Chapter

2. Background and Related Work

2.1 Breast Cancer

The breast cancer starts in the breast and is dominant in the female breast. It is therefore necessary to understand the anatomy of the female breast.

2.1.1 Anatomy of Breast

The human breast is functionally part of reproductive system and highly complex. It undergoes many phenomenal changes right from birth to menarche and from pregnancy to breastfeeding till menopause. This lateral aspect of pectoral region is supported and attached to the chest wall by ligaments and pectoralis major muscle. It is located vertically between 2\textsuperscript{nd} to 6\textsuperscript{th} rib and horizontally it extends from lateral border of sternum to the mid of axial. The breast is surrounded by superficial fascia and rest on deep fascia. A conical projection called nipple is present at the level of fourth intercostals space [16]. The nipple contains no fat, hairs or sweat glands. Figure 2.1 shows anatomy of the female breast.

![Anatomy of the female breast](image)

Figure 2.1 Anatomy of the female breast (Courtesy: Terese Winslow LLC, U.S. Govt.)
The different components of breast are explained as below.

**Lobes and Milk Ducts:** Each breast consists of 12–20 lobes of secretory tissues as shown in Figure 2.2. Each lobe is a cluster of number of small milk producing glands called as lobules. Each lobe has one lactiferous duct to carry milk towards nipple. Lobes and ducts are radially embedded in connective tissue and adipose of superficial fascia.

**Adipose Tissue:** The major portion of the breast is occupied by the group of fat cells known as adipose tissues. These tissues are spread across 2\textsuperscript{nd} to 6\textsuperscript{th} rib in the chest and are extended from collarbone down to the armpit. Fatty tissue surrounds the breast surface, fills spaces between lobes and determines shape and size of the breast.

![Figure 2.2 Ducts and lobules in the breast (Courtesy: Breast Cancer-Komen Greater New York)](image-url)
The Lymph System: It is an important component of immune system made up of channel of lymph nodes, lymph vessels, nerves etc. The white blood cells and excessive fluid useful to fight against diseases is carried through network of lymph nodes spread across the breast and armpit as well as entire body. It is clinically significant as it filters out cancerous or abnormal cells and keeps them away from normal tissue [17].

2.1.2 Symptoms of Breast Cancer

The human breast is made up of billions of microscopic healthy cells as like other parts of the body. These cells continuously divide, multiply, grow, die and new cells replace the dead cells in an orderly manner. The entire cellular phenomenon is regulated by the genes. Sometimes the cells start behaving abnormally when the changes in the gene called mutations starts taking place. Because of mutations, cells start dividing and multiplying in an uncontrolled and rapidly manner. The progressive abnormal growth of cells forms a tumor. In its early stage, tumor is very small, cannot be felt and shows no symptoms. The abnormality can be picked up with careful observation on the mammogram. Eventually tumor starts growing which can be felt as a lump inside the breast. All such lumps are not cancerous all the time.

There is variety of symptoms caused by different types of breast cancers. The most common symptoms are enlisted below [18].

- Recently developed thick breast tissue or lump different than other tissues
- Sudden change in size, shape, texture with puckering and dimpling of breast
- Red, pitted skin, peeling, scaling, or flaking of skin over entire breast or around the nipple leading to overall change in the appearance
- Swelling all over the breast
- Blood discharging through nipple without squeezing
- Inverted nipple or nipple gets pulled inside
- A lump or swelling under arm or around collarbone
- Pain in the breast or around the armpit

Once any of these symptoms are observed by self breast examination (sensitivity is just 26%), the person should consult the physician for clinical examinations and tests [19]. Mere symptoms do not mean cancer; rather there can be any other reason. A breast lump can be a water body called as cyst which is not cancerous.
2.1.3 Risk Factors of Breast Cancer

In general, one woman in eight faces breast cancer in their lifetime. Risk of breast cancer incidence is higher or lower, for each person. There are certain risk factors which are responsible for increasing the chances of development of the breast cancer. A few factors are hereditary which cannot be avoided but some can be avoided. The environmental, hormonal, and lifestyle related risk factors [20] associated with breast cancer include:

- **Genes:** Those having mutations of BRCA1 and BRCA2 gene
- **Gender:** Being female chances are 100 times greater than men.
- **Age:** possibility of cancer rises with age, especially after 55
- **Early menarche and late menopause:** First menstruation cycle before age 12, and no menopause even after age of 55
- **Inherited risk:** More chances with a close female relative including your mother, grandmother, sister, or daughter. Breast cancer can develop without family history as well
- **Breast density:** Mammograms of dense breast are hard to interpret
- **Previous breast cancer:** Chances of recurrence are more
- **Late age delivery:** Female delivering a child after age 35 has more risk
- **Never being pregnant:** Women who never carried full-term pregnancy
- **Drinking alcohol:** Alcohol consumption in excessive amounts.
- **Hormone therapy:** Persons undergoing progesterone and estrogen medications after menopause
- **Radiation exposure:** Those who undergo radiation treatments on chest at childhood or young age
- **Obesity:** Being obese is always risky
- **Long menstruation History**

2.1.4 Biomarkers of Breast Cancer

Cancer Biomarkers or tumor markers substance found in person’s urine, blood other body fluids or on tumors produced by the body as a response to cancer or its treatment. When the person is diagnosed with cancer, his/her doctor suggests the
testing of cancerous tissues for finding the various biomarkers [21]. Biomarkers help the doctors to decide the course of actions during treatment of cancer [22]. The most common biomarkers of breast cancer are listed below.

**Estrogen Receptor (ER) and Progesterone Receptor (PR):** There are certain special proteins inside and on the surface of breast cells known as cell receptors. These receptor proteins receive instructions from hormones regarding what to do, when to divide, how much to multiply, how long to continue etc. If the growth of cancer cells is fueled by estrogen hormone in the body, it is known as ER-positive. If the growth of cancer cells is fueled by progesterone hormone in the body, it is known as PR-positive. Testing for ER and PR response of cancer cells helps the doctor to decide hormonal or endocrine therapy which may work.

**Human Epidermal growth factor Receptor 2 (HER2):** A varying amount of HER2 protein is found in all breast cells. An unusually high level of HER2 can fuel the rapid growth and breast cancer spread quite faster.

Triple-negative types are about 10-20% of total breast cancers which means that tests for progesterone, estrogen receptors and excess HER2 proteins are negative [23]. These triple-negative breast cancers don’t respond to endocrine therapy medicines as the cancer growth is not fueled by progesterone, estrogen, or HER2 hormones. It is rare, grows and spreads more quickly than others and is difficult to treat. The doctors and researchers are trying to find out new medications that can treat triple negative breast cancer.

### 2.1.5 Stages of Breast Cancer

Breast cancer progresses through five main stages as shown in Table 2.1. The severity of cancer increases from earlier stage to higher one. The nature of cancer being invasive or non-invasive, size of tumor (Tumor T stage), involvement of lymph nodes (Node N stage) and whether it has spread to other organs (metastasis M stage) or not etc. determines the TNM staging [24]. This TNM staging is as shown in description column Table 2.1 and allows the radiologist to define the cancer clearly for further treatment plan.
<table>
<thead>
<tr>
<th>Stage</th>
<th>Location of tumor</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>DCIS</td>
<td>Basement membrane of cells from ducts and lobules by is infected in a limited way. Cancer is limited to lobules and ducts only. It is bilateral in 20-40 of women and more common among young.</td>
</tr>
<tr>
<td>1A</td>
<td></td>
<td>A small invasive tumor: 2 cm wide or less. No spread to lymph nodes.</td>
</tr>
<tr>
<td>1B</td>
<td></td>
<td>Tumor size is less than 2cm and it is closer to lymphatic nodes.</td>
</tr>
<tr>
<td>2A</td>
<td></td>
<td>Tumor is smaller than 2 cm. Starts affecting nearby tissues. Spread towards lymph nodes 1–3cm nearby.</td>
</tr>
<tr>
<td>2B</td>
<td></td>
<td>Tumor size is in between 2-5 cm. Cancer is closer to armpit, not reached to axillary lymph nodes. It’s larger than 5 cm.</td>
</tr>
<tr>
<td>3A</td>
<td></td>
<td>Tumor is beyond 5cm in size. Cancer spread to 4–9 axillary lymph nodes or any breast bone nodes.</td>
</tr>
<tr>
<td>3B</td>
<td></td>
<td>Tumor invades chest wall or skin and may invade up to 9 lymph nodes.</td>
</tr>
<tr>
<td>3C</td>
<td></td>
<td>Cancer is spread over 10 or more internal mammary nodes and lymph nodes near collarbone.</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Tumor grows to any size; nearby and distant lymph nodes are infected. Distant organs are also infected.</td>
</tr>
</tbody>
</table>
2.1.6 Types of Breast Cancer

The breast cancer is divided in two types: non-invasive and invasive. These types and their subtypes are briefly explained below [25]. The figures in the parenthesis show the percentage of cancer incidences.

- **In situ carcinoma (15-20 %)**
  - **Ductal Carcinoma In Situ (DCIS):** It starts developing at the basement membrane of ducts. It does not invade the lining of breast duct and covers 80% of total in situ carcinomas.
  - **Lobular Carcinoma In Situ (LCIS):** It is known as lobular neoplasia cancer and starts developing in the lobules. It doesn’t invade lobules and is found in 20% of total in situ cases. Close follow up is suggested if found.

- **Invasive Carcinoma (70-85 %)**
  - **Ductal Carcinoma (IDC):** Very common cancer found among 79% invasive cases that begins in milk ducts and spread into neighboring tissue in the breast. If not controlled, it can infect other organs in the body.
  - **Lobular carcinoma (ILC- 10%):** Initiates inside lobules. ILC incidence is increasing among post menopause women.
  - **Tubular Carcinoma (6%)**
  - **Mucinous (Colloid) Carcinoma (2%)**
  - **Medullary Carcinoma (2%)**
  - **Papillary Carcinoma (1%)**
  - **Metaplastic Carcinoma (< 1%)**

- **Inflammatory Breast Cancer (IBC):** In this aggressive type, the cancer cells stops the lymph vessels from draining the fluid and white cells to fight against any attack on organs. It reflects through swelling, reddish skin with thick orange pits on warm breast.

Some types of breast cancer which are least common include:

- **Paget disease on nipple:** affects the nipple, areola and skin.
- **Phyllodes tumor:** A very rare cancer inside connective tissue.
- **Angiosarcoma:** A cancer on the lymph or blood vessels.
2.2 Imaging Modalities for Breast Cancer Detection

One cannot prevent the breast cancer but can be aware of its early signs. Some of the signs can be observed on the breast as symptoms. Imaging tests are useful in determining the rough nature and position of abnormality in the breast. A few imaging modalities approved for breast cancer detection includes mammography, ultrasound, MRI and thermal imaging. The work undertaken is focused on mammographic images.

2.2.1 Mammography

Mammography is the most frequently used non-invasive imaging test for the detection of breast cancer. It uses low-energy X-ray dose (usually 21.5-30 keV) of ionizing radiation to capture the picture of breast structure on a film [26]. Advanced full-field digital mammography quickly produces digital images which can be transferred or stored for a longer period. The high quality sharp images can be analyzed carefully for precise location and size of the abnormality. This results into improved detection accuracy with reduced false positives.

According to American College of Radiology, every woman above age of 40 or after menopause should undergo annual screening mammography [27]. Early detection only can help to avoid further consequences of this deadly disease. Following are some notable benefits of life saving mammographic screening test.

- It allows to identify small suspicious lesions showing abnormal growth
- It improves diagnostic accuracy
- It offers no side effects and it is not harmful

2.2.2 Ultrasound

It is the examination of choice in high risk young women and is valuable as a supplementary tool in the assessment of mammographically ‘dense’ breast. At the minimum 7.5 MHz linear array probe should be used [28]. The original role of breast ultrasound is in the differentiation of cystic and solid lesions. The role of ultrasound complements both clinical examination and mammography. Ultrasound plays a significant role in the triple assessment of symptomatic lesions on the dense breast.
2.2.3 Breast MRI

It is a non-invasive medical imaging test to produce a detailed picture of breast abnormality in breast cancer patients with high risk. It is specifically used to clarify the abnormalities which are not seen on mammographic or ultrasound image. However, it is not available everywhere, is expensive and has high false positive rate [29].

2.3 Mammography Views, Density and BIRADS

2.3.1 Mammography Unit

It comes with a package consisting of Digital mammography system, Image processing and diagnosis workstation, and associated safety equipment. A digital mammography system includes X-ray generator, X-ray tube, acquisition workstation/operation console, digital detector, image processing and diagnosis workstation, various softwares. Workstation is capable of displaying multiple images and priors for comparison purposes. It also has multi-modality viewer capability for display of ultra sound, X-ray, digital mammography, MRI, PET, CT on a third color monitor. Dedicated breast imaging software comes along with at least the following functions: magnifying, zooming, panning, windowing, brightness adjustment, contrast adjustment, distance measurement, histogram display, contrast enhancement etc. The software has capability to export unprocessed and processed Full Digital Imaging and Communications in Medicine (DICOM) 3.0 compatible images.

2.3.2 Mammography Views

A typical view of the mammography unit is as shown in the Figure 2.3. With the help of mammography unit, the different views of both left and right breasts are captured and the same are described in Table 2.2.
Table 2.2 Views of mammography

<table>
<thead>
<tr>
<th>View</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medio Lateral Oblique (MLO)</td>
<td>Taken from mid of chest to lateral side with angle of 45°. It is most common and clinically vital as more changes are occurring in upper outer quadrant during cancer growth.</td>
</tr>
<tr>
<td>Latero Medial Oblique (LMO)</td>
<td>Taken from lateral part of body to mid of chest with angle of 45°. It gives best possible view including pectoral muscle and nipple.</td>
</tr>
<tr>
<td>Cranio caudal (CC)</td>
<td>Taken vertical from upper to lower part of breast. It covers entire breast profile including nipple.</td>
</tr>
</tbody>
</table>

The sample MLO and CC views of mammograms are depicted in Figure 2.4. Usually, the radiologists watch the two mammographic views side by side. This means a MLO or CC views of left and right breast of the same patient together. These adjacent views of the two different breasts of the same patients allow radiologists to identify the difference readily. On the other hand, they use another display screen with the CC and MLO views of the same breast of a patient kept side by side. This arrangement allows them to observe the single abnormality on two different views of the same breast.
How does a mammogram look?

The background of the mammogram is black, and the profile of the breast is shown in gray shades. Usually the fatty part is captured as black while the glandular tissue and ligaments connectives appear as white. The dense cluster of glandular and connective tissue appears as bright white shades on mammogram. Finding the abnormalities in such dense breasts is a difficult task. A dense breast is not necessarily prone to cancer while fatty breast may develop cancer. Every person is different and so the mammogram. American College of Radiology has described scale to measure mammographic density (MD) of breast. This scale is termed as Breast Imaging Reporting and Data System (BI-RADS) which is used commonly by the radiologists to define abnormality without ambiguity, classify MD and facilitate concise disease communication. Type A image is of almost entirely fatty breast with lowest density. Breasts of around 40 % of total patients with scattered density of fibro-glandular tissues belong to Type B. Type C breast are heterogeneous density with small obscured mass. Around 10 % of women found with Type D are having a very dense cluster of glandular and connective tissue [30]. The sample images showing measure of density are depicted in Figure 2.5.

The radiologists are expected to observe the mammograms of dense breast very carefully to pick up the abnormality without adding more into the false positives. The radiologists are supposed to prepare a mammogram report for which BIRADS [31] has provided the guidelines as presented in Table 2.3.
Design and Development of Algorithms for Automatic Breast Cancer Detection

2.4 Mammogram Segmentation

Image segmentation is most important common process of separating several constituent components of the scene. It plays a vital role in medical image analysis by identifying meaningful regions of interest (ROI) which indicate the particular disease. These identified ROIs usually belong to a cluster of homogeneous pixels with respect to some criteria pertaining to values of pixels or their relationships. Homogeneity or similarity and discontinuity form the basis of segmentation. Homogeneity of pixels usually is determined using pixel value threshold or mean etc. Discontinuity based methods capture the sudden changes in pixel’s gray levels to divide the image [32].

There are different frameworks adopted by segmentation methods for variety of applications. There methods can be categorized based on thresholding region growing, data clustering, edge contouring etc. Seeded and unseeded are the two main

---

Figure 2.5 Mammographic breast density (a) almost entirely fatty (b) scattered fibroglandular density (c) heterogeneously dense (d) extremely dense

Table 2.3 BIRADS category of abnormality on mammogram

<table>
<thead>
<tr>
<th>Category</th>
<th>Assessment</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Incomplete study</td>
<td>Go for prior studies or additional imaging</td>
</tr>
<tr>
<td>1</td>
<td>Negative</td>
<td>Suggested regular Routine screening</td>
</tr>
<tr>
<td>2</td>
<td>Benign</td>
<td>Suggested regular Routine screening</td>
</tr>
<tr>
<td>3</td>
<td>Probably benign</td>
<td>Suggested follow up to confirm</td>
</tr>
<tr>
<td>4</td>
<td>Suspicious abnormality</td>
<td>Suggested to undergo Biopsy</td>
</tr>
<tr>
<td>5</td>
<td>Highly malignancy</td>
<td>Suggested to take Appropriate action</td>
</tr>
<tr>
<td>6</td>
<td>Known malignancy</td>
<td>Suggested to take Appropriate action</td>
</tr>
</tbody>
</table>
types of region growing methods. The research undertaken is focused on modifying seeded region growing algorithm and it is covered in detail as below.

### 2.4.1 Region Growing Algorithm

The procedure of region growing starts with selecting the initial seed pixel from the image. One has to decide first the number of regions the image to be segmented in. Accordingly, each pixel starts gathering the neighboring pixels to form the region. Each region continues expanding as long as it meets certain similarity criteria. Neighboring pixels are selected by comparing their values with pixel value of seed. The procedure followed by the algorithm is as follows.

1. Select initial pixel as \( C_1, C_2, \ldots C_k \) as seed of regions \( R_1, R_2, \ldots R_k \)
2. Measure the difference between seed with neighboring pixel. If the difference is smaller that threshold, include the pixel in the region
3. Re-compute the boundary of each region \( R_i \)
4. Repeat steps 2 and 3 until all pixels in the image are allocated to their suitable region

Region growing algorithm is an effective way of segmenting mammograms for detecting the breast cancer regions, hence, better diagnosis [33]. Its advantages include simplicity, robustness and no tuning parameters are required. However, selection of initial seed and recursive selection of neighboring pixels are the major stumbling blocks in the performance of the algorithm.

### 2.4.2 Evaluation of Segmentation

The segmentation results into the identified lesions with contour. This lesion must be compared with the lesion marked by the radiologist on the original mammogram image to find out the area overlap. This overlap is assessed by two ways: one subjective or qualitative by the radiologist and other is objective or quantitative using some mathematical computations. The quantitative performance of segmentation is assessed by several metrics. Images on which the ground truth of lesions is marked by the radiologist are labeled as \( T \), while the segmented results of each image are converted into binary images labeled as \( L \), without changing image dimensions or
pixel resolution. The quantitative metrics for assessment of segmentation results include:

**Intersection over Union:** It is the ratio of intersection of two sets with union of those two sets. It is used to compare similarity of two sets and also known as Jaccard similarity index or Jaccard index [34]. It is defined as

$$O(L, T) = \frac{|L \cap T|}{|L \cup T|} = \frac{|L \cap T|}{|L| + |T| - |L \cap T|}$$

(2.1)

$O(L, T) = 1$ if $L$ and $T$ both are empty, otherwise $0 \leq O(L, T) \leq 1$

**Dice Coefficient:** Classically the Dice coefficient [35] is defined as

$$D(L, T) = \frac{2|L \cap T|}{|L| + |T|}$$

(2.2)

$D(L, T) = 1$ if $L$ and $T$ both are empty, otherwise $0 \leq D(L, T) \leq 1$

**Hausdorff Distance:** It is a maximum distance of a lesion $L$ to the nearest point in another lesion $T$ [36] and it is defined as

$$h(L, T) = \max_{l \in L} \left\{ \min_{t \in T} d(l, t) \right\}$$

(2.3)

where $d(l, t)$ is any metric distance (e.g. Euclidean) between pixels $a$ and $b$.

### 2.5 Feature Extraction and Classification

The suspicious lesions are segmented and then evaluated using Jaccard index. The lesions determined as positive are considered for feature extraction and classification.

#### 2.5.1 Feature Extraction

The segmentation stage identifies and delineates the suspicious lesions on the mammograms. The important information, visual contents or data pertaining to these lesions is then captured and quantified in a redundant manner. This quantifiable redundant information represents the lesion and is known as characteristic feature. A feature represents the whole entity entirely instead of its constituent pixels as individual variables. Thus features combine these variables as single entity and reduce
the amount of processing. Thus the feature extraction is a process of describing the large data set completely and accurately without losing important relevant information. A group of features pertaining to single entity or lesion forms a feature vector. Feature extraction yield redundant data for effective analysis of given image. This in turn results in the fast training of machine learning algorithms and facilitates the decision making process such as classification.

Several quantitative methods have been developed for characterizing the size, shape and texture of abnormal lesions on the mammograms. These methods are important for discriminating breast abnormalities. Because of the different pathologies of breast abnormalities, the geometric and textural properties can be used to discriminate between malignant tumors and benign masses. In this research study, geometric parameters such as area circularity, perimeter, circularity ratio, convexity, compactness, roughness, extent, eccentricity, circular density etc. are measured. Additionally, the selected six textural features based on gray level co-occurrence matrix (GLCM) [37] are also considered for better characterization of breast abnormalities.

2.5.2 Classification

Once the lesions are characterized, the task of classifying the lesion into specified class as either benign mass or malignant tumor becomes easy. The classification is achieved through an algorithm namely classifier which organizes unlabeled data into discrete labeled classes. The vague unknown data values are interpreted by the classifier by using some predefined set of rules. The accuracy of the classifier to discriminate the unknown entity is determined by the set of extracted features. Classifiers can manipulate the classification by employing probability estimates. Classifiers play vital role in pattern and object recognition, especially in medical image analysis.

Classification is based on machine learning algorithm (MLA) which makes use of three data sets: training, validation and testing. A MLA is a mathematical form that learns to identify input lesion patterns based on their characteristic features. Once a MLA learns all the known input patterns from the training data, it must be tested on completely unknown patterns belonging to the same set as that of the training data. If the model performs well on the test data then it is considered as a ML model that
generalizes your dataset of interest. Usually, classifier is trained on 'training set', its parameters are fine-tuned based on 'validation set' and finally the classifier performance is evaluated with unknown set. Enough care must be taken to ensure that the known training set and unknown test set are altogether distinct to avoid biasing.

### 2.5.3 K-Fold Cross-Validation

There is no 'one' way of choosing the size of training/testing set and people apply heuristics such as 10% testing and 90% training. However, this heuristic doesn’t guarantee the unbiased classification. K-Fold cross validation is a well accepted statistical method to estimate how accurately a predictive classifier model performs in practice [38]. The given data set is randomized and then partitioned into K almost equal sized groups. Then choose Kth partition for testing and K-1 partitions for training the classifier. Within the training set you can further employ another K-fold cross validation to create a validation set and find the best parameters. And repeat this process K times to get an average of the metric to get rid of classifier 'bias'. One has to make sure that your test set meets the following two conditions:

- It is large enough to yield statistically meaningful results
- It data set is from the same population

Assuming that your test set meets the preceding two conditions, your goal is to create a model that generalizes well to new data. Our test set serves as a proxy for new data. Notice that the model learned for the training data is very simple. This model doesn't do a perfect job - a few predictions are wrong. However, this model does about as well on the test data as it does on the training data. In other words, this simple model does not overfit the training data. The general process of cross validation is explained as follows.

1. Arrange the contents of dataset randomly
2. Divide the randomized dataset into K equal sized partitions
3. For every distinct partition:
   i. Select any one partition as test data set
   ii. Let remaining K-1 partitions as a training data set
iii. Test the classifier model on training set and evaluate it on test set
iv. Store the performance score and repeat the same K times

4. Calculate the average performance using stored scores

The classifier performance can be optimized by selecting the proper value of K. There is no formal rule as such but K=10 and K=5 are selected widely. The value of K should be preferably large enough to ensure the appropriate representation of data set through each of its partitions.

2.6 Classifiers

The following subsections describe the two classifiers used in this research work.

2.6.1 SVM

Support Vector Machine (SVM) is the simple, most popular yet strong classifier for a small set of images [39]. It focuses more on confidence of classification to discriminate the benign mass and malignant tumors using the respective feature vectors as input. The decision function is defined as Eq. (2.4)

\[ f(x) = \sum_{i=1}^{S} y_i \alpha_i K(x_i, x) + b \]  

where \( x_i \in \{-1,+1\} \) represents training vectors, \( S \) is the size of training set, \( \alpha_i \) is the parameter for optimizing the margin, \( K(x_i, x) \) is the kernel function and \( b \) is the regularization parameter for minimizing training error [40]. In this study, radial basis function (RBF) is taken as the kernel function and the same is defined in Eq. (2.5)

\[ K(x_i, x) = \exp \left( \frac{||x_i - x||^2}{2\sigma^2} \right) \]  

where \( ||x_i - x||^2 \) is the squared Euclidean distance between the two feature vectors and \( \sigma \) is a free parameter. The value of RBF kernel ranges between zero and one and decreases with distance. A variety of combinations of \( (\alpha_i, b) \) can be obtained once at fixed interval values of \( \alpha_i \) and then with fixed interval values of \( b \). The problem of
SVM related to trade-off between the large margin and few misclassifications can be solved by optimized combination of \((\alpha_i, b)\) parameters that achieves highest accuracy.

### 2.6.2 \(k\)-NN

K nearest neighbors (\(k\)-NN) is a simple pattern classification algorithm that stores all available cases and classifies new cases based on a distance functions as similarity measure. It is non-parametric algorithm as it never assumes the distribution of underlying data. Hence it is preferred when there is little or no prior information about the data to be classified, especially nonlinear data. The classification is based on how closely the new data resembles with the existing data features. The \(k\)-NN is widely used algorithm which classifies a new feature vector by calculating its distance from all the other feature vectors of the segmented suspicious lesions. The classification accuracy depends on the number of neighbors (k) selected to classify the new feature vector and the choice of the distance. In this study, we study and analyze several distances such as Manhattan distance, Euclidean distance, Minkowski distance, Correlation distance etc. and different values of the nearest neighbor’s parameter k. The values that give the best results and minimize the classification error are chosen for further classification. However, \(k\)-NN stores the entire training dataset which makes it memory intensive [41].

The steps generally used to implement a \(k\)-NN classifier are as follows:

1. Load the given data
2. Initialize the suitable value of k
3. Repeat the following steps for all training data points
   i. Calculate the distance between test data and each row of training data using either of Euclidean, Minkowski, Correlation distance etc.
   ii. Sort the in ascending order the calculated distance values
   iii. Select first k rows from the sorted array
   iv. Select the most frequent class of these rows
   v. Return the value of predicted class

The optimal value of k can be calculated based on the segregation of the training and validation from the initial dataset with the help of validation error curve.
2.6.3 Performance Measurement of Classification

Accuracy is the most common metric to measure performance of classifier which gives percentage of correct classifications yielded by the classifier. This metric is useful in comparing the different classifiers. However, there are some other factors ignored during assessing the performance of classifiers without bias. Actually in given population or data related to any disease, the test conducted may come out with true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) as shown in Figure 2.6. The performance of the classifier can be assessed more honestly by taking into account all of these above numbers.

![Figure 2.6 Metrics using confusion matrix](image)

The common metric based on these numbers are defined as below.

\[
Sensitivity = \frac{TP}{TP + FN} \tag{2.6}
\]

\[
Specificity = \frac{TN}{TN + FP} \tag{2.7}
\]

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}
\]

\[
Positive \ Predictive \ Value = PPV = \frac{TP}{TP + FP} \tag{2.9}
\]

\[
Negative \ Predictive \ Value = NPV = \frac{TN}{TN + FN} \tag{2.10}
\]
The results of the test conducted in a case study of 100,000 patients with 200 actual cancer patients, are depicted in the Table 2.4.

<table>
<thead>
<tr>
<th>Test Status</th>
<th>Disease status ↓</th>
<th>Test Positive</th>
<th>Test Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Diseased</td>
<td>160 (TP)</td>
<td>40 (FN)</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>Patient Healthy</td>
<td>29940 (FP)</td>
<td>69860 (TN)</td>
<td>99800</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30100</td>
<td>69900</td>
<td>100000</td>
<td></td>
</tr>
</tbody>
</table>

The matrix formed with TP, FP, TN, FN is known as confusion matrix. The performance metrics are calculated as below.

- **Sensitivity** = True Positive rate = \(\frac{160}{160 + 40}\) = 80.0%
- **Specificity** = False Positive rate = \(\frac{69860}{69860 + 29940}\) = 70.0%
- **Accuracy** = \(\frac{70020}{100000}\) = 70.02%
- **PPV** = \(\frac{160}{160+29940}\) = 0.53%
- **NPV** = \(\frac{69860}{69860+40}\) = 99.94%

Only sensitivity and specificity are not enough to assess the classifier’s performance. Thus, all the above mentioned metrics together give the much better insights to honestly assess the performance of classifiers.

### 2.6.4 ROC Curve

In the clinical practice, the radiologist has his own way of interpreting the abnormalities on mammograms and the same may vary from radiologist to another. These interpretations are about what type of tissues are positive and what are negative. The varying decisions give rise to a range of thresholds which govern clinical decisions. Thus, sensitivity and specificity depend on the radiologist’s threshold. As the proportions of TP and TN vary, the decision threshold may shift to right or left as shown in Figure 2.7. This reveals that there is always a tradeoff between sensitivity and specificity. A few radiologists may have lose or less stringent and others may have more stringent criteria for an abnormality test. An illustrative example in Table 2.5 may help us to understand the instance.
Table 2.5 Inter observer variations

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Radiologist-1</th>
<th>Radiologist-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>True Negatives</td>
<td>72</td>
<td>85</td>
</tr>
<tr>
<td>False Positives</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>False Negatives</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Sensitivity %</td>
<td>80</td>
<td>60</td>
</tr>
<tr>
<td>Specificity %</td>
<td>90</td>
<td>94.44</td>
</tr>
<tr>
<td>Overall result</td>
<td>High FP: more biopsies; Found more cancers</td>
<td>Less FP: less biopsies; Found less cancers</td>
</tr>
</tbody>
</table>

By perusing the contents of Table 2.5, it is clear that sensitivity and specificity are radiologist specific. Receiver Operating Characteristic (ROC) analysis is an appropriate tool for quantification of impact of inter-observer variations [42]. The solution for this tradeoff is to begin with the threshold 0.0, classify accurately all positive cases, and classify inaccurately all negative cases. The threshold is then moved from 0.0 to 1.0, progressively increasing true positives and decreasing false positives. ROC is a graph of TPR (sensitivity) plotted against FPR (1 – specificity) for each decision threshold. Figure 2.7 depicts a sample ROC curve.

![Figure 2.7 Example of ROC curve](image)

The diagonal line shows the classifier with no discrimination power and the values above the diagonal shows that the classifier is performing. Thus, the above illustration shows how the ROC analysis helps to select optimal classifier model.
Threshold Selection

It is quite natural to select a threshold for a classifier that maximizes true positives while reducing false positives. For a mammographic screening test radiologists usually prefer to get high true positive rate though it results with high false positives. This is because they don’t want to miss a single cancer. On the other hand radiologists prefer to set the threshold to minimize false negatives.

2.6.5 Area Under ROC Curve (AUROC)

ROC curves help us to assess the performance of the classifier over its entire operating range. The most widely-used measure is the area under the ROC curve (AUROC). As depicted in Figure 2.7, the AUROC for a classifier with random guessing is 0.5 while for the perfect classifier it is 1.0. A lower AUROC closer to 0 indicates that the results are opposite to that of expected. A wrong setup of classifier results in a value less than 0.5.

Classifier Comparison

The AUROC is also useful tool in comparing the performance of two or more classifiers. Generally, the researchers select a single threshold at a time and performance of both the classifiers at that point is compared. The overall AUROC also can be considered for comparing performance.

Computing the AUROC

The results of the probabilistic binary classifier are tabulated in the form of confusion matrix as shown in Table 2.6. Assume that among the 50 cancer patients that are classified, 45 are correctly classified and the 5 are misclassified.

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class 1</td>
</tr>
<tr>
<td>Class 1</td>
<td>10 true positives (TP)</td>
</tr>
<tr>
<td>Class 0</td>
<td>3 false positives (FP)</td>
</tr>
</tbody>
</table>
The sensitivity (true positive rate) and specificity (false positive rate) is calculated at many different threshold (for example 0.00;0.01,0.02,…,1.000.00;0.01,0.02,…,1.00) for the given classifier. A graph of sensitivity on vertical axis and specificity along X axis is plotted. The resulting curve is called ROC curve, and the area covered by this curve is known as AUROC. Figure 2.8 shows the AUROC graphically.

![Receiver operating characteristic example](image)

**Figure 2.8 Area under ROC curve**

As shown in Figure 2.8, the blue area corresponds to the Area Under the curve of the Receiver Operating Characteristic (AUROC). The dashed line in the diagonal we present the ROC curve of a random predictor: it has an AUROC of 0.5. The random predictor is commonly used as a baseline to see whether the model is useful. The results of AUROC comparison are quoted in terms: “The AUROC of Classifier 1 is 0.87, and of classifier 2 is 0.83, so it is clear that classifier 1 performs better. The P test conducted shows that the differences between AUROC of these two classifiers are statistically not significant.”

### 2.6.6 Confidence Interval

The classifier performance with increased confidence significantly impacts its success or failure in many applications. Classifier evaluation involves applying one or more evaluation metrics to the confusion matrix. The criterion widely used to measure the ranking quality of a classification algorithm is the area under an ROC curve (AUROC). To measure and report the AUROC properly, it is crucial to determine an
interval of confidence for its value as it is customary for the error rate and other measures. Confidence intervals are usually interpreted as margin of errors because they provide magnitude and precision [43]. It is also important to make the computation of the confidence interval practical by relying only on a small and simple number of parameters. Classifier evaluation involves estimating to what extent, if any, the resulting classifier is capable of predicting the class labels of the unseen instances. In the context of the above control-studies, the issue is measuring the difference between distributions of classifier predictions and class labels. In case-control studies, classification testing examples undergo two classifications, labels and predictions. The issue of evaluation becomes: at a particular level of confidence (95%), is there a statistically significant difference between the error proportions of classifications. One way to assess this ability is to measure the difference in error proportions between instance labels and classifier predictions by using Tango’s test. Medical statisticians prefer the use of confidence intervals rather than p-values to present results. Confidence intervals have the advantage of being close to the data and on the same scale of measurement, whereas p-values are a probabilistic abstraction. Hence, p values are calculated in this work.

2.7 Commercial CAD Systems

There are a few commercial CAD systems widely used by the radiologists. The performance of two CAD systems tested on the set of 126 mammographic digital cases, having been independently diagnosed by two senior radiologists is depicted in Table 2.7 [44]. The researchers have used five different setting levels such as high sensitivity, normal sensitivity, standard, normal specificity and high specificity.

<table>
<thead>
<tr>
<th>Parameter / CAD System</th>
<th>Cyclopsus CAD</th>
<th>SecondLook</th>
</tr>
</thead>
<tbody>
<tr>
<td>version</td>
<td>mammo (v. 6.0)</td>
<td>(v. 6.1C)</td>
</tr>
<tr>
<td>Produced by</td>
<td>Cyclopsus CAD Ltd. Italy</td>
<td>iCAD Inc. (OH, USA)</td>
</tr>
<tr>
<td>Overall sensitivity</td>
<td>83.1 %</td>
<td>66.2%</td>
</tr>
<tr>
<td>Average FPs/l</td>
<td>1.38</td>
<td>0.47</td>
</tr>
<tr>
<td>sensitivity for mass lesions</td>
<td>76.9% @ 0.73 FP/im</td>
<td>61.5% at 0.28 FP/im.</td>
</tr>
<tr>
<td>Sensitivity of microcalcifications</td>
<td>76.2% @ 0.64 FP/image</td>
<td>61.9% @ 0.19 FP/im.</td>
</tr>
</tbody>
</table>
2.8 Available Databases

The clinical data in this study is taken from Tata Memorial Centre (TMC), Mumbai, India. The usage of all the images of the biopsy-proven breast cancer patients is approved by the Institutional Research Ethics Committee of TMC, Mumbai. The identity of all the patients is anonymized before analysis by the researchers. This dataset consists of 360 FFDM images including 180 CC views and 180 MLO views of left and right breast of 90 randomly selected cancer patients. The patients are screened on ‘Hologic Selenia System’ and on ‘GE Medical Senograph System’ leading to images in DICOM format. Specifically, the dataset includes 160 images (CC and MLO views of 80 breasts of 80 patients) with verified malignant tumors, 20 images (CC and MLO views of 10 breasts of 10 patients) with benign mass and 180 (CC and MLO views of 90 breasts of 90 patients) images are of normal breasts. Three patients have malignant tumors along thoracic wall and pectoral muscle. Four senior radiologists have outlined the lesions on CC and MLO views creating Ground Truth (GT) files of respective mammograms. These GTs are confirmed using Histo-Pathological Reports (HPR) of the breasts of the respective patients. The opinions recorded by the radiologists are used to validate and evaluate the proposed algorithms. The 130 images from Mammographic Image Analysis Society (MIAS) dataset are also included for some comparison purpose [45]. Each of these mammographic images is having a size of 1024x1024 pixels with 8 bits per pixel. The spatial resolution of each pixel is 200mm per pixel. Another dataset of randomly selected 74 patients with CC and MLO views of either left or right breast is taken from widely used publicly available benchmark Digital Database for Screening Mammography (DDSM) dataset [46]. This first DDSM dataset in our study consists of 50 images with malignant tumors (25 patients), 10 image with benign mass (5 patients) and 40 images of normal breasts (20 patients). Another set consists of 148 mammograms selected from DDSM dataset. This set comprises 38 pairs of mammograms with benign masses and 36 pairs of mammograms with malignant tumors visible on both CC and MLO views. The lesions on all these images are delineated by the radiologists in the DDSM dataset.
2.9 Chapter Summary

This chapter has described the anatomy of the breast, symptoms, risk factors, stages, and types of breast cancer. The different imaging modalities used for identifying breast cancer are explained briefly. The detailed description about mammography, views of mammograms and BIRADS categories etc. are then covered. The brief information about segmentation, region growing algorithm, feature extraction, classification, K-fold cross-validation, $k$-NN, SVM, ROC, confidence interval etc. are included for readers ease. The chapter ends up with available commercial CAD systems and datasets used for experimental work.