Chapter Four

ALGORITHM METHODOLOGY AND IMPLEMENTATION

This chapter expounds the main methodology and materials exploited for the accomplishment of the project.

4.1 System Overview

In the course of achieving the proposed objectives of this research, the algorithm implementation mainly comprises of the capabilities of MATLAB Toolboxes. A computer system with a graphical user interface was created for the purpose of performing pattern recognition and classification on raw mammographic images. The latter part was achieved using (centers and radii of masses) associated within the information provided by mini-MIAS (Mammographic Image Analysis Society) Mammographic database [46]. The system as will be detailed later, concerns with loading the raw mammographic images that contain malignant (spiculated, well-defined, ill-defined) masses and benign masses that were taken from the mini-MIAS and were passed into a segmentation stage, known as the watershed transform. Then calculate a dataset of features for the process of feature selection and extraction. The resulting features were then used to train the ANFIS system using backpropagation algorithm in order to discriminate the pattern of the masses into two classes; benign and malignant.

4.2 Proposed Flowchart Algorithm

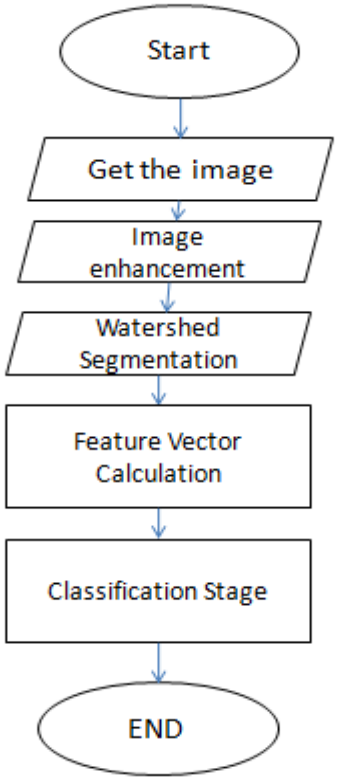

Fig. 4.1 illustrates the proposed algorithm flow chart

4.3 Data Source

The mammogram images used in this experiment were taken from the mini mammography database of MIAS. In this database, the original MIAS database was digitized at 50 micron pixel edge and has been reduced to 200 micron pixel edge and clipped or padded so that every image is 1024 x 1024 pixels.

All images are held as 8-bit gray level scale images with 256 different gray levels (0-255) and physically in portable gray map (pgm) format. The database provides information about ground truth on breast density. The three categories are:

• Fatty (F)

• Glandular (G)

• Dense (D)

There are several different categories of abnormalities that could be present in the breast tissue, if existent. Some of which are the following;

• CALC - Calcification

• CIRC - Well-defined/circumscribed masses

• SPIC - Spiculated masses

• MISC - Other, ill-defined masses

• ARCH - Architectural distortion

• ASYM - Asymmetry

• NORM – Normal

With the severity of abnormality classified as either benign or malignant. The existing data in the collection also includes image-coordinates of the centre of abnormality and the approximate radius (in pixels) of a circle enclosing the abnormality. The algorithm implementation concerns with the creation of pattern recognition and classification system as detailed below using ANFIS. The MATLAB GUI (graphical user interface) was used to create and represent the system.

4.4 Image Enhancement

This stage is developed for the purpose of image processing and enhancement. The process of enhancing the quality of images assists in producing reliable representation of breast structures and thus, renders an effective means for proper diagnosis and prognosis for radiologists. Once the raw mammographic image gets loaded onto the GUI figure, the user can press the enhancement button which allows the system to perform 2-D adaptive wiener filtering of the image followed by contrast limited adaptive histogram equalization. The resulting output is the enhanced image.

4.5 Pattern Recognition and Classification

Disagreement or inconsistencies in mammographic interpretation motivates utilizing computerized pattern recognition algorithms to aid the assessment of radiographic features [47].

This study is concerned with the detection of masses and their classification into either, benign or malignant tissue. Therefore, a total of 51 mammograms comprising 18 ill-defined, spiculated, circumscribed masses and 33 benign cases were taken into consideration.

4.5.1 Image Segmentation

Image segmentation is the process of isolating objects in the image from the background, i.e., partitioning the image into disjoint regions, such that each region is homogeneous with respect to some property, such as grey value or texture [48].

4.5.1.1 The Watershed Transform

The watershed transform can be classified as a region-based segmentation approach. The intuitive idea underlying this method comes from geography: it is that of a landscape or topographic relief which is flooded by water, watersheds being the divide lines of the domains of attraction of rain falling over the region [49].

An alternative approach is to imagine the landscape being immersed in a lake, with holes pierced in local minima. Basins (also called `catchment basins') will fill up with water starting at these local minima, and, at points where water coming from different basins would meet, dams are built. When the water level has reached the highest peak in the landscape, the process is stopped. As a result, the landscape is partitioned into regions or basins separated by dams, called watershed lines or simply watersheds. When simulating this process for image segmentation, two approaches may be used: either one first finds basins, then watersheds by taking a set complement; or one computes a complete partition of the image into basins, and subsequently finds the watersheds by boundary detection. To be more explicit, we will use the expression `watershed transform' to denote a labeling of the image, such that all points of a given catchment basin have the same unique label, and a special label, distinct from all the labels of the catchment basins, is assigned to all points of the watersheds [50]

Assume that the image **F** is an element of the space C(D) of real twice continuously differentiable functions on a connected domain D with only isolated critical points Then the topographical distance between points p and q in D is defined by.

$$T_f(p,q) = inf \int_\gamma \|\nabla f(\gamma(s))\| \, ds \qquad (4.1)$$

Where the infimum is over all paths (smooth curves) γ inside D with $\gamma(0) = p$, $\gamma(1) = q$.

Let f belongs to C(D) have minima $\{m\}_{k \text{ belongs to } I}$, for some index set I. The catchment basin $CB(m_i)$ of a minimum mi is defined as the set of points x belongs to D which are topographically closer to m_i than to any other regional minimum m_j :

$$CB(m_i) = \{x \in D \big| \forall_j \in I\{i\}: f(m_i) + T_f(x, m_i) < f(m_j) + T_f(x, m_j)\} \qquad (4.2)$$

The watershed of f is the set of points which do not belong to any catchment basin:

$$W_{shed}(f) = D \cap (\bigcup_{i \in I} CB(m_i))^c \quad (4.3)$$

So the watershed transform of f assigns labels to the points of D, such that

a) Different catchment basins are uniquely labeled, and
b) A special label W is assigned to all points of the watershed of f.

4.5.2 Feature Selection

Feature selection methodologies analyze objects and images to allow the extraction of the most prominent features that are representative of the various classes of objects. The features are then used as inputs to classifiers that assign them to the class that they represent and are therefore, the most deciding factor in the precision of classification. Feature selection prevails to be a pervasive concern in pattern recognition and classification. The main issue addressed by feature selection is to choose the optimum subset of features that leads to the smallest classification error and that can achieve the best performance in terms of accuracy and computation time. It is desirable and therefore, imperative that only a few features be selected and extracted for the classification stage. Large number of features would increase computational needs and degrade the classifier's performance [27]. The standardized (300x300) window ROIs extracted from the mammographic images was used for the feature selection process. Features selected according to their discrimination power which is measured using the MATLAB built-in function **ttest**. Features that demonstrated high discrimination power between malignant and benign tissue were selected.

The following section details the information on the extracted features obtained from the feature selection stage.

4.5.2.1 Extracted features

The following selected features formed the input dataset for classifier after being processed from the ROI.

1 Entropy

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy, ***h*** can also be used to describe the distribution variation in a region. Overall entropy of the image can be calculated as:

$$h = -\sum_{k=0}^{L-1} Pr_k \, (\log_2 Pr_k) \quad (4.4)$$

Where, Pr_k is the probability of the k-th grey level, which can be calculated as. **Zk/MxN** is the total number of pixels with the k-th grey level and ***L*** is the total number of grey levels.

2 Sum Entropy

Sum entropy, otherwise known as entropy of pair-sums is similar to the concept of entropy. The new addition in this case is the definition of a new probability function. The function represents a probability of a particular pair-sum of pixel values denoted as $\mathbf{P_{x+y}(i)}$. Substituting **Prk** by the new probability function in equation 4.4, we obtain the sum entropy.

$$-\sum_{i=2}^{2N_g} [\boldsymbol{P}_{x+y}(\boldsymbol{i}) \log[\boldsymbol{P}_{x+y}(\boldsymbol{i})]] \quad (4.5)$$

3 Difference Variance

Difference variance, otherwise known as variance of pair-differences, has a probability function for all possible pair-differences initially defined. This is denoted by $\mathbf{P_{x+y}(i)}$. Defining the means of pair-differences by ***f'***

$$f' = \sum_{i=0}^{N_g-1}[iP_{x-y}(i)] \quad (4.6)$$

The difference variance is finally calculated as

$$P_{x-y} = \sum_{i=0}^{N_g-1}[(i-f')^2 P_{x-y}(i)] \quad (4.7)$$

4 Energy

Energy, also known as uniformity or the angular second moment returns the sum of squared elements in the Grey Level Co-Occurrence Matrix (GLCM). The range of energy is [0 1] and is 1 for a constant image. The formula for finding energy is given in Equation 4.8.

$$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\{P(i,j)\}^2 \quad (4.8)$$

5 Contrast

Contrast returns a measure of the intensity contrast between a pixel and its neighbor over the whole image. It is a measure of the amount of local variations present. The contrast is weighted by the square of the gray level differences. When ***i*** and ***j*** are equal, the cell is on the diagonal. These values represent pixels entirely similar to their neighbor, so they are given a weight of 0, and this renders a constant image. The range of contrast is [0 (size (GLCM, 1)-1)2]. In contrast calculation, the weights increase exponentially (0, 1, 4, 9 etc.) as ***(i - j)*** increases. It is calculated by using Equation 4.9.

$$\sum_{n=0}^{N_g-1} n^2 \left\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\right\} \quad (4.9)$$

With ***p(i, j)*** denoting the normalized co-occurrence matrix and ***Ng*** the number of discrete gray levels of the images.

Both contrast and dissimilarity have weights related to the distance ***(i-j)*** from the GLCM diagonal. Therefore, they are a function of amplitude differences ***(i-j),*** rather than amplitudes (***i*** and ***j***) which makes them insensitive to the mean value of the amplitude and hence, are a measure of texture independent of how strong or weak the average amplitude may be [51] [52] [53].

4.5.3 Classification stage

The high degree of non-linearity employed in the discrimination of the input data with regard to diagnosis prediction suggests that automatic diagnosis systems should implement robust pattern recognition models of non-linear and highly adaptive architecture such as adaptive-neuro fuzzy inference system.

Classification problems aim to identify the characteristic that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave.

An ANFIS based classifier is presented as a diagnostic tool to aid physicians in the classification of breast cancer. ANFIS using a strategy of hybrid approach of adaptive neuro-fuzzy inference system that compose of two intelligent approaches, and it will achieve good reasoning in quality and quantity. In other words have fuzzy reasoning and network calculation. The objective of classification is to classify the breast masses into either malignant or benign; the features vector was applied as the input to the ANFIS classifier. The ANFIS network has a total of 243 fuzzy rules and one output, the classification by ANFIS was performed using MATLAB.

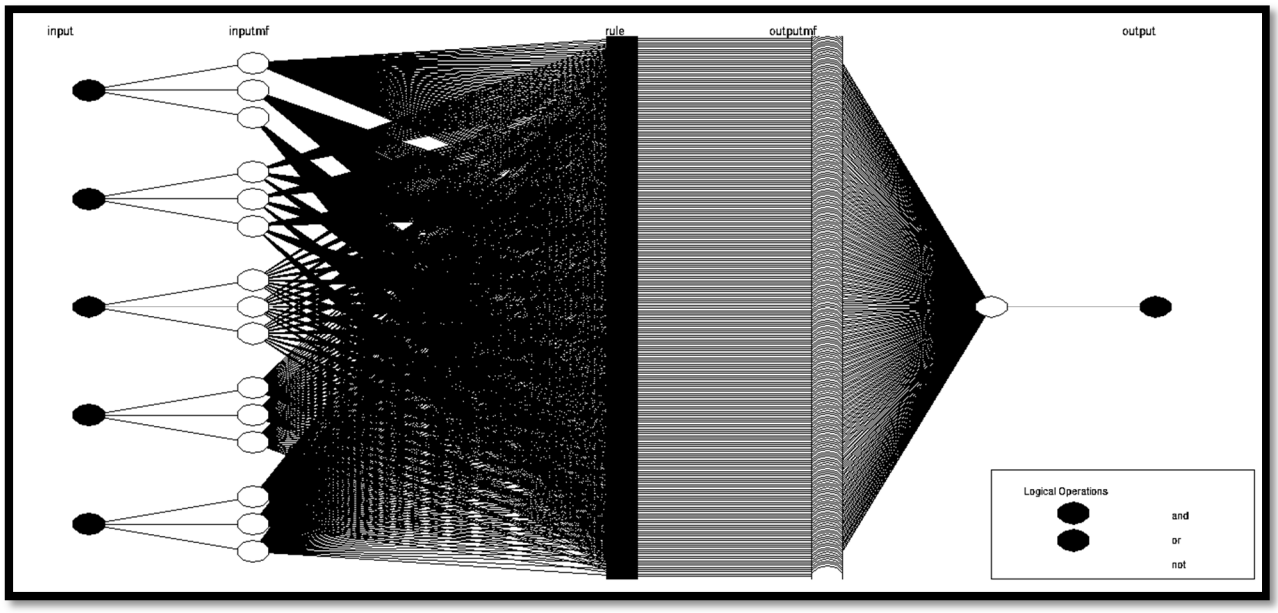

Fig. 4.2 shows the ANFIS structure that used in the study

The basic structure of the type of fuzzy inference system seen thus far is a model that maps input characteristics to input membership functions, input membership function to rules, rules to a set of output characteristics, output characteristics to output membership functions, and the output membership function to a single-valued output or a decision associated with the output. Considered only fixed membership functions that were chosen arbitrarily and applied fuzzy inference to only modeling systems whose rule structure is essentially predetermined by the user's interpretation of the characteristics of the variables in the model.

In this dissertation ANFIS Editor GUI in the toolbox is used. These tools apply fuzzy inference techniques to data modeling. As have seen from the other fuzzy inference GUIs, the shape of the membership functions depends on parameters, and changing these parameters change the shape of the membership function. Instead of just looking at the data to choose the membership function parameters, choose membership function parameters automatically using these Fuzzy Logic Toolbox applications.

The neuro-adaptive learning method works similarly to that of neural networks. The modeling approach used by ANFIS is similar to many system identification techniques. First, you hypothesize a parameterized model structure (relating inputs to membership functions to rules to outputs to membership functions, and so on). Next, collect input/output data in a form that will be usable by ANFIS for training and then use ANFIS to train the FIS model to emulate the

training data presented to it by modifying the membership function parameters according to a chosen error criterion.

In this research, a sugeno type ANFIS having **five inputs** with **three Two-sided** Gaussian membership functions for each input, **one output**, and **243 rules** was used.