255).In this analysis utilized 400 benign, 40 normal also 40 malignant mammograms which are thick, fat and fatty glandular moreover have irregularities like surrounded asymmetry, masses,, ill-defined masses moreover all the mammograms are changed toward **.JPG** format.

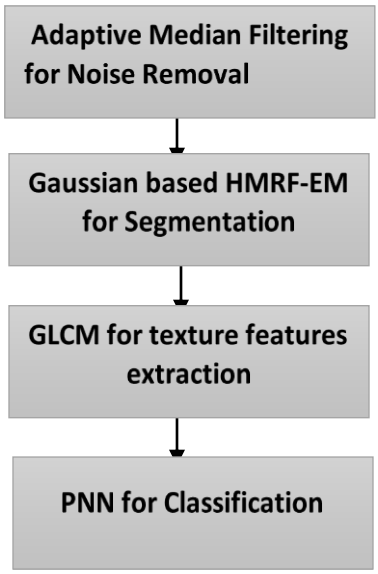
Steps in Implementation

We have implemented the steps shown in design phase. But we implemented them with following approaches.

- Pre-processing is done with **Adaptive Median Filter**.
- Segmentation is done with a novel approach as **Gaussian based Hidden Markov Random Filed with Expectation Maximization.**
- The difference between normal tissue and cancerous tissue is very small in some cases. So, the features of the tumor area in the image have key importance for automatic classification. Using only one feature or using a few features leads to poor classification results because of the small difference between the textures. So, for Feature extraction is done with **GLCM algorithm.**

- In training phase, the extracted features discovered are then passed to **Probabilistic Neural Network Classifier** for Classification.

A workflow of approaches used for classification



1. Pre- Processing

Pre-Processing is very important step to adjust and correct the mammogram image for further study and analysis. The goal of pre-processing is to enhance the signal to noise ratio between masses and normal breast tissues in mammograms by using techniques such as filtering. Different filters are available for image enhancement and noise reduction.

Importance of Pre- Processing for Mammogram Images

We know that mammogram can be classified as malignant or benign after diagnosis but the most important characteristics that tell whether it benign or malignant is the tumor are its shape (round, irregular) and margins (circumscribed, ill-defined). The tumor with round and regular shapes is usually benign or normal and with irregularities shows malignancy. The accuracy of this step determines the probability of success of the remaining steps such as segmentation, classification etc. Unknown noise, poor image contrast, inhomogeneity, weak boundaries and unrelated parts are usual traits of clinical images. The issues arise in this classification and in features extraction is mammogram images low contrast and noise.

- **Noise** maybe expressed as any change in the mammogram that do not relate to variations in the X-ray attenuation of the object being imaged. X-ray quantum noise is the significant type of noise which is denoted by a Poisson distribution. Such type of noise is generally introduced during mammograms image acquisition because of a smaller number of x-ray photons. It reduces the contrast of the mammogram image mainly which results in the low value of signal-to-noise ratio. Resultantly, the contrast of mammograms becomes low because of quantum noise whose detection is not only a complicated job but also challenging for the radiologists.
- **Low Contrast** Sometimes it becomes quite complicated to distinguish between benign and malignant masses because of low contrast. So, it is pertinent to apply any automatic technique to enhance the visual quality of mammographic image in order to fore front the malignancies, if exists.

Adaptive Mean Filtering

For Mammogram image processing we opted adaptive mean filtering because this filter will protect the edges of the picture and remove noise. It used to smooth the non-repulsive noise from two-dimensional signals without blurring edges and preserved images. This makes, it particularly suitable for enhancing mammogram images.

Working

The **Adaptive Median Filtering** has been applied wide as an advanced de-noising technique compared with traditional median filtering. The adaptive Median filter executes spatial processing to determine which pixels in a Mammogram image have been affected by noise.

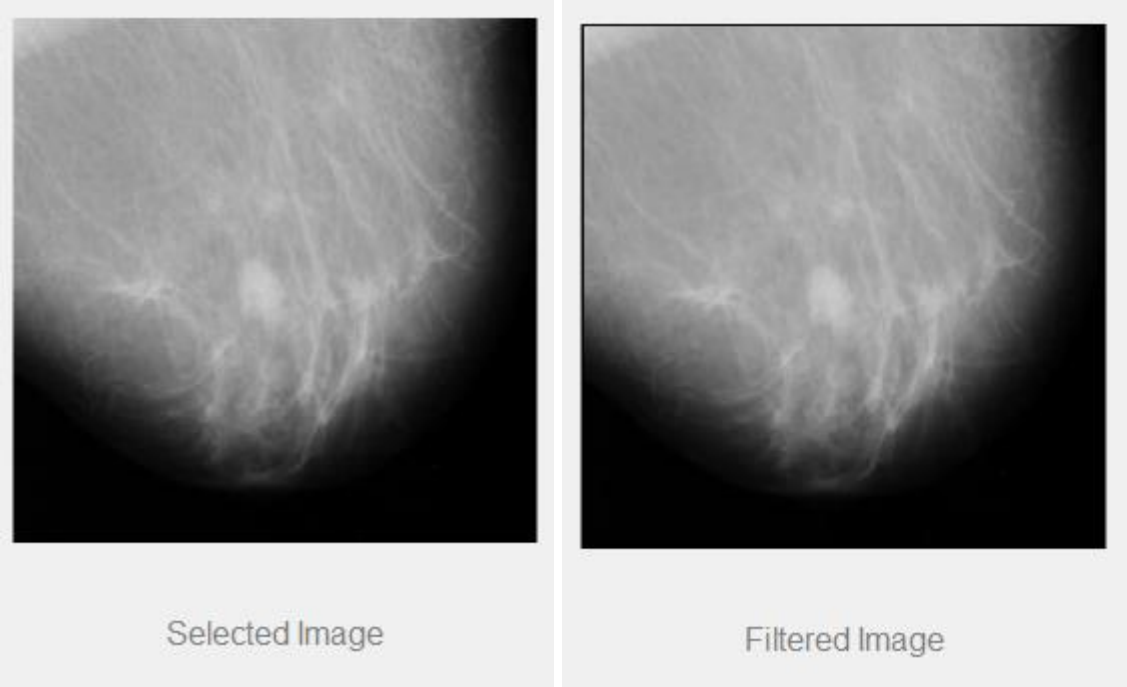
- The Adaptive Median Filter classifies pixels as noise by comparison each pixel in the MR image to its surrounding neighbor pixels. The size of the neighborhood window is adjustable, as well as the threshold for the comparison. A pixel that is different from most of its neighbors, as well as being not structurally aligned with those pixels to which it is

similar, is labeled as noisy pixel. These noisy pixels are then exchange by the median value of the pixels in the neighborhood that have passed the noise labeling test.

- Adaptive median filter changes the size of the neighborhood window through operation. But, in classic median filter; the neighborhood window is constant through the operation.

For that, the standard median filter does not perform well when the impulse noise density is high, while the adaptive median filter can better handle these noises. Also, the adaptive median filter preserves Mammogram image details such as edges and smooth non-impulsive noise, while the standard median filter does not.

Implementation



We have given an imput image labeled as slected image and applied an adaptive filter through which we obtained a better de-noised image for furthur processing steps.

2. Segmentation

Image segmentation divides an image into regions such that pixels within a region are homogeneous with similar properties based on some predefined condition. Mammogram image segmentation is one of the critical and challenging tasks. Segmentation methods are challenged for Mammogram images as the tumors area to be segmented have non-rigid anatomical structure, complex shape which varies in size and position among images.

Challenges in Mammogram image segmentation

A tumor is a pathology occupying a certain area, ranging from medium-grey to white shades in the mammogram. The smallest tumors visible in mammograms are approx. 0.5 cm in diameter.

The most significant features indicating whether the tumor is malignant or benign are its shape and the nature of its margins. Digital mammograms frequently contain strong noise that's why pre-processing also done in order to enhance the process of segmentation while cancerous tumors are of varied shapes and appearances. Furthermore, the contrast of suspicious-looking regions of mammograms is frequently low and heterogeneous, and the margins between masses are fuzzy and difficult to identify. All of this means that the segmentation of the lesions is an important and frequently very difficult task.

A Gaussian based HMRF-EM approach towards Mammogram Images Segmentation

Gaussian Mixture Model (GMM) has been widely applied in image segmentation. However, the pixels themselves are considered independent of each other, making the segmentation result sensitive to noise. To overcome this problem for the segmentation, process a proposed method where a mixture model using Markov Random Filed (MRF) that aims to incorporate spatial relationship among neighborhood pixels into the GMM. The proposed model has a simplified structure that allows the Expectation Maximization (EM) algorithm to be directly applied to the log-likelihood function to compute the optimum parameters of the mixture model.

In general, algorithms used for the detection and segmentation of masses as well as their further possible classification can be divided into two approaches: supervised segmentation and unsupervised segmentation.

- Supervised segmentation mainly includes model-based methods. Model-based methods use previously acquired (e.g., defined or learned) knowledge of objects and background regions that are being segmented. Previous knowledge is used to determine whether specific regions occur in the image or not. Supervised segmentation methods also include template matching approaches, in which the training set contains templates or patterns of objects that can be detected. Unfortunately, the main limitation of model or template-based methods is their reduced effectiveness in case of irregular masses with speculated margins that are difficult to distinguish.
- Unsupervised segmentation methods work by dividing the image into areas that are different or uniform about defined features, such as grey levels, their texture or colour.

How it is useful for Mammogram images?

The major goal of the work is to perform segmentation of Mammogram images in order to delineate tumor region. Gaussian Mixture Model (GMM) has been widely applied in image segmentation. However, the pixels themselves are considered independent of each other, making the segmentation result sensitive to noise. But according to our goal, we must extract the tumor region and as it has irregular shape, so we need to consider the neighboring pixel. The adjacent pixels of the same tissue have nearly similar properties and that is why HMRF can segment similar tissues based on this principle. Also segmented regions are not intermixed with neighboring segments as HMRF imposes strong spatial constraint based on neighborhood property which results into smooth segmentation.

Working

The following steps has been followed in segmentation of the ROI from the image

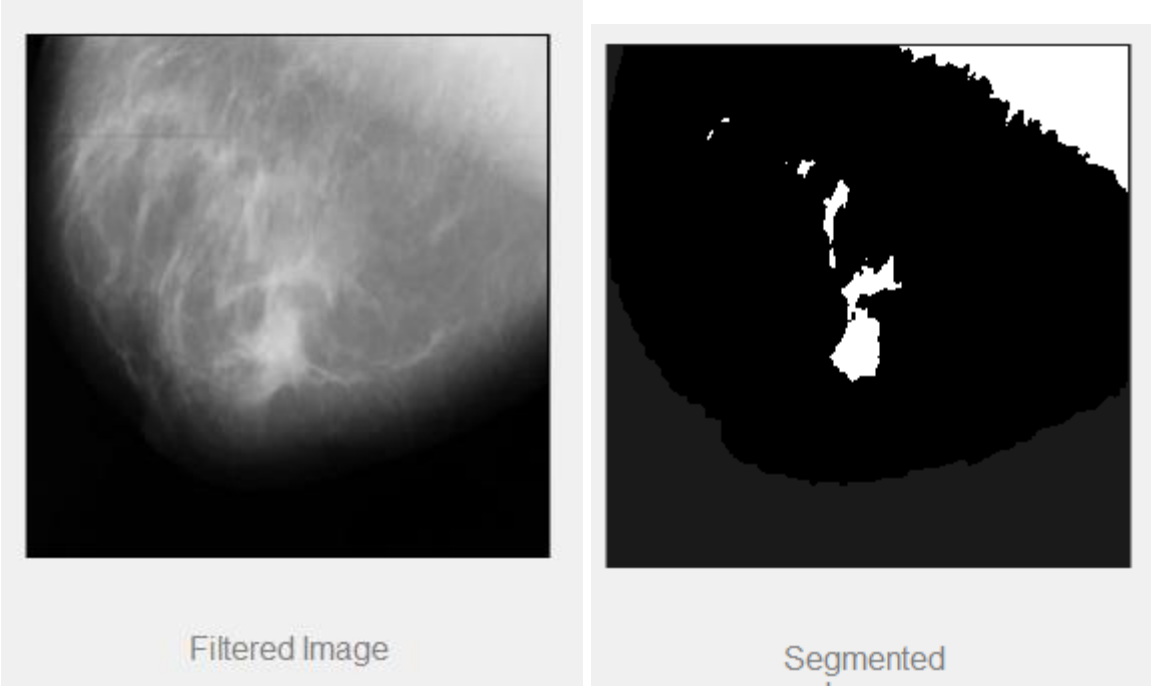
- GMM has been used to model intensity distribution of tissues and tumor in Mammogram images where its parameters are represented using mean, covariance and Gaussian components
- Each Gaussian component represents individual object. Estimated parameters of GMM are then given to HMRF-EM framework to predict class label.
- Parameters of HMRF model are mean and standard deviation of each GMM. MAP and EM algorithms have been used to learn HMRF parameters and class labels alternatively as both are dependent on each other. Parameters are learned by maximizing probability of class labels and by minimizing total posterior energy
- The basic idea of the EM algorithm is to begin with an initial model, to estimate a new model. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence condition is satisfied. In our experiment when no significant change in total energy observed or maximum EM iterations are reached, algorithm converges. EM algorithm has been used in our experiment to estimate the parameter set for HMRF model.
- As our approach is to optimize the initial conditions to be used by the **HMRF-EM.** When initial condition is very different than normal, mainly initial segmentation then EM may not give proper results. Hence, to optimize initial conditions k-means clustering is used to generate initial clusters and for approximation of intensity distribution in each segment Gaussian Mixture Model has been used instead of single Gaussian component.

Steps of Gaussian HMRF-EM the above described framework can be simplified as:

- Perform initial intensity-based clustering using k-means depending upon number of
- clusters specified.
- Estimate intensity distribution of Mammogram image using Gaussian Mixer Model with parameter set
- Estimate the class labels by HMRF-MAP estimation
- Update parameter set using EM algorithm
- Do MAP estimation, such that minimizes the total posterior energy
- Repeat step 4 till maximum EM iterations reached or there is no significant change in the
- Then we binarize the images showing dense tumor area and background.

The values for Gaussian components, Number of iterations, beta and k-clusters are set according to [1].

Implementation



3. Feature Extraction

The feature extraction and selection from an image plays a critical role in the performance of any classifier. Features, characteristics of the objects of interest, if selected carefully are representative of the maximum relevant information that the image has to offer for a complete characterization a lesion. Feature extraction methodologies analyze objects and images to extract the most prominent features that are representative of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent.

What kind of features helpful in Mammogram Images?

Because digital mammography images are specific, not all visual features can be used to correctly describe the relevant image patch. All classes of suspected tissue are different by their shape and tissue composition. Therefore, the most suitable visual feature descriptors for this kind of images are based on shape and texture. Textural features remain the best type of feature to be extracted from gray level images such as mammographic images, this is because these variables constitute texture which are: difference in gray level values; coarseness (scale of gray level differences); and directionality or regular pattern or lack of it. We can use different feature extraction methods and test them on variety of classifiers. That's why we use GLCM features extractor. It is widely used in much texture analysis application.

Texture Features

Texture is a repeated pattern of information or arrangement of the structure with regular intervals. In a general sense, texture refers to surface characteristics and appearance of an object given by the size, shape, density, arrangement, proportion of its elementary parts. Texture classification produces a classified output of the input image where each texture region is identified with the texture class it belongs. Texture can be defined as

- **Structural** texture is a set of primitive texels in some regular or repeated relationship.
- **Statistical** Texture is a quantitative measure of the arrangement of intensities in a region. This set of measurements is called a feature vector.

Gray Level Co-Occurrence Matrix (GLCM)

The GLCM is a tabulation of how often different combinations of gray levels co-occur in an image or image section. The GLCM contains a square matrix of size N × N, where N is the number of different gray levels in an image. An elements p (i, j, d, θ) of an image indicate the relative frequency, where i is the gray level of pixel p at location (x, y) and j is the gray level of a pixel located at d distance from p in the orientation θ. The texture information is used to classify the lump to be either benign or malignant.

The 22 features are extracted from the GLCM as follows:

1. Energy
2. Entropy
3. Dissimilarity
4. Contrast
5. Inverse difference
6. Correlation
7. Homogeneity
8. Autocorrelation
9. Cluster shade
10. Cluster prominence
11. Maximum probability
12. Sum of squares
13. Sum average
14. Sum variance
15. Sum entropy
16. Difference variance
17. Information variance of correlation
18. Difference entropy
19. Information measure of correlation
20. Maximal correlation coefficient
21. Inverse difference normalized
22. Inverse difference moment normalized.

Implementation

After Applying GLCM, we have got following features:

1x1 struct with 23 fields

Field	Value
autoc	1.4041e+04
contr	6.7910e+03
corrm	0.0399
corrp	0.0399
cprom	1.8302e+08
cshad	3.7171e+05
dissi	62.2573
energ	3.7723e-05
entro	10.5475
homom	0.0504
homop	0.0182
maxpr	5.4843e-05
sosvh	1.7929e+04
savgh	235.8069
svarh	6.0251e+04
senth	5.7999
dvarh	6.7910e+03
denth	5.1015
inf1h	-0.0414
inf2h	0.6036
indnc	0.8240
idmnc	0.9215

4. Classification

The basic of our proposed solution is to classify a tumor segmented from mammogram image as **Normal, Malignant** or **Benign**. The extracted features are used as input to the classifier.

Probabilistic Neural Network

A probabilistic neural network (PNN) is a feed forward neural network, which is widely used in classification and pattern recognition problems. A PNN that is probabilistic neural network is a feed forward neural network derived from the Bayesian network and a Statistical algorithm called Kernel Fisher discriminate analysis. The operations in PNN are organized into a multilayered feed forward network with four layers:

- Input layer
- Hidden layer
- Pattern layer/Summation layer
- Output layer

Probabilistic Neural Network is a network formulation of 'probability density estimation'. It is a model based on competitive learning with a winner takes all attitude and the core concept based on

Figure 1.Architecture of Probabilistic neural network

multivariate probability estimation. The development of PNN relies on the Parzen window concept of multivariate probabilities. The PNN is a classifier version, which combines the Bayes' strategy for decision-making with a nonparametric estimator for obtaining the probability density function. The input layer simply distributes the input to the neurons in the pattern layer and does not perform any computation.

Why PNN for Mammogram Classification?

A PNN is predominantly a classifier since it can map any input pattern to several classifications. Among the main advantages that discriminate PNN is: Fast training process, an inherently parallel structure, guaranteed to converge to an optimal classifier as the size of the representative training set increases and training samples can be added or removed without extensive retraining. Accordingly, a PNN learns more quickly than many neural networks model and have had success on a variety of applications. Based on these facts and advantages, PNN can be viewed as a supervised neural network that can use it in system classification and pattern recognition.

Working

- The input layer neurons distribute input measurements to all the neurons in the pattern layer.
- The second layer has the Gaussian kernel function formed using the given set of data points.
- The third layer performs an average operation of the outputs for each review class.
- The fourth layer performs a vote, selecting the largest value and class label is then determined.

The features extracted from GLCM will be stored and first converted to feature vector and then given to PNN. In this way a new PNN structure is formed in MATLAB.

Implementation: Training and Testing

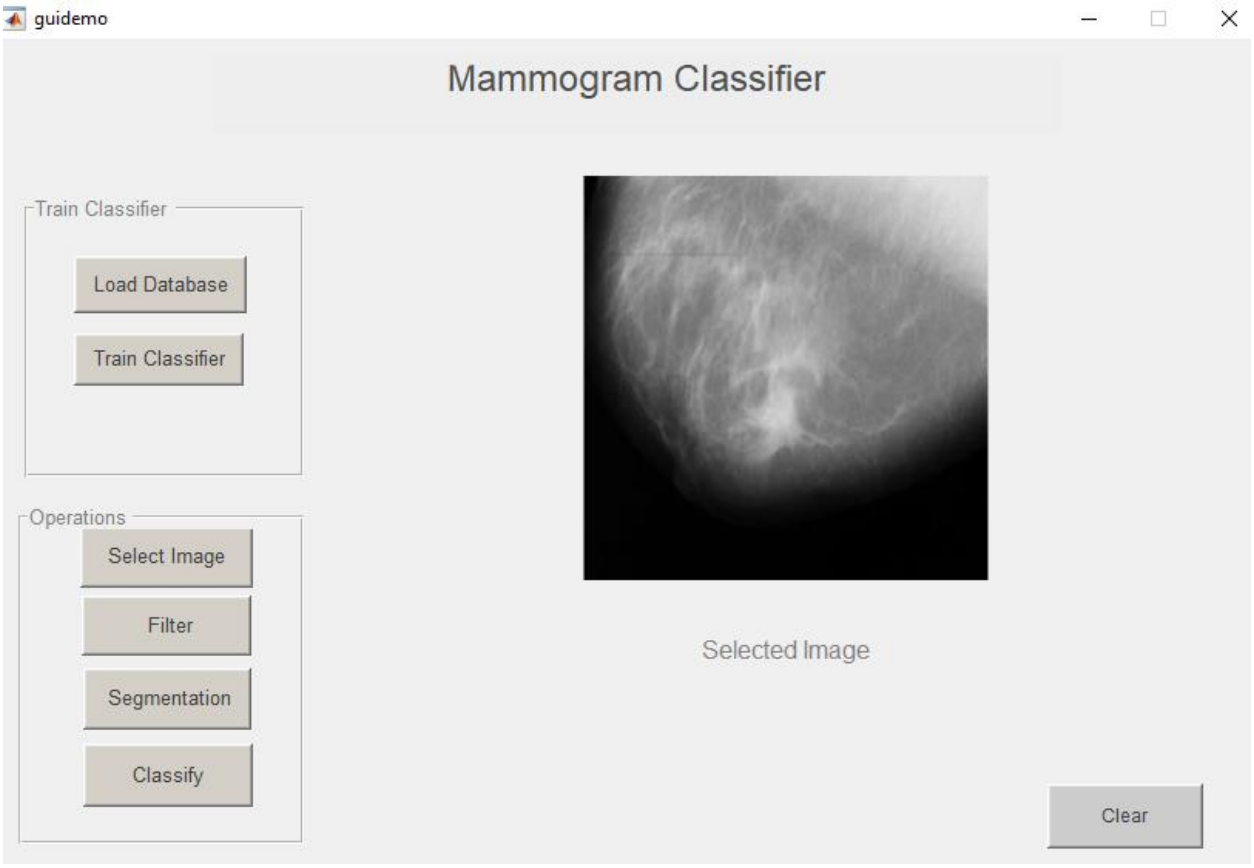
Classification is further divided into

- Training phase
- Testing phase

In **training phase**, the database of 120 images including Normal, Malignant and Benign is loaded and used to train PNN classifier.

In **testing phase** the PNN is tested with the rest of database and shown a good accuracy, further input image can be classified into either of the class.

Design



(a)

(b) Input image in (a) classified as Malignant

In the above design option **Load Database** load the database of mammogram images and done its processing till segmentation.

Then **Train Classifier** option train the loaded database and train the **PNN** classifier.

The further options can be use for testing with the left dataset, Image can be given and can perform all operations filtration, Segmentation and then can be tested as to classify the class as either **Normal, Benign** or **Malignant.**

Conclusion

We have proposed solution for classification of tumor in mammogram images. We done it with help of image dataset from MIAS dataset. But it has one limitation, the size of images is small. The chosen classifier PNN was useful for it as has one big advantage over other classifiers it has faster training and give better results when a new image is given as input. The pre-processing and segmentation and features extraction techniques came out to be helpful because they are the input for classifier and classifier has given a better performance.

As it has discussed breast cancer needs to detect early and should treat with state-of-art treatment for cancer, for early detection mammogram is useful tool. Breast cancer that's found early, when it's small and has not spread, is easier to treat successfully. But more better strategies need to be designed and work on for preventing deaths from breast cancer.

References

1. A. Singh. June 2017.”Mass Classification of Mammogram Images using Selected Textural Features with PNN Classifier”, Volume 5, issue 4.
2. N.Safdarain. October 2019, “Detection and Classification of Breast Cancer in Mammography Images Using Pattern Recognition Methods “. Volume 3, Issue 4.