Automated Image Analysis Techniques for Screening of Mammography Images

Enda Molloy

B.E. Electronic Engineering Project Report

EE4BN

March 2009

Abstract

This project proposes the use of automated image analysis techniques for screening of mammography images to help radiographers in the diagnosis of breast cancer. This will be achieved by performing analysis techniques on a database of mammography images and identifying which cases contain suspicious masses which may be indicative of the presence of breast cancer, and therefore warrant further investigation by a physician.

In this system the images are pre-processed, features representing regions of interest are extracted and finally, classified using an artificial neural network into one of three classes - benign, malignant or normal.

The system also provides a facility that would allow a doctor to upload and view images via a web browser, so that the images could be accessed remotely.

The report explains the physiological and the major technical areas relevant to the project. Tests are conducted using the mini MIAS database, available online. Analyses of the techniques selected and implemented in the system are discussed as well as issues that arose during the course of the project. Results indicate that the best performance is obtained using a multilayer perceptron neural network architecture with intensity features as input vectors.

Declaration of Originality

I declare that this thesis is my original work except where stated.

Signed: _____ Date: _____

Enda Molloy

Acknowledgements

I wish to thank all the members of the Electronic Engineering Department and a special thanks to my supervisor, Dr. Edward Jones for his guidance, advice and for always making himself available to discuss the project.

I would also like to thank my family and fellow classmates, for all their support and encouragement throughout the duration of this project.

Table of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction

1.1 Project Background

Breast cancer is the second most common cause of cancer in Irish women after non-melanoma skin cancer, and the most common cause of cancer death in Irish women. There are nearly 1,900 new cases of breast cancer in Ireland every year [1].

Breast cancer is defined as an abnormal growth of cells in the breast that multiply uncontrollably, new cells grow when they are not needed and old cells don't die when they should. It is widely believed that the main factors which cause breast cancer are hormonal and genetic. Breast cancer can occur in both men and women, however for every one hundred females with breast cancer only one male will contract the disease, as a result only female breast cancer is discussed in this project.

Masses are quite subtle, and often occur in the dense areas of the breast tissue, they have smoother boundaries than microcalcifications, and have many shapes such as circumscribed, speculated, lobulated or ill-defined. The circumscribed ones usually have a distinct boundaries, 2–30 mm in diameters, and are high-density radiopaque; the speculated ones have rough, star-shaped boundaries; and the lobulated ones have irregular shapes [2]. Masses/Tumors can be either benign or malignant.

A **benign tumor** is not cancerous because:

1. They do not grow in an aggressive manner and if removed normally don't grow back.
2. Benign tumors do not invade healthy surrounding tissue.
3. Benign tumors do not metasize i.e. spread to other parts of the body.

A **malignant tumor** is cancerous because:

1. If the tumor is removed it can still grow back.
2. Malignant tumor cells can invade and damage surrounding tissue.
3. Malignant tumor cells can metasize.

1.2 Project Specification

The system can be divided into two main stages, feature extraction and classification. Feature extraction involves analysing the images using image processing techniques and deriving features representing abnormalities from them. Classification on the other hand refers to further analysis of the images after which a decision is made as to whether regions of interest of stage one, are deemed to be suspicious or not.

At the front end of the system will be the image processing stage, which will be developed using previously established techniques for image analysis. The system should also include additional image processing techniques to reduce noise which can sometimes appear on mammograms. Feature extraction techniques should be investigated and suitable options implemented. Once features have been extracted, different classification architectures (based on techniques that have been used in literature) should be examined and a suitable architecture chosen. From this a basic system can be built and tested for the screening of mammograms. Web server functionality may also be added to the system to enable a radiologist to access the data on their server from any location by means of a web browser (e.g. in an out-patient situation, or if a "local" therapist wants to get a second opinion from a "remote" colleague). A second front end processor could also be chosen (using the same classification architecture) and a comparative study carried out between this front end processor and the first front end processor. Both approaches would be evaluated from the point of view of performance (percentage of data correctly screened) and complexity (in terms of the computation required).

1.3 System Overview

The basic system structure is shown in Figure 1.1. A digitised mammogram is initially fed into the system where it's contrast is enhanced using CLAHE (Contrast Limited Adaptive Histogram Equalisation), any areas representing regions of interest are cropped from the image and further analysis is carried out on these. Intensity and textural features are calculated and extracted from the cropped ROI's, by using first order statistics features and wavelet coefficients respectively. These features are then passed through a trained, multilayer perceptron neural network classifier which determines which category the region of interest falls into to – benign, malignant or normal.

The system also incorporates a basic online database system which allows authorised medical personnel to view images and screening results from a web browser.

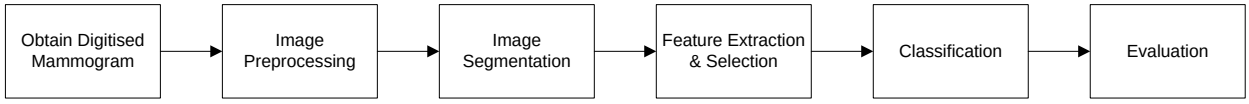


Figure 1.1 System Overview

1.4 Report Layout

The layout of the remainder of the report is as follows:

Chapter 2: Background information, discusses some of the relevant background information important for this project.

Chapter 3: Front-End Processing, deals with the different techniques used in the front-end processing of the images and how they were implemented.

Chapter 4: Classification, this chapter details background information on the classification architecture used in the project and also how it was implemented.

Chapter 5: Classifier Results, in this chapter the performance of the two front-end processors as inputs to the neural network are compared.

Chapter 6: Database Web Application, this chapter discusses how the application was designed and created.

Chapter 7: Conclusion, this chapter gives a brief summary of the project outcomes and also discusses any further developments that could be carried out.

Chapter 2: Background Information

In this Chapter some of the relevant background information with respect to breasts and breast cancer is discussed. This includes a brief description of some of the present screening techniques. The source of mammograms used in this project is also discussed.

2.1 Breast Structure

The breast is held in place by the chest muscles that cover the ribs. Each breast is made up of fifteen to twenty lobes with lobules containing smaller lobules. Tiny glands within these lobules are what produce milk, which flows from the lobules through thin tubes called ducts to the nipple. The nipple is in the center of a dark area of skin called the areola. Fat fills the spaces between the lobules and ducts. The breasts also contain lymph vessels which lead to lymph nodes. The lymph nodes trap bacteria, cancer cells, or other harmful substances. Most breast cancers begin in the ducts, some in the lobules and the rest in other tissues [3]. Figure 2.1 below depicts the internal structure of the breast.

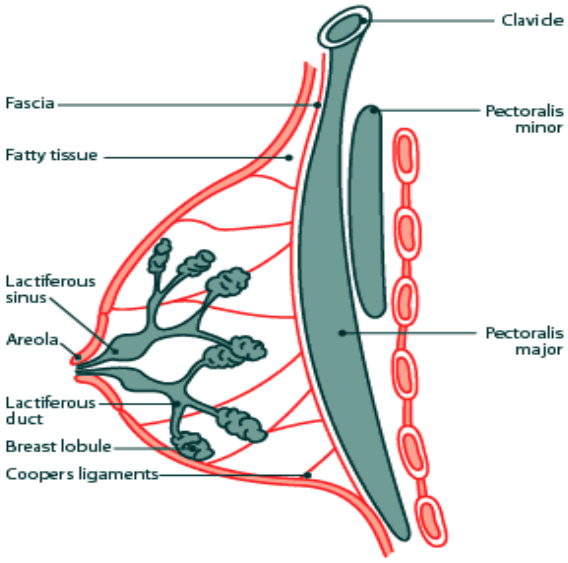


Figure 2.1 Internal breast structure [4].

2.2 Breast Cancer Symptoms

Breast cancer can have a number of different symptoms, the most common of which is a lump in the breast although some 90% of breast lumps are benign. Other symptoms include a change in how the breast or nipple feels, a change in the shape or size of the breast, nipple turned inwards towards the breast, change of skin colour or a fluid discharge through the nipple. These symptoms are shown graphically in Figure 2.2 below.

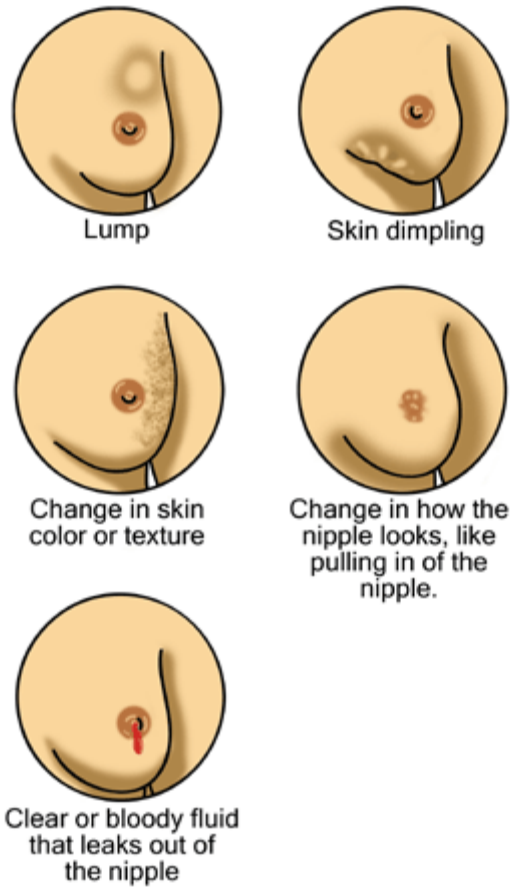


Figure 2.2 Breast cancer symptoms [5]

2.3 Mammography

Mammography is at present the most reliable and widespread method for early detection of breast cancer. It is estimated that mammography can show changes in the breast for up to two years before a patient or clinician can feel them. Mammography works by using low dose x-ray to examine breasts, this is performed by a special type of x-ray machine which compresses the breast. Using x-rays as an imaging tool involves exposing the body to a small dose of ionizing radiation to produce the images. The x-rays are absorbed at different rates as they pass through various types of tissue. It is these variations in absorption rates that provide the details of the internal breast structure. Currently both film and digital mammography are in clinical use today. Film mammography is the traditional method whereby the x-ray is made on high resolution, high contrast film. On the other hand, Digital mammography, or full-field digital mammography (FFDM), is coming to the fore with continuing developments and advances in technology. Digital Mammography is a mammography system where the film is replaced with solid-state detectors, similar to those in digital cameras, which convert x-rays into electrical signals. These electrical signals are sent to a computer which interprets them and displays the image. One of the main advantages of using digitised mammograms is that it allows for the use of a computer-aided diagnostic (CAD) tool.

2.4 CAD

Computer-Aided Diagnosis (CAD) can be defined as a diagnosis that is made by a radiologist who uses the output from a computerised analysis of medical images as a ‘second opinion’ in detecting lesions and in making diagnostic decisions. The final diagnosis is made by the radiologist [6].

Although mammography is most reliable and widespread method for early detection of breast cancer, it is estimated that some 10-30% of all malignant cases are misdiagnosed, two thirds of which are retrospectively evident on the mammogram [7]. This shows how difficult it is for radiologists to continuously scrutinize large numbers of images with only a small number actually showing abnormalities. As well as this, the high percentage of false positives, which is 80-90% of surgical biopsies performed in women, necessitates the need for a cheaper and less invasive detection method, while also removing human error. Computer aided diagnosis could be used to eliminate or reduce these errors and this is what this project has attempted to do.

2.5 MIAS Database

The source of the mammograms used in this project is the MIAS database [8].

The Mammography Image Analysis Society (MIAS) is an organisation of UK research groups interested in the understanding of mammograms who have produced a digital mammography database for research purposes.

The X-ray films in the database have been carefully selected from the United Kingdom National Breast Screening Programme and digitised with a Joyce-Lobel scanning microdensitometer to a resolution of 200 μm × 200 μm, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. Every image is 1024 X 1024 pixels in size. The database contains left and right breast images for 161 patients. Its quantity consists of 322 images, which belong to three classes such as normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant (cancerous) images. It also includes expert radiologist's markings on the locations of any abnormalities that may be present. For each image, experienced radiologists have given the type, location, scale, and other useful information of them.

The database includes a readme file, which details the (i) type of abnormality, whether it is a radial lesion, circumscribed mass, or microcalcification, (ii) the class of the abnormality i.e. benign or malignant and (iii) the location of the center of the abnormality and it's diameter. A typical mammogram from the MIAS database is shown in Figure.2.3.

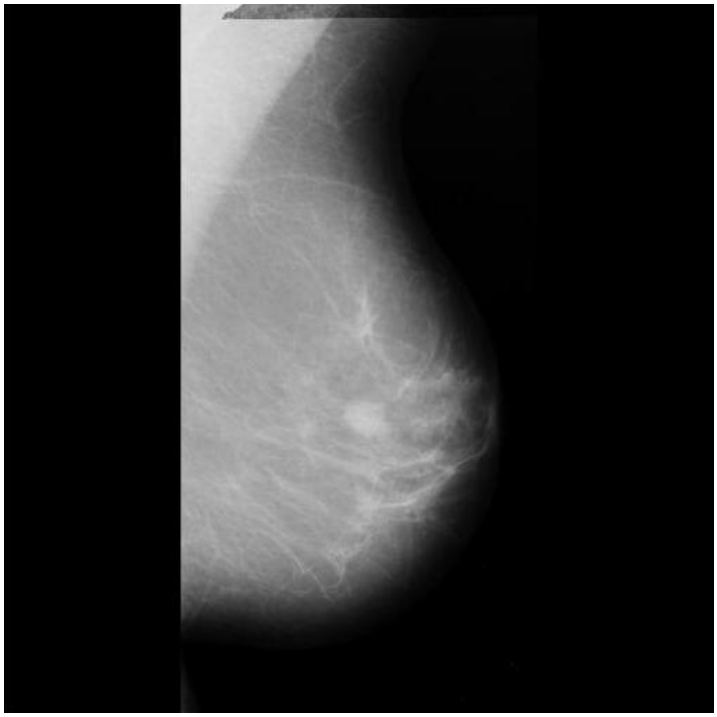


Figure 2.3 Mammogram from MIAS database

Chapter 3: Front – End Processing

This chapter describes techniques investigated and used in the front-end processing. Front-end processing consists of image enhancement, image segmentation, image de-noising as well as feature extraction and selection. In addition the process of choosing certain techniques is discussed along with reasons for emitting others. Also discussed are the implementation decisions made and problems encountered.

3.*1 MATLAB*

The majority of the system is designed using MATLAB [9]. MATLAB is described by Mathworks, the software creator, as a high-level computing language with technical applications and environment for algorithm development, data visualization, data analysis, and numeric computation. By using MATLAB for these areas of programming the product as a whole, language and environment, can be used to great effect as extensive specialised libraries of usage definable functions are available to the user. These are implemental by simply naming and passing parameters to their function of choice.

MATLAB can be used in a large spectrum of applications, including but not limited to signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology. The functions which allow users to control and build their algorithms with such ease are stored in toolboxes. These are collections of MATLAB functions which relate to a particular application of area. For example, the image processing toolbox is used in this project. The large range of toolboxes demonstrates the extent and range of

situations to which the MATLAB environment can be applied to solve particular problems in its application areas.

MATLAB also provides a number of features for documenting and sharing work. MATLAB code is compatible with other languages and applications and contains specific functions for integrating MATLAB based algorithms with external applications and languages, such as C, C++, FORTRAN, Java, COM, and Microsoft Excel [9]. This allows developers to use solutions developed in MATLAB on existing legacy systems without difficulty. The Image Processing Toolbox, Wavelet Toolbox and the Neural Network Toolbox were used extensively in this project.

3.2 Image Enhancement

The principle object of image enhancement is to process an image so that the resulting image is more suitable than the original one for a given application. Image enhancement techniques fall into two main categories, namely spatial domain methods and frequency domain methods. Spatial domain methods work on the principle of directly manipulating image pixels whereas frequency domain methods are based on altering the Fourier transform of the image. As Mammograms are black and white x-rays of a compressed breast, they are low contrast images, so it is important to pre-process the images. The reason images need to be pre-processed is so that intensity differences between objects and background can be increased and to enable clearer views of breast structures. The aim is to enhance the textures and features of masses.

Through experimentation it was decided that spatial domain techniques provided the best results and so were used in this project. Histograms are the basis for most spatial domain processing techniques. If the image is dark then the components of the histogram will be mostly on the low/dark side of the grayscale, whereas if an image is bright they will be on the high/bright side.

An image with low contrast will be relatively narrow and centered towards the middle of the grayscale. Below, Figure 3.1 depicts a mammogram (record number mdb010 from the MIAS database) and its histogram.

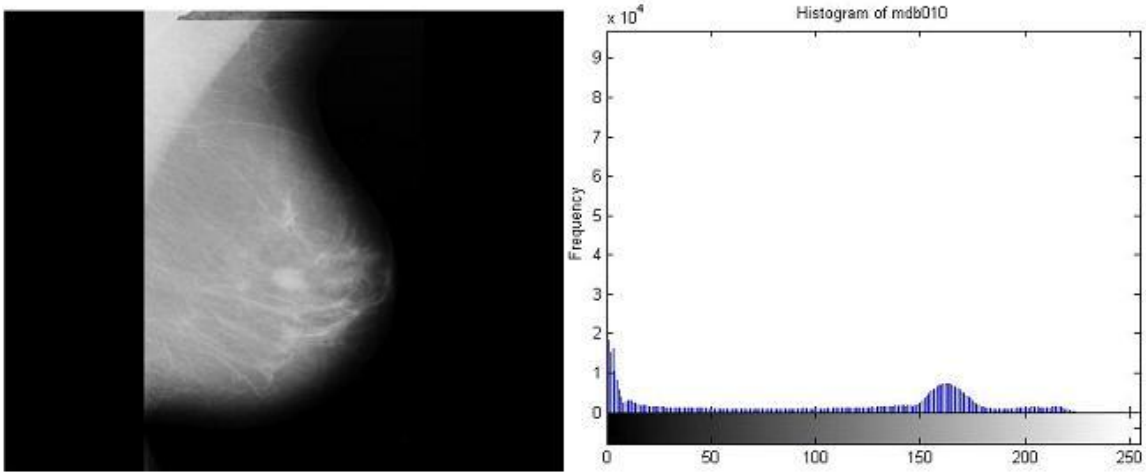


Figure 3.1 Mammogram (mdb010) and its histogram

As can be seen from Figure 3.1, the actual breast area is very dull and the histogram confirms this as most of the breast intensity values are between 150 and 170. A solution to this problem is histogram equalisation. Histogram equalisation works by basically spreading out the most frequent intensity values over the entire grayscale, 0 to 255 in this case. This allows areas of lower local contrast to gain a much higher contrast without affecting the global contrast. The downside to histogram equalisation is that the contrast of background noise may also be increased. This problem can be overcome through the use of CLAHE (Contrast Limited Adaptive Histogram Equalisation) which is an extension of histogram equalisation. CLAHE operates on small regions in the image called tiles instead of the whole image. Each tiles contrast is enhanced so that the histogram of the output region approximately matches that of a flat histogram. The neighboring tiles are then combined using bilinear interpolation to eliminate

any artificial boundaries. The results of CLAHE on the original mammogram shown in Figure 3.1 can be seen in Figure 3.2.

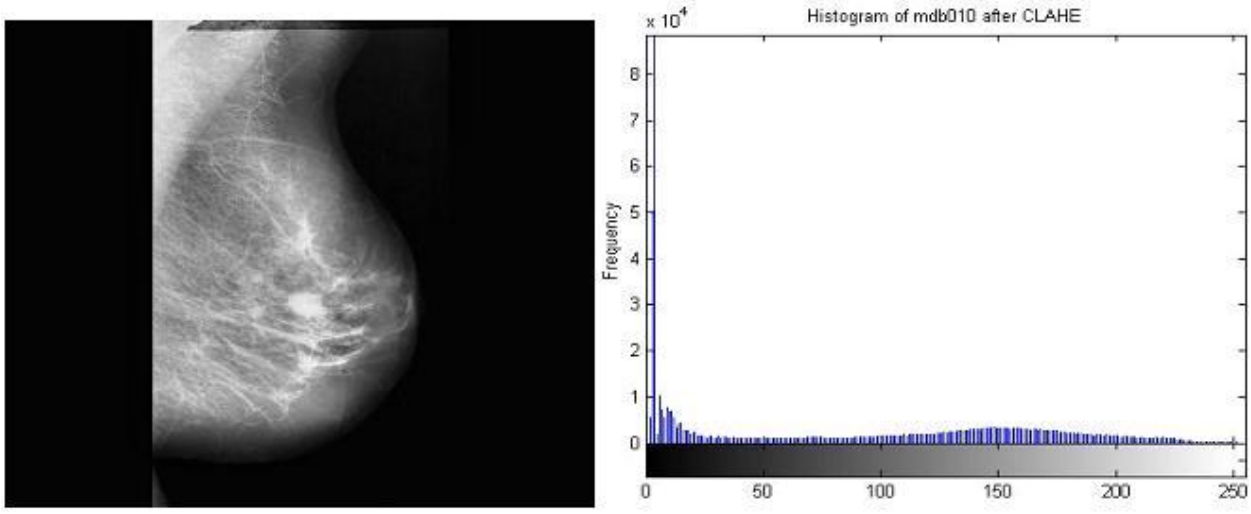


Figure 3.2 Mammogram (mdb010) and it's histogram after CLAHE is applied.

3.3 Image Segmentation

The next logical step in the system is to perform image segmentation, this involves separating the suspected areas they may contain abnormalities from the image. The suspicious area is an area that is brighter than its surroundings, has almost uniform density, has a regular shape with varying size, and has fuzzy boundaries [10]. A number of different segmentation techniques were investigated and are outlined in this section.

3.3.1 Global Thresholding

Global thresholding is one of the more basic segmentation techniques, and is based on the image histogram. Since masses are in general brighter than their surrounding tissues it makes thresholding useful for the type of segmentation needed. The segmentation involves setting a threshold, an intensity value, such that all pixels whose intensity values are less than the threshold belong to one category and the remainder belong to the other. In Figure 3.3 it can be seen that values below a certain threshold were set to one (black) and values above the threshold set to zero (white), leaving a binary image containing ROI's. However the downside to this technique is that the threshold value varies from image to image.

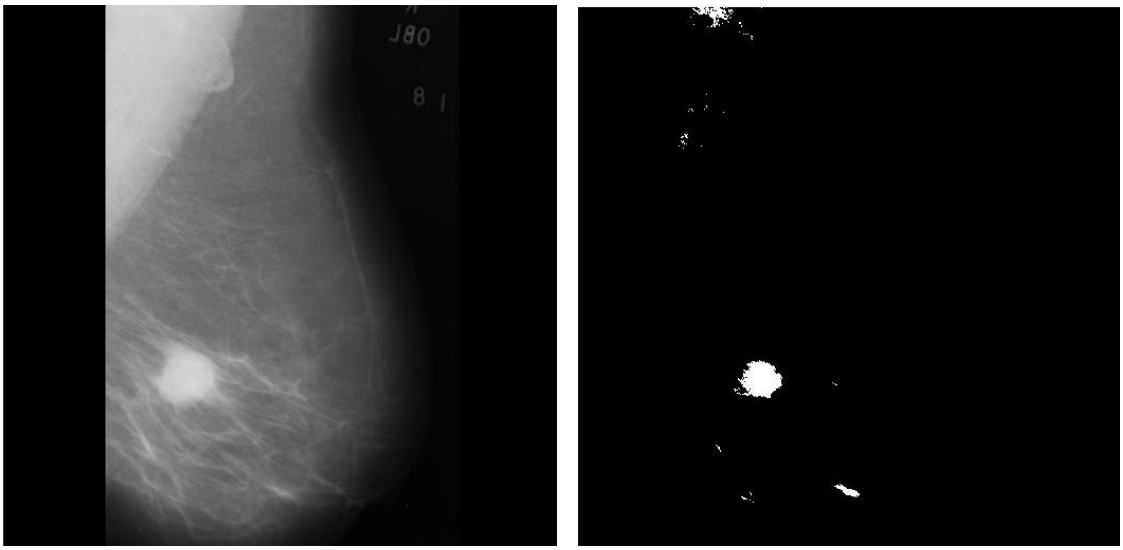


Figure 3.3 Original mammogram (mdb028) on the left and segmented ROI's on the right.

3.3.2 Region Growing

Another technique investigated was region growing. Region growing is a region based segmentation approach which finds the region directly. The method by which image regions are formed is as follows:

$(a) \bigcup_{i=1}^{n} R_i = R.$

$(b)\ R_i \text{ is a connected region}, i = 1,2, \ldots., n.$

$(c)\ R_i \bigcap R_j = \emptyset \text{ for all } i = 1, 2, \ldots, n.$

$(d)\ P(R_i) = TRUE \text{ for } i = 1, 2, \ldots, n.$

$(e)\ P\left(R_i \bigcup R_j\right) = FALSE \text{ for any adjacent region } R_i \text{ and } R_j.$

$P(R_i)$ is a logical operator defined over the points in set $P(R_k)$ and $\emptyset$ is the null set. Condition (a) indicates that the segmentation must be complete i.e. every pixel must be in a region. Condition (b) requires that points in a region must be connected in some predefined sense. Condition (c) indicates that the regions must be disjoint. Condition (d) deals with the properties that must be satisfied by the pixels in a segmented region-for example $R_i = TRUE$ if all pixels in R_i have the same gray level. Also condition (e) indicates that region R_i and R_j are different in the sense of predicate P [11].

Region growing is a procedure that groups pixels into larger regions based on a set of predefined criteria. The method works by first taking a set of seed points, which mark the object or objects to be segmented, and growing the regions iteratively by joining neighbouring pixels to the seed, if and only if they contain similar properties to that of the seed. This process continues until all the pixels in the image have been allocated a region. The following function was used in MATLAB to implement basic region growing:

```
[G] = regiongrow(f, S, T)
```

Where *f* is the input image to be segmented, *S* is a intensity value which defines all the points in f that will become seed points and *T* is defined as the global threshold value, which is used to test how similar a pixel is to that of the seed and if should be connected to it or not. A pixel is said to be similar to the seed if the difference between it and the seed pixel's intensity value is less than or equal to the threshold value. In Figure 3.4, through inspection of pixel values the seed point, *S*, was chosen to be 220 and the threshold value, T, to be 15. The output image, *G*, is also shown. As with thresholding the issue that arises with this method of segmentation is that an initial seed point is needed, and as this seed point varies from image to image, some sort of an adaptive seed point selection would be needed.

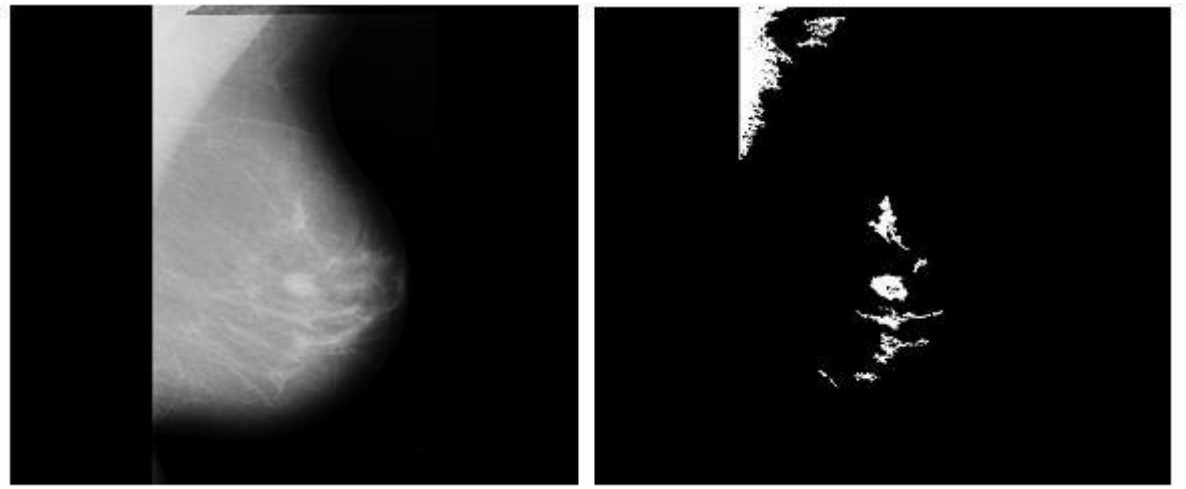


Figure 3.4 Original mammogram (mdb010) on the left with segmented ROI's on the right.

3.4 Noise Reduction

Sometimes mammograms can be affected by noise which can be random in nature i.e. white Gaussian noise, as a result of the X-ray system or the digitising camera used. Although the

images in the MIAS database do not suffer from noise, Gaussian noise was added to the images to simulate the effect. Noise reduction is then preformed using wavelet analysis.

3.4.1 Wavelets

Wavelets are a relatively new mathematical tool which has contributed significantly to image and signal analysis over the past twenty years. A wavelet can be defined as a mathematical function used to divide a given function into different scale components and each scale component can then be studied with a resolution that matches its scale. A wavelet transform is then a representation of a function by wavelets [12]. There are two types of wavelet transform: the continuous wavelet transform (CWT) and the discrete wavelet transform (DWT). The continuous wavelet transform of a function *f* using a wavelet function basis is defined as:

$$f(a,b) = \int f(x)\psi x_{a,b}(x)dx \qquad (3.1)$$

Where $\psi(x)$ is the mother wavelet function. The basis of the wavelet function is obtained by scaling and shifting a signal mother wavelet function.

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right) \quad ; \quad a > 0 \qquad (3.2)$$

Where a is a scale factor and b is the shift value. The DWT is obtained by taking $a = 2$ and $b = Z$. In both cases the transform, transforms the function into a function of scale and translation. While the Fourier transform uses sinusoids of infinite duration to decompose the wavelet transform uses wavelets of finite duration to perform the same operation.

When it comes to dealing with images the discrete wavelet transform is the transform chosen, the reasons are outlined in the coming section. The type of wavelet mother function chosen is that of the Haar wavelet, its mother function $\psi(t)$ can be described as follows:

$$\psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2} \\ -1 & \frac{1}{2} \leq t < 1 \\ 0 & otherwise. \end{cases} \quad (3.3)$$

Its scaling function *φ(t)* can be described as

$$\varphi = \begin{cases} 1 & 0 \leq t < 1 \\ 0 & otherwise. \end{cases} \quad (3.4)$$

3.4.2 Discrete Wavelet Transform (DWT)

The discrete wavelet transform is used for image processing as it gives a detailed insight to an image's spatial and frequency characteristics, unlike the Fourier transform or CWT which deal only with an image's frequency characteristics. The first stage of image de-noising is image decomposition.

The DWT of an image is most efficiently calculated by passing it through a series of filters, called a filter bank, as shown in Figure 3.5. In this figure, the image is represented by x[n], the low pass filter is represented by G[n], the high pass filter is represented by H[n] and the down sampling operator represented by ↓. At each level, the high pass filter produces detail coefficients for that level while the low pass filter produces approximation coefficients which are fed into the next level, and the process continues for as many levels as are required. The approximation coefficients are the high scale, low frequency components of the image whereas the detail coefficients (vertical, horizontal and diagonal) are the low scale, high frequency components. Image decomposition of the noisy mammogram can be seen in Figure. 3.6.

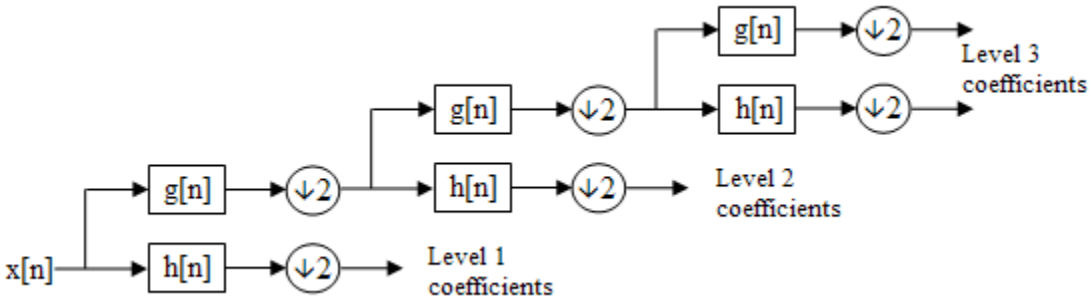


Figure 3.5 Three level filter bank [13].

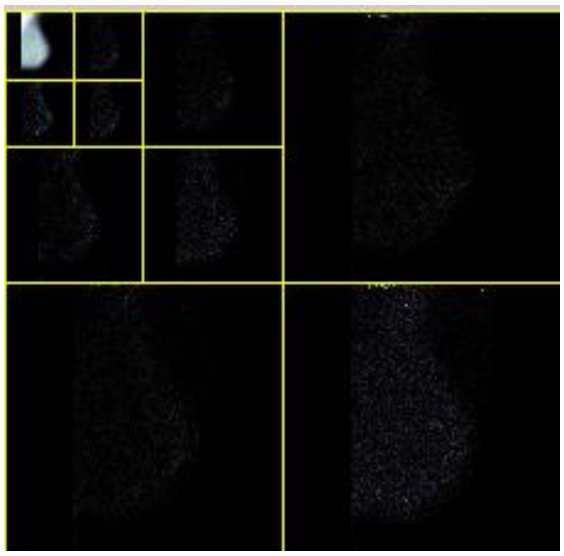


Figure 3.6 Three level decomposition of mammogram

3.4.3 Thresholding

The next stage in de-nosing the image is to threshold the coefficients. This involves selecting and applying a threshold to the detail coefficients for each level from 1 to N. This can be

achieved using hard thresholding, setting to zero all the elements whose absolute values are lower than the threshold or by soft thresholding which first sets to zero all the elements whose absolute values are less than the threshold, and then shrinking the non-zero coefficients towards zero. Soft thresholding is used in this project, although no visual difference was observed when hard thresholding was used.

3.4.4 Wavelet Reconstruction

The final step is to perform wavelet reconstruction using the original approximation coefficients of level N and the modified detail coefficients of levels 1-N. This is achieved by using the inverse discrete wavelet transform (IDWT). Where wavelet decomposition uses down sampling and filtering the reconstruction process consists of up-sampling and filtering. The noisy image is shown alongside the final de-noised image in Figure 3.7.

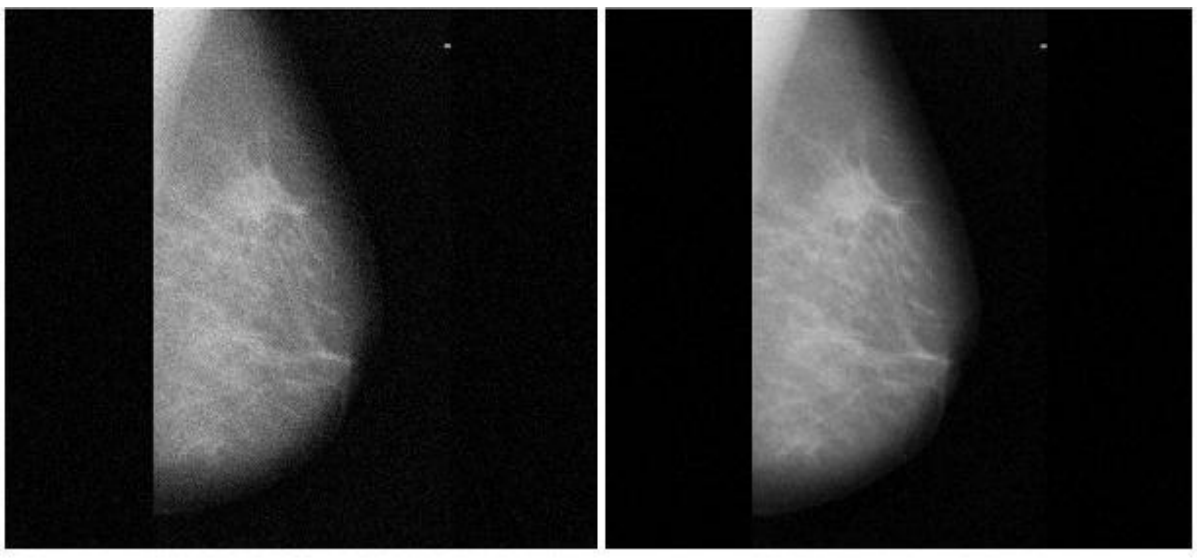


Figure 3.7 The noisy image on the left and the de-noised image on the right.

3.5. Feature Extraction and Selection

This section outlines the methodologies used to derive features representing abnormalities from the images.

3.5.1 Feature extraction

Mammograms in general contain a lot of different information depicting tissue types, vessels, and the X-ray film among others. All of this information would be too complex for any system to deal with and as a result a reduced representation set of features is used. Features in general can fall into three main categories, intensity features, geometric features and texture features. Two main approaches are examined in this system:

- First order statistical features (intensity features).
- Textural features using wavelet decomposition.

3.5.2 Feature Selection

Feature selection on the other hand, is defined as the process of selecting an optimum subset of features from the enormous potential features available in a given problem domain after the image segmentation [14]. A large number of features will in general require a large amount of memory and computational power, as well as possibly degrading the performance of the classifier, so often only a subset of the calculated features are used.

3.5.3 First Order Statistical Features

In this system, first order statistical features are used to extract features based on the intensity histogram of the ROI. This is achieved through the use of statistical moments based on the histogram. A similar approach is taken in literature, where Alolfe *et al*. [15], presents a computer aided diagnostic system for the detection of malignant tumors on digital mammograms, using first order statistical features. Using this approach as a basis the following statistical features were calculated: average intensity, average contrast, skewness, uniformity, entropy, and kurtosis. The expression for the nth moment about the mean is given by

$$\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n \, p(z_i) \qquad (3.5)$$

Where z_i is a random variable indicating intensity, $p(z)$ is the histogram of the intensity levels in a region, L is the number of possible intensity levels and

$$m = \sum_{i=0}^{L-1} z_i \, p(z_i) \qquad (3.6)$$

is the mean (average intensity). The remainder of the moments are listed and described in Table 3.1. The moments were calculated in MATLAB on ROI's from the MIAS database, where the location of the abnormality is known. Where there is no abnormality i.e. a normal breast then a large area of the breast was selected, to achieve a full representation of the breast.

Moment	Expression	Measure of Intensity
Standard Deviation	$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}$	A measure of average contrast.
Third Moment	$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 \, p(z_i)$	Measures the skewness of a histogram.
Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$	Measures the uniformity of intensity in the histogram.
Kurtosis	$\mu_4 = \sum_{i=0}^{L-1} (z_i - m)^4 \, p(z_i)$	Measures the relative flatness of the intensity in a region.
Entropy	$e = -\sum_{i=0}^{L-1} p(z_i) \, \log_2 p(z_i)$	A measure of randomness.

Table 3.1 Table describing intensity measures based on the intensity histogram

3.5.4 Wavelet Decomposition

The second approach used to extract features is by using wavelet decomposition. This is based on texture analysis of the mammogram. The texture of an image can be described as the spatial variation in gray levels. As expected, different types of mammograms have different types of texture. If the texture types were all the same some sort of autocorrelation function could be used however as the each mammogram is different this is not an option. A solution to this is proposed by Ferreira [16], whereby the discrete wavelet transform is chosen to decorrelate the data. The DWT is chosen because it can decorrelate the data without affecting the main distinguishable characteristics of that data. For this reason it was chosen to be used in this

project. The process of wavelet decomposition was previously described in detail in section 3.4.2.

As in literature [16], low frequency approximation coefficients are chosen as a representation of texture for each image as these coefficients contain the most information about the image. In this project, wavelet decomposition through the use of the Daubechies 4 (DB4) wavelet was applied to areas of 64 x 64 pixels with the abnormality centred, as can be seen in Figure 3.8. The DB4 wavelet has been proven to perform well in literature [16]. The reason behind choosing a window of fixed size was so that the same number of wavelet coefficients was computed for each image. Applying one level of decomposition produced hundreds of wavelet coefficients. It was decided only to use the hundred biggest coefficients as an input to the classifier to reduce complexity.

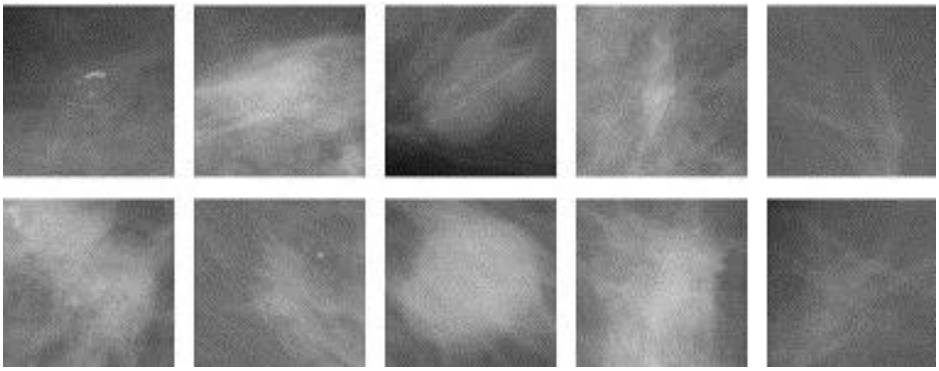


Figure 3.8 Samples of ROI's used.

Chapter 4: Classification

This chapter provides some background information on neural networks, describes the classification architecture used in the system and also how it was implemented.

4.1 Artificial Neural Networks

An Artificial Neural Network (ANN) is a computational model for processing information, it is inspired by the way the biological nervous system operates. In the human brain for example, a neuron, as seen in Figure 4.1, will receive electrical signals from other neurons and when this input exceeds a certain threshold the neuron will fire i.e. output an electrical pulse. Learning occurs by modifying the connections between these neurons. In an ANN a biological neuron is represented by an artificial equivalent.

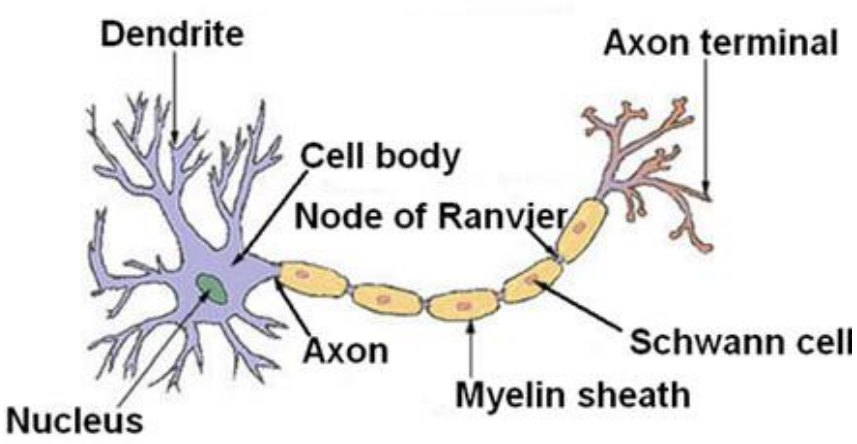


Figure 4.1 Biological Neuron [17].

An artificial neuron is modeled in Figure 4.2. In this model, '*p*' is the input vector of '*R*' elements and '*b*' is a scalar bias. The input to the transfer function is the sum of the product '*Wp*' (weighted input) and the bias '*b*'. This in turn gets passed to the activation function '*f*' which produces the neurons output '*a*'. This output is a scalar and mirrors the firing of a biological neuron.

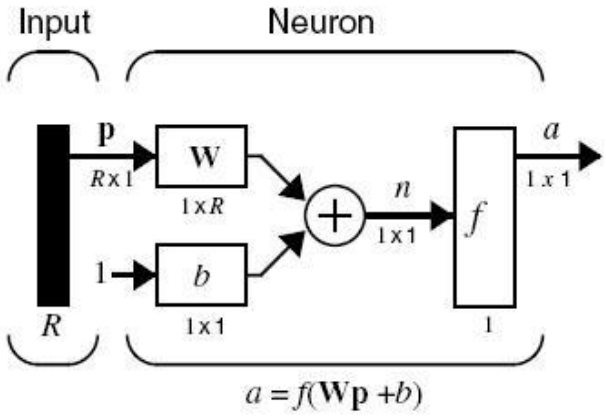


Figure 4.2 Model of an artificial neuron in MATLAB [9].

There are three main steps in using ANN's. Firstly, a suitable architecture or network structure is chosen depending on the type of specific data as well as the application. It is important that the structure chosen isn't overly complicated for the task at hand, otherwise it will lead to learning difficulties. The second step involves training. The neural network is trained using a training algorithm and a training data set. Initially the weights applied to the interconnections are randomly chosen, but as training continues the weights are constantly adjusted until best performance is achieved. Training can be performed in one of two ways, either supervised training or unsupervised training. Supervised training is used in this system, this means that during training each input knows what its corresponding output should be and the network tries to map them accordingly. ANN's are like people in that they learn from doing. The final step is testing. After the network has been trained, its structure is saved and evaluated using a different

data set (test set), from the same source of course, to ensure that proper training has indeed occurred.

One of the main advantages ANN's have over standard programming methods is that they are a non-linear model, meaning they are very good at solving problems that relate to pattern recognition and prediction. As a result ANN's are used in a variety of applications including medical diagnosis, signal analysis and interpretation and even in stock price prediction. For these reasons it was decided that a neural network would be a suitable classifier for this project.

4.2 The Multilayer Perceptron

A Multilayer Perceptron (MLP) is a feed forward supervised neural network. The MLP in its most simplistic view consists of a network of processing nodes arranged in layers. There must be three or more of these layers in an MLP, including an input layer to accept the input vectors, one or more hidden layers, and an output layer with one node per class. Every node in one particular layer is connected to every node in the layer above and below it, as can be seen if Figure 4.3. The connections carry the weights that are adjusted during training.

The operation of the network can be broken down into two main stages, the forward pass and back propagation. In the forward pass, an input vector is presented to the network and the output of the input layer nodes is the components of the input pattern. For successive layers the input to each node is then the sum of the scalar products of the incoming vector components with their respective weights, as detailed earlier. Back propagation on the other hand is an iterative process whereby the multilayer network is trained. In the learning phase, the set of training data is presented to the input layer together with their corresponding desired outputs, which as previously stated represent the input classification. Initially with random weights, for each input the network must adjust the weights attached to the connections so that difference between the network output and that of the desired output for that input vector is reduced. This process

continues until a stable set of weights is achieved. A flow diagram of this process is included as Appendix A.

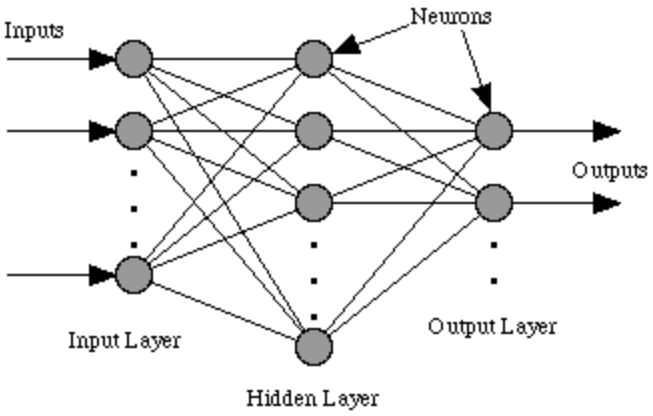


Figure 4.3 Multilayer Perceptron [18].

4.3 Neural Network Implementation

Implementation of the MLP neural network was carried out in MATLAB, using it's neural network toolbox. This made creating a simple MLP relatively easy as only knowledge of a few basic commands was needed.

The first of these commands is 'newff', this creates a feedforward network object. This function requires four input parameters.

```
net = newff(minmax(Inputs_trg), [hidden_nodes output_nodes], {'tansig'
'purelin'},'traingdx');
```

The first argument is a matrix of input vectors, which for this project was a set of features, as discussed in sections 3.5.3 and 3.5.4. The second argument defines the number of hidden and output nodes to be used. The third argument contains the names of the transfer functions used in the network, the tan-sigmoid function is used in the hidden layer while the linear transfer

function is used in the output layer. The final argument defines the training algorithm used, in this instance it is 'traingdx' which is a training algorithm that changes weight and bias values depending on the gradient descent momentum and the learning rate.

Once all of the biases and weights are initialised the next step is to train the network. Training is preformed using the 'train' function. This function performs batch training which means that the weights and biases of the network are only altered after the complete training set has been passed to the network.

```
[net, tr] = train(net, Inputs_trg, Targets_trg);
```

The first argument is the new network that has been created. The second argument is the set of input vectors as mentioned previously and the third argument is the set of network target vectors. Each target vectors contains three numbers, -0.9, -0.9 and 0.9. The target vectors have to be representative of the input data. For example, if an input vector represents a cancer then the corresponding target vector should be [-0.9, 0.9, -0.9], likewise for benign it should be [-0.9, -0.9, 0.9] and hence for normal [0.9, -0.9, -0.9].

The final stage in the development of the MLP is quite simply to simulate the network. This performed using the 'sim' command:

```
outputs = sim(net, inputs);
```

Again the first argument is the new network created by the 'newff' function, and the second argument is the input vectors. The trained network is first simulated using the training data to ensure that training has taken place, once it can be verified that it has, the network can then be simulated using validation and test data. The results can be seen in the next chapter.

Chapter 5: Classifier Results

In this section performance and complexity of both front ends, as inputs to the classifier are compared.

5.1 Performance of the front ends

The accuracy of the system, using both front ends, is measured using three parameters: performance, specificity and sensitivity. Performance is defined as the percentage of correctly identified cases. The formula's for specificity and sensitivity are shown below.

$$Specificty = \frac{TN}{TN+FP} \ , Sensitivity = \frac{TP}{TP+FN}$$

Where TN is the rate of true negative, which represents the cases that don't have cancer and have been correctly classified. FP is the rate of false positive, this represents the cases that don't have cancer but have been incorrectly classified as having cancer. TP is the rate of true positive, this represents the number of cancer cases correctly classified. Finally, FN is the rate of false negative, this represents the cases that have cancer but have been incorrectly classified as not having so. A high sensitivity means that few cancers will be missed by the system. Sensitivity is however the more important parameter, as a false negative means there has been an error in recognising cancer which could be life threatening. While a false positive isn't life threatening, it causes the patient great anxiety and can lead to a waste of valuable resources. The results are presented in the form of a confusion matrix. A confusion matrix is simply a matrix that displays the number of correct and incorrect classifications made by the network.

5.2 First order statistics and the network

The first order statistics, described in chapter 4, based on histogram intensity were used as input vectors to the MLP, along with corresponding output targets. Initially the system was trained and tested, however it was clear that some modifications needed to be made in order to achieve optimum performance. There were two main parameters that had the biggest impact on the performance of the network: the number of hidden nodes and the learning rate. If the number of hidden nodes is too few, the network is not be able to handle the complex computations it has to perform, but if the number of hidden nodes is too many it results in overtraining. Likewise, if the learning rate is too small the network will spend to long adjusting the weights and might never reach the desired goal. However if the learning rate is too large, the system may miss slight discrepancies in information which could affect the overall performance. Experiments were carried out to find values for both the number of hidden nodes and learning rate that would enable the network achieve optimum performance. The results are graphed for hidden nodes and learning rate in Figures 5.1 and 5.2 respectively.

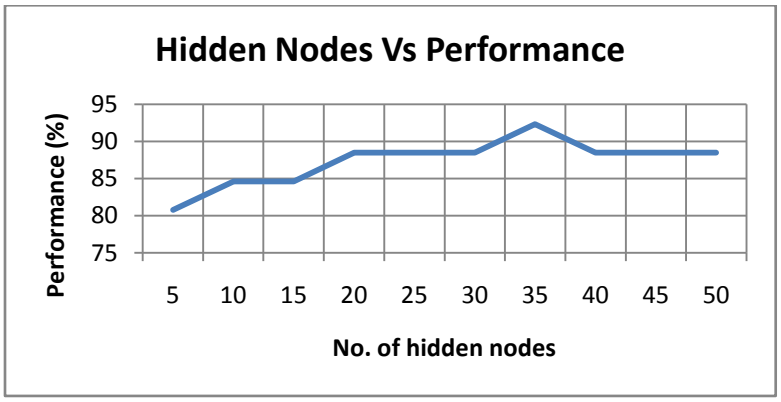


Figure 5.1 Graph of Hidden Nodes Vs Performance

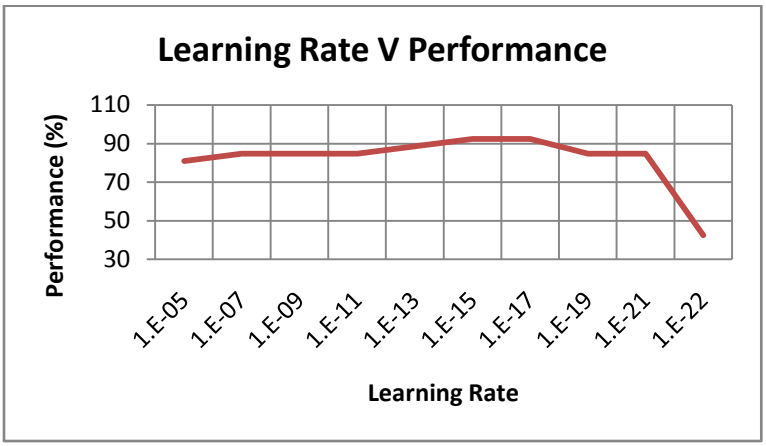


Figure 5.2 Learning Rate Vs Performance

The optimum number of hidden nodes was found to be 35, and the optimum learning rate was at $1\ exp^{-15}$. This number of hidden nodes is relatively large, and could be reflected by the fact that only six features are used at the input.

With these parameters the network was trained and tested. The confusion matrix for the test data is shown below in Figure 5.3, where class one represents normal, class two represents malignant and class three represents benign.

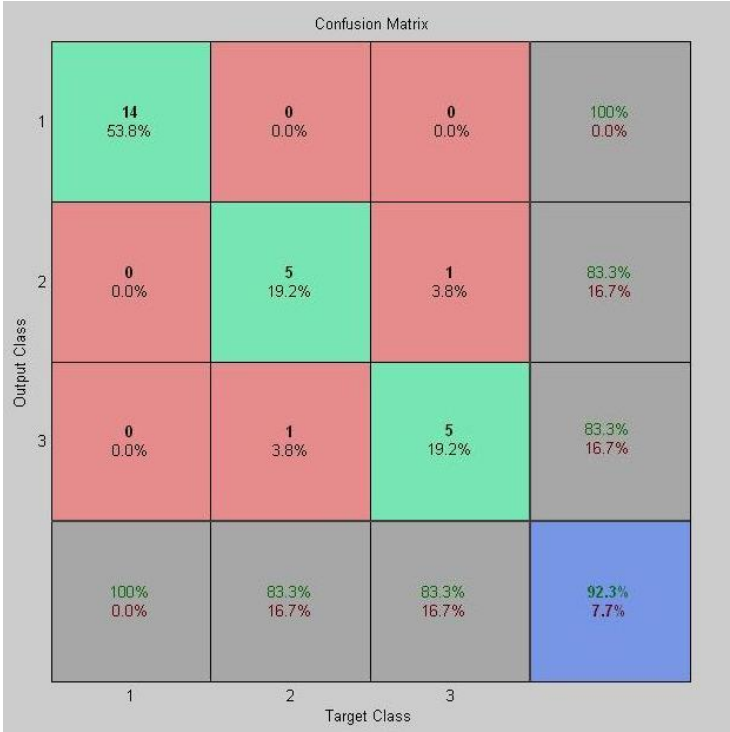


Figure 5.3 Confusion matrix of test data (MLP with statistics).

As can be seen from Figure 5.3 the model correctly predicts 100% of the normal cases, it correctly predicts five out of the six malignant cases, although it classifies one of them as benign and finally it predicts five out of six benign cases but incorrectly classifies one of them as malignant. This results in an overall successful classification rate of 92.3%, specificity of 95% and sensitivity of 83.3%.

5.3 Wavelet coefficients and the network

For the second study, the wavelet coefficients described in chapter 4, are used as input vectors to the MLP, along with their corresponding output targets as before. As this model was trained and tested it became apparent that no learning was actually taking place. As an experiment it was decided to only use two output classes to see if the network could distinguish between say normal and malignant cases. This experiment proved to be successful and so it was decided to split up this second MLP in to two MLPs, one for classifying normal and tumorous tissue and the second for classifying malignant and benign cases.

The first model created was to distinguish between normal and tumorous tissue. The same experiments with regards to finding optimum values for hidden nodes and learning rate were carried out as previously mentioned. This resulted in five hidden nodes and a learning rate of $1\ exp^{-15}$. The reduction in the number of hidden nodes needed in comparison to the initial MLP is mostly likely caused by the increase in input features available to the network. The results are illustrated in the confusion matrix in Figure 5.4, class one represents the normal cases and class two represents the tumorous cases.

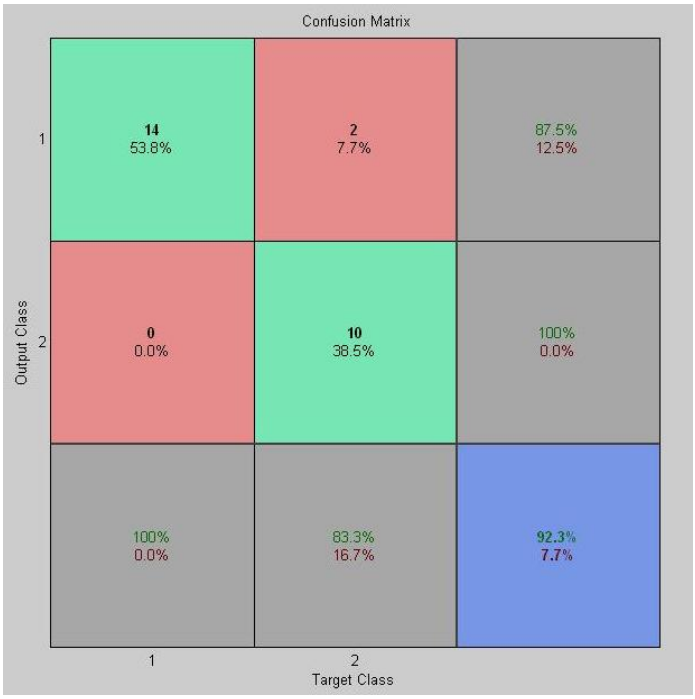


Figure 5.4 Normal Vs tumorous tissue, confusion matrix of test data (MLP and wavelets coefficients).

As can be seen from Figure 5.4 the model correctly predicts 100% of the normal cases. It does however incorrectly classify two of the tumorous cases as normal. This results in a successful classification rate of 92.3%, specificity of 100% and sensitivity of 83.3%.

The second model was created to distinguish between malignant and benign cases. The optimum number of hidden nodes and learning rate value were derived as before and were found to be 5 and $1\ exp^{-7}$ respectively. The results are illustrated in the confusion matrix in Figure 5.5, class one represents the malignant cases and class two represents the benign cases.

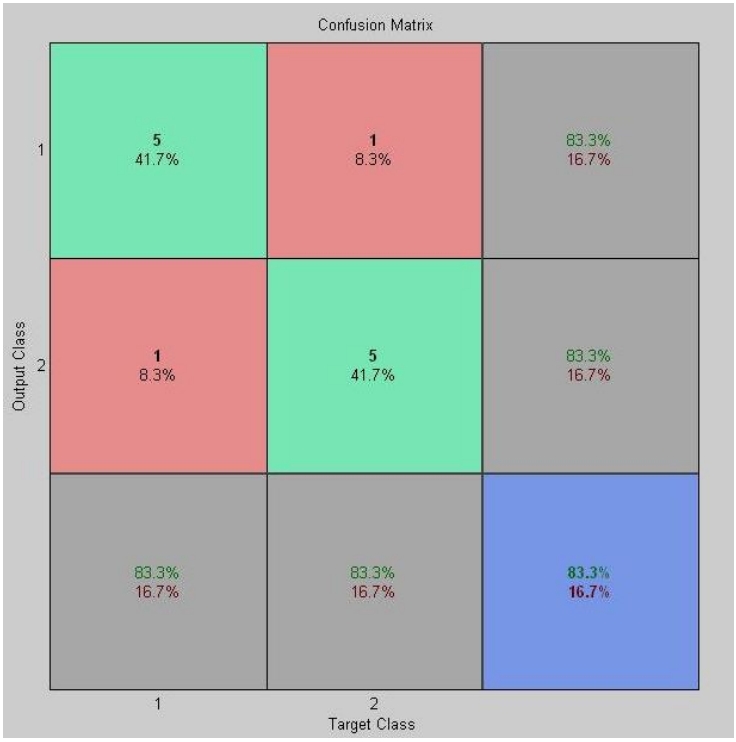


Figure 5.5 Malignant Vs Benign, confusion matrix of test data (MLP and wavelet coefficients).

As can be seen from Figure 5.5 the model correctly predicts five out of six of the malignat cases and also five out of six of the benign cases. This results in a successful classification rate of 83.3%, specificity of 83.3% and sensitivity of 83.3%.

The overall performance of the wavelet coefficient based MLP can be determined by multiplying the results of the first MLP (normal and tumorous tissue) and the second MLP (benign and malignant). This results in an overall performance of 76.9%, overall specificity of 83.3% and overall sensitivity of 69.4%.

5.4 Conclusions

The results show that the first order statistics perform better than the wavelet coefficients as input vectors to the network. When the first order statistics were used as the front-end processor the performance was 92.3%, specificity of 95% and sensitivity of 83.3%. While a performance of 76.9%, specificity of 83.3% and sensitivity of 69.4% was achieved when the wavelet coefficients were used. Thirty five hidden nodes were needed to produce the results for the first order statistics approach and five hidden nodes for each of the wavelet based networks.

Chapter 6: Database Web Application

This chapter describes the design and implementation of the online database for physicians to view images and upload images. It initially details the technologies used and the reasons behind using them.

6.1 Database Selection

The database is used to store all of the patient information along with the images so it is important that a suitable database was selected.

Structured Query Language (SQL) is effectively a standard programming language for creating, updating and retrieving information that is stored in relational database management system. A rational database is a database divided into logical units called tables, where tables are related to one another within the database, this allows for large amounts of data to be broken down into smaller more manageable parts. An SQL database was an obvious choice, there are however a number of SQL databases on the market each with their own advantages, these include Microsoft SQL Server, MYSQL, Oracle, SQLite and PostgreSQL among others. With respect to speed and functionality, Microsoft SQL Server outperforms the rest however it is very expensive and would be over the top for a basic database, which is all that is needed for this project.

After all the options were weighed up, it was decided to use MYSQL Database Software for a number of reasons. MYSQL is the world's most popular open source management database developed by MYSQL AB. It has over ten million installations worldwide, including sites like Google and Facebook, and last year it was bought by Sun Microsystems for one billion US dollars! The main reason this software was chosen was due to its simplicity and obvious popularity. Without ever knowing anything about SQL or MYSQL, anyone with basic programming knowledge could create simple databases with a few basic commands. As a result

of its popularity there are endless amounts of tutorials, support and information available. Another advantage and reason for choosing MYSQL is that it is free under the GNU General Public License. MYSQL was also designed with speed in mind and includes an Open Data Base Connection (ODBC) which means that a connection can be made to the database regardless of the programming language used. MYSQL also comes with numerous tools and literature to help make sure that creating and managing the database is as easy as possible.

MYSQL will be used to store the patients ID, their physician, the hospital, the image and also any comments which may be relevant.

6.2 PHP and HTML

The database needed to be accessible and viewable via a web-browser, so a web based interface needed to be designed.

PHP (Hypertext Pre-processor) is a scripting language designed specifically for use on the internet. More often than not it is used for server side scripting and was developed for dynamic web pages. PHP allows the standard hypertext web pages to become dynamic and for this reason it was chosen for this project, as the webpage will constantly be updated with new images. PHP is an embedded scripting language i.e. it's code is embedded within HTML, it works seamlessly with HTML and writing PHP tags around PHP scripts along with saving the file with extension ".php" is all that is required. PHP scripts can parse data submitted by HTML forms and communicate with databases which make it ideal for this application. A big advantage of PHP is that it will work on virtually all web servers, all operating systems and like MYSQL is openly available free of charge under the GNU General Public License. PHP provides the application part while MYSQL provides the database part of the Web database application. PHP is used in this project to manage the user login, handle the uploading of patient information and also is used to query the database and in turn generate dynamic web pages to view the results.

HTML (Hypertext Markup Language) is a programming language that was designed for the creation of web pages. HTML was used alongside PHP in the implementation of the database application. HTML is written in the form of tags, the tags tell the browser how to display the images or text, this is known as "Markup".

6.3 WAMP

WAMP, which stands for Windows Apache MYSQL and PHP is used along with PHP and HTML to develop the web application. When WAMP is installed on a computer it installs the list of applications mentioned below for a windows environment. The equivalent installation in Linux is known as LAMP.

- Apache
- MYSQL
- PHP
- PHPmyadmin
- SQLitemanager
- Wampserver service manager

The service manager that comes with WAMP allows full control of the server and local projects. All programs can be started and stopped as required using the service manager. Probably the most useful aspect of WAMP is that it allows the user to test scripts locally on a computer before they are uploaded to an actual server.

6.4 Designing and Creating the Database

The first step in this process was to decide what the database would like and what information it needed to hold. The system would basically need two tables, one for storing physician login information and another to store the patient information and images.

Initially the table to contain the list of usernames and passwords was created. Obviously the table needed to contain username and password variables, but an ID number was also included to ensure that the table could be used for other applications if so desired. The ID number was chosen as the primary key of type INTEGER, and will increment each time a new user is added. The username and password are created as type VARCHAR (40) meaning that it can include numbers and characters up to a length of forty. To ensure that neither the username nor password was left empty, a safety net was included to catch such exceptions by declaring that the username or password cannot be NULL. A graphical representation of the designed table structure is shown in Figure 6.1 where 'PK' denotes the primary key.

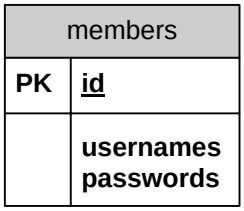


Figure 6.1 members table

Once the structure of the table had been designed it could be coded. As MYSQL code is run in real time it was not possible to save it onto the attached CD, as a result the code is shown in Figure 6.2. The name of the database is 'project' and the name of the table is 'members'.

```
CREATE TABLE `members` (
`id` int(4) NOT NULL auto_increment,
`username` varchar(40) NOT NULL default '',
`password` varchar(40) NOT NULL default '',
PRIMARY KEY (`id`)
) TYPE=MyISAM AUTO_INCREMENT=2 ;
```

Figure 6.2 MySQL code for members login table

To populate the table with users the following command was used,

```
INSERT INTO `members` VALUES (1, 'enda', md5('1234'));
```

In this example 'id' is given the value '1', username is created as 'enda' and the password as is an encrypted MD5 version of '1234'. MD5 (Message-Digest algorithm 5) is a widely used internet standard method of encryption. It makes use of 128-bit hash table to perform its encryption. It was decided to use it for this application to increase security.

The second table to be created would contain all of the patient information. This table, which is called 'patientinfo', is shown in Figure 6.3. As with the previous table an ID number was chosen as the primary key of type INTEGER, auto incrementing. The patients number is defined as 'patient no', of type INTEGER and has a NOT NULL clause attached to it. When it came to deciding how to store the images, there were two options available. Firstly, store the actual image in the table or secondly, store the image name in the table. If the image were to be stored in the table, it would be of type BLOB (Binary Large Object). There are a number of disadvantages associated with storing BLOB's in databases, namely that if the database crashes all of the images could be lost and also its much slower to take a file out of the database than point to a file that is already located on the server and. For these reasons it was decided to store the image name in the table, with the actual images stored in a folder on the server. The other entries include doctor, who is the physician of the patient, hospital, which gives the name of the hospital the patient belongs to, image, which contains the name of the image, and finally notes, which the physician uses to describe the image.

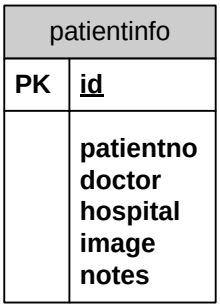


Figure 6.3 patientinfo table

The code used to create this table is as follows:

```
CREATE TABLE `patientinfo` (
`id` int(4) NOT NULL auto_increment,
`patientno` varchar(40) NOT NULL default '',
`doctor` varchar(40) NOT NULL default '',
`hospital` varchar(40) NOT NULL default '',
`image` varchar(40) NOT NULL default '',
`notes` varchar(40) NOT NULL default '',
PRIMARY KEY (`id`)
) TYPE=MyISAM;
```

Figure 6.4 patientinfo table

6.5 Web Application

Once all of the necessary MySQL tables were created, the next step involved creating the web application which would allow physicians to remotely access, view and upload images. The

application needs to be secure as so that only authorised personnel can view and make alterations to the database. The application doesn't need to be overly complex and so should be kept as simple as possible to avoid confusion, while still achieving its objective. A flow diagram illustrating the flow of the application can be seen in Appendix B.

As previously mentioned one of the most import aspects of the application is security, especially seeing as patient-doctor confidentiality may be at risk. The connection to the database on the server can be achieved through an ODBC connection (described earlier), which means the connecting password is not shown anywhere in the code making it very secure. When someone tries to access the application they are presented with a log in screen, where they enter their user name and password, the systems checks these against values in the members table and the person will only be allowed through if the records match.

One of the disadvantages in coding in PHP is that it is one of the more prone languages to hacking. Injection attacks could have catastrophic effects if they were allowed to get through. This system is protected against SQL injection, which is a hacking technique that exploits weakness in security in the database layer of the web application. A successful SQL injection exploit can read sensitive data from the database, modify database data (Insert/Update/Delete) and even execute administration operations on the database, such as shutdown the DBMS [19].

Often web applications contain a design flaw whereby the link to the page following the log in screen is displayed in the HTML code, therefore a hacker could just view the HTML source code for the log in and copy the URL into the address bar of the web browser. To protect against this PHP sessions are used. When a user successfully logs in, a session is started and registered in that users name, whereas if a user simply copy's the URL the session will not be registered and the application redirects the login screen. This ensures that only somebody with a correct username and password will have access to the system.

6.5.1 Application Design

The design of the application started with the design of the login screen. The login screen consists of a basic HTML form which is improved upon through the use of CCS (Cascading Style Sheets). CSS was used to enhance the appearance of the form and give it more of a professional look, the login screen can be seen in Figure 6.5. After a user enters their username and password, and clicks the Login button, the members table is queried and if the credentials match then the user is allowed to view the image database. If their credentials do not match then they are presented with a message telling them so.

Figure 6.5 Login form

The login page links to the main image database page. This page is basically a HTML table that displays all of the information contained in the table 'patientinfo', with a link to view each of the images. This is achieved by connecting to the database to get all of the information and then

using a loop to continuously output a new table row for each record until they have all been displayed. This page includes an option for the user to logout, whereby their session is terminated and they are returned the log in screen, a flow diagram is included as Appendix B.1. Another option available to the user is to upload an image to the database by clicking on the appropriate link. This link will bring the user to another page where they are presented with a HTML form in which to enter patient and image information. A PHP script takes care of writing the information to the database and saving the image to the server, also if there is an error uploading the image an appropriate message is displayed, a flow diagram illustrating the steps involved in uploading an image is included as Appendix B.2.

Chapter 7: Conclusion

This chapter gives a brief summary of the project and further areas of development are discussed.

7.1 Project Summary

In this project, a study was undertaken to try and automate image analysis techniques for the screening of mammography images. A brief overview of the outcomes is detailed in this section.

Image enhancement was the first area examined, and the results from the performance of CLAHE are visually very good. Numerous image segmentation techniques were investigated however no one technique was found to be sufficient at segmenting ROI's from a selection of different mammograms. Image de-noising is also performed using wavelet decomposition and reconstruction and the results from this are quite promising.

Two methods of computer aided classification are presented in this project. The first method, makes use of statistical features as input vectors to an MLP, the results are relatively good however they are limited by the size of the database. The model achieves an overall classification rate of 92.3%, specificity of 95% and sensitivity of 83.3%. The second method makes use of DB4 wavelet coefficients however the results are not quite as good. This model achieves an overall classification rate of 76.9%, specificity of 83.3% and sensitivity of 69.4%. This may be due to the fact that a window of fixed size was used which meant that information representing areas other than that of the abnormality was also processed.

The project was also extended to provide for a web database application, this was implemented using the rational database management system, MySQL and the server side scripting language, PHP. This application allows the user to view images and upload images as well as providing the necessary security to do so.

7.2 Further Developments

There are a number of further developments that could be incorporated into this project. Firstly there are many more image processing techniques that could be examined but due to time constraints with this project the surface was only scratched.

The classification performance of the network could be tested on a different database. It was initially hoped to test it on a second database, the DDSM University of South Florida database [20], however this database has not been active for the past number of years and there were issues with uncompressing downloaded images, so this could not be tested.

With regards to the front ends of the classifiers, the number of input statistics could be increased, this may or may not impact positively on the system. Experiments could be carried out using different types of wavelets, changing the number of coefficients selected, the class of coefficient selected, for example rather than using the low frequency approximation coefficients use the vertical high frequency coefficients, or possibly even a selection of both.

Also experimentation could be carried out using different types of classifier. In this project the MLP is used but other classifiers such as K-Nearest Neighbour (K-NN) could be used, or moving away from artificial neural networks something like a binary decision tree could be used.

References

[1] http://www.vhi.ie/hfiles/hf-018.jsp

[2] I. Christoyianni, E. Dermatas and G. Kokkinakis, Fast detection of masses in computer-aided mammography, *IEEE Signal Process. Mag.* **17** (2000) (1), pp. 54–64

[3] http://www.medicinenet.com/breast_cancer/article.htm

[4] http://www.breastcancersource.com/

[5] www.answers.com/topic/breast-cancer

[6] http://www.wma.net/e/publications/pdf/2000/giger.pdf

[7] M. L. Giger, "Computer-aided diagnosis," *RSNA Categorical Course Phys*., pp. 283-298, 1993.

[8] http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html

[9] http://www.mathworks.com

[10] S.M. Lai, X. Li and W.F. Biscof, On techniques for detecting circumscribed masses in mammograms, IEEE Trans. Med. Imaging l8 (1989) (4), pp. 377–386

[11] R. C. Gonzalez and R.E. Woods, *Digital Image Processing 2nd Edition*, Prentice Hall, New Jersey, 2002.

[12] http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html

[13] http://upload.wikimedia.org/wikipedia/commons/2/22/Wavelets_-_Filter_Bank.png

[14] M. Sameti, R.K. Ward, J. Morgan-Parkes and B. Palcic, A method for detection of malignant masses in digitized mammograms using a fuzzy segmentation algorithm, *Proceedings of the 19th International Conference IEEEE/MBS* (2000), pp. 513–516

[15] M. Alolfe *et al*, "Development of a Computer-Aided Classification System for Cancer Detection from Digital Mammograms", *Proc. National Radio Science Conference*, Egypt, 2008.

[16] Ferreira and Borges, 2003 C.B.R Ferreira and D.L. Borges, Analysis of mammogram classification using a wavelet transform decomposition, Pattern Recognition Lett. 24 (2003), pp 973-982.

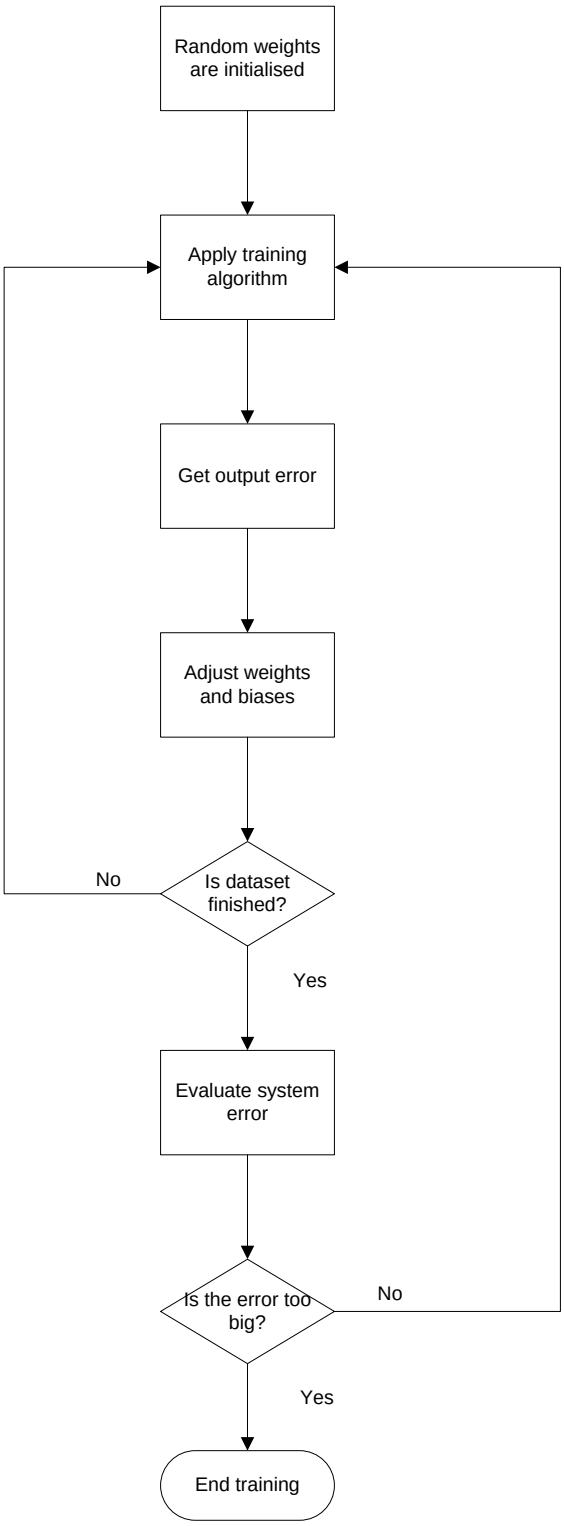
[17] http://www.health-fitness.com.au/the-brain-cell-neurons-and-neuroglia/

[18] http://www.geocomputation.org/2000/GC016/GC016_01.GIF

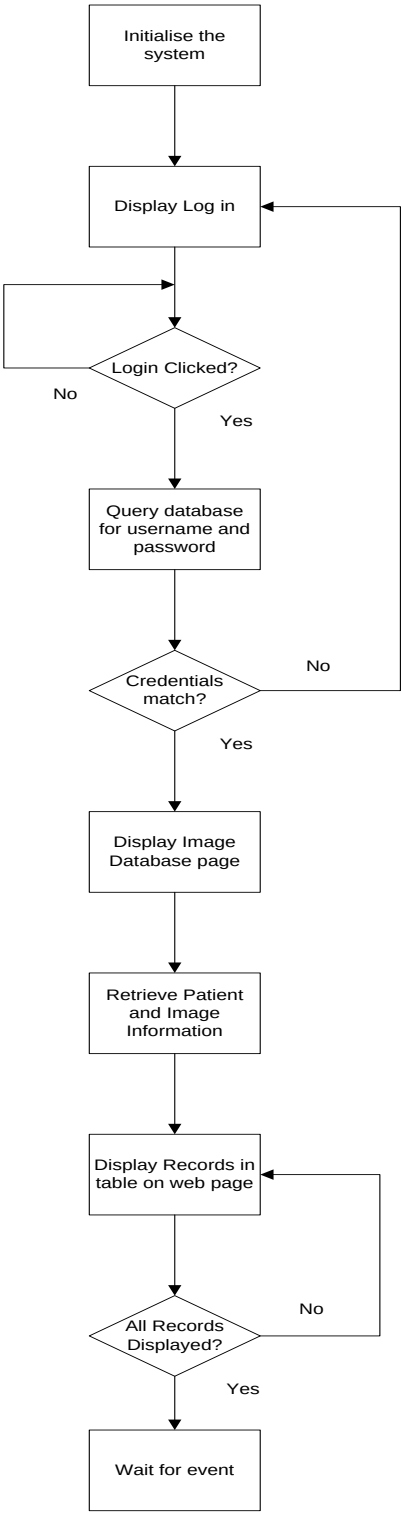
[19] http://www.owasp.org/index.php/SQL_injection

[20] http://marathon.csee.usf.edu/Mammography/Database.html

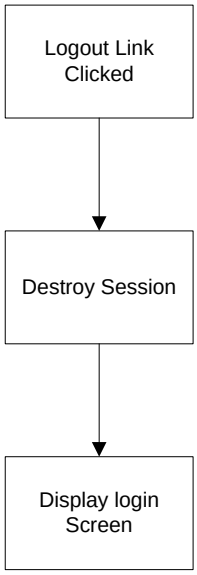
Appendix A – MLP Training Flow Diagram



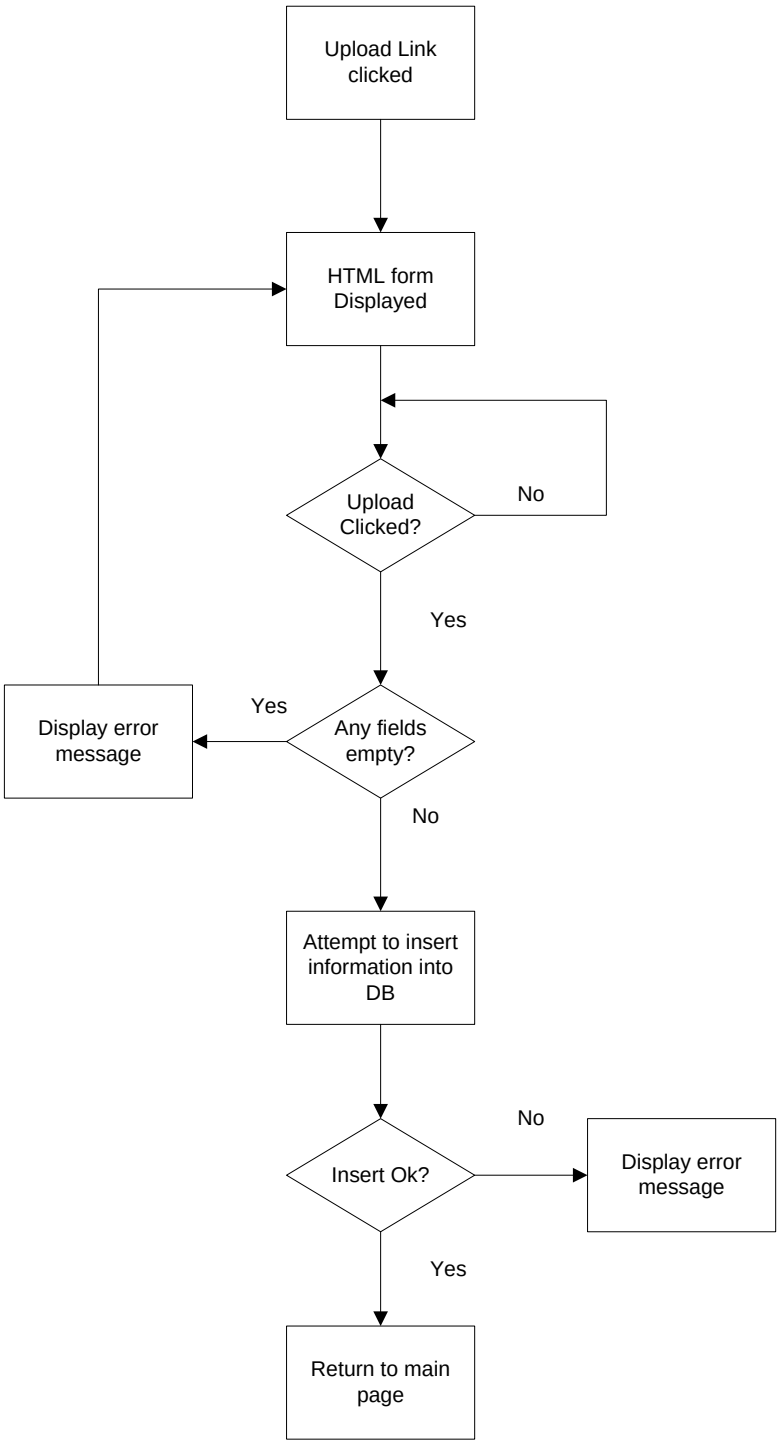
Appendix B – Web App. Database Flow Diagram



Appendix B.1 – Log out Event Flow Diagram



Appendix B.2 – Image Upload Event Flow Diagram



Electronic Appendix

The attached CD contains the following:

MATLAB folder:

1. **imsegment.m:** program that segments an image resulting in a binary image.
2. **regiongrow.m:** function that performs region growing.
3. **statmoments.m:** program that calculates statistical moments.
4. **intenstat.m:** function that returns six statistics on a given image.
5. **imdenoise.m:** program that performs noise reduction using wavelets.
6. **Imagedecomp.m:** function that performs wavelet decomposition of an image, return the approximation coefficients for a specified level.
7. **ann_stat.m:** neural network code that creates and tests the architecture with statistical features as inputs.
8. **test_net.m:** function which computes and returns a confusion matrix and performance rate.
9. **statistics.mat:** contains the statistical input vectors and targets for ann_stat.mat.
10. **ann_tumVnorm.m:** neural network code that creates and tests the architecture with wavelet features as inputs, to distinguish between normal and tumorous tissue.
11. **tumVnorm.mat:** contains wavelet input vectors and targets for ann_tumVnorm.m.
12. **ann_benVmal.m:** network code that creates and tests the architecture with wavelet features as inputs, to distinguish between benign and malignant tumors.
13. **tumVnorm.mat:** contains wavelet input vectors and targets for ann_benVmal.m.

Database Web Application folder:

1. **index.php:** script that displays the login screen
2. **loginproc.php:** script that checks to see if the username and password entered are in the database, allows the user to the next stage if they are.

3. **view.php:** script that displays an html table containing all of the relevant information in the database.
4. **viewImage.php:** script that displays the relevant image when the user clicks a link in view.php
5. **upload.php:** script that displays an html form that allows the user to submit information.
6. **addToDB.php:** script that enters the information from upload.php into the database.
7. **logout.php:** script that logs the user out, by destroying the session.