Chapter 4

Dataset

4.1. Introduction
Dataset is a collection of similar and related data stored for processing. Further this can be defined as a collection of data that contains individual data units organised in a specific format. A data set generally contains a collection of many types of data. Medical dataset can be defined as a collection of pieces of information, especially those that are part of a collection to be used in an analysis of a problem, such as the diagnosis of diseases. These pieces are stored in a specific format and can be used for future research and training purpose or may be processed by a CAD.

4.2. Standardisation
Medical images of the human body or parts of the body, intended for clinical purposes for revealing or diagnosis of disease in medical science defines medical imaging. Digital X-Ray, mammogram, Ultrasound (USG), Computed Tomography (CT), Magnetic Resonance Imaging (MRI) are some well accepted imaging techniques used for clinical diagnosis. Each of these imaging techniques results in images that have specific characteristics. Moreover, images obtained also depend on the make of the imaging device, parameters set by the operators and individual characteristics of the patients. Generalised feature extractions
from such images are difficult as they involve diverse technologies. The only commonality among all images is the intensity features exhibited by them. DICOM (Digital Imaging and Communication in Medicine) image format is universally well accepted for all above imaging technologies. In general, the greyscale image pixel is represented by 8 bits in DICOM format having $256 (2^8)$ greyscale colour intensities.

4.3. Experimental Datasets

In this research, two types of datasets are used. The mammographic datasets are needed to experiment and establish the medical image processing algorithms that are proposed in this thesis related to breast cancer screening like abnormality detection, asymmetry analysis, volume calculation, density estimation etc. On the other hand the datasets containing digitised Histopathological slides of biopsy are applied to those pre-processing algorithms which are related to the confirmatory test of breast cancer.

4.4. Mammographic Datasets

The mammographic datasets are collection of digitised mammographic slides of patients consisting of different age groups with normal and different types of abnormal slides in a specific file format. The benchmarking of an algorithm can be drawn by using a standard test dataset to directly compare the results. Very few mammographic datasets are available in public domain for academic research activity. The most easily, well organised and the most commonly used datasets are the Mammographic Image Analysis Society (MIAS) dataset, the Digital Database for Screening Mammography (DDSM) dataset and Lawrence Livermore National Laboratories (LLNL) dataset. Besides, there are currently few projects developing new mammographic image datasets as well. The proposed algorithms have used the aforesaid three dataset to obtain the experimental result and estimate the
accuracy of the same. In the standard mammographic datasets two different views of images are common namely medio-lateral oblique (MLO) and cranio caudal (CC) views. The medio-lateral oblique (MLO) view is popular among the researchers than cranio caudal (CC) due to more tissues are visible in this projection. But both are equally important when volumes are to be extracted. The datasets which are used in the dissertation are briefly described below to justify their potentiality as standard dataset to be used for experimental purpose.

4.4.1. **Mammographic Image Analysis Society (MIAS) Dataset**

The Mammographic Image Analysis Society (MIAS) is an organisation of UK research groups interested in the understanding of mammograms and has developed a database of digital mammograms. Films taken from the UK National Breast Screening Programme have been digitised to 50 micron pixel edge with a Joyce-Loebl scanning microdensitometer, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains 322 digitised films and is available on 2.3GB 8mm (Exabyte) tape. It also includes radiologist's "truth"-markings on the locations of any abnormalities that may be present. The database has been reduced to a 200 micron pixel edge and padded/clipped so that all the images are 1024*1024. Mammographic images are available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex [262][161].

The 322 MLO mammogram images are organised in pair where each pair represents the left breast with even filename numbers and right breast with odd filename numbers of a single patient. The mammograms are classified according to the character of background breast tissue i.e. Fatty, Fatty-glandular and Dense-glandular tissue. The database consists of all sorts of common abnormalities to analyse. The abnormalities are calcification, well-
defined/circumscribed masses, spiculated masses, ill-defined masses, architectural
distortion and asymmetry. The severity of abnormality is also mentioned i.e. benign and
malignant. The locations of abnormalities are also indicated in the database.

4.4.2. The Digital Database for Screening Mammography (DDSM)
The Digital Database for Screening Mammography (DDSM) is a repository for use by the
mammographic image analysis research community. The project was supported by a grant
from the Breast Cancer Research Program of the U.S. Army Medical Research and Materiel
Command. The DDSM project is a joint work involving Co-Principal Investigators at the
Massachusetts General Hospital (D. Kopans, R. Moore), the University of South Florida (K.
Bowyer) and Sandia National Laboratories (P. Kegelmeyer). Additional cases from
Washington University School of Medicine were provided by Peter E. Shile, MD, Assistant
Professor of Radiology and Internal Medicine. Additional collaborating institutions include
Wake Forest University School of Medicine (Departments of Medical Engineering and
Radiology), Sacred Heart Hospital and ISMD, Incorporated.

The primary purpose of the database is to facilitate advanced research in the development
of computer algorithms to aid in screening. Secondary purposes of the database may
include the development of algorithms to aid in the diagnosis and the development of
teaching or training aids. The database contains approximately 2,500 studies. Each study
includes two images of each breast, along with some associated patient information like
age at time of study, ACR breast density rating, subtlety rating for abnormalities, ACR
keyword description of abnormalities and image information. Images containing suspicious
areas have associated pixel-level "ground truth" information about the locations and types
of suspicious regions [102][101].
4.4.3. LLNL/UCSF Dataset

Lawrence Livermore National Laboratories (LLNL) along with University of California at San Francisco (UCSF) radiology department has developed a mammogram database. The library contains 198 films from 50 patients each containing 4 views per patient containing CC and MLO views of left and right breast, selected among a wide range of subjects of interest. These films were digitised to 35 microns and all pixels were sampled to 12 bits of greyscale. For each digitised film image, they provided 2 associated "truth" images. They also provided a file with case history, radiologist's comments and other information.

The films were selected to present 5 normal, average, healthy cases with previous normal mammograms and no history of ultrasound, magnification views, biopsy, etc., 5 normal but difficult cases with either dense or fibrous breasts, implants or asymmetric tissue, 20 cases of obviously benign micro-calcifications with at least 3 years of follow-up without change or developing cancer, 12 cases of suspicious, benign micro-calcifications and 8 cases with a malignant cluster of micro-calcifications, biopsy proven [284].

Instead of these three organised datasets, the proposed algorithms are also tested with other unorganised data sources like downloaded information from internet or live data received from medical organisations which are available in public domain not violating the privacy of patients and any other legal infringements. The experts in this related field, suggested that the numbers of test cases are sufficient to prove the accuracy and efficiency of the novel algorithms which are proposed in later chapters to identify different types of abnormalities.
4.5. Histopathological Datasets

Later part of the dissertation is dealing with Histopathological slide for confirmation testing. The abnormalities can be identified by the screening methods but confirmation can only be drawn through analysing the Histopathological slide produced by the biopsy procedure. If Histopathological slides are not considered, the CAD system will be restricted only up to the detection phase. The CAD system will become a diagnostic tool only if it is capable of handling the Histopathological slide images also.

The proposed dissertation has also tried to put some light on the true diagnosis phase by analysing the Histopathological slide image generated from clinical biopsy. Although, it is introductory and simple in nature but it has immense scope for future researches. In this regards, the freely downloadable biopsy slide available in OriGene and other sites are used to evaluate the proposed algorithms [202].

4.6. Discussion

Among the three mammographic database discussed above MIAS dataset is sufficiently large to conduct experimental analysis. Moreover, the dataset contains 322 mammogram images of different size, shape and morphology. The images are also classified and benchmarked by an expert team of radiologist of MIAS into three distinct categories based on their parenchymal density. Further, the radiologist have provided all relevant information regarding any abnormality present. The images are classified into normal, asymmetry, presence of mass, calcification etc. The mass is further classified according to the type of mass present, like, well-defined/circumscribed masses, spiculated masses, ill-defined masses etc. The choice of MIAS database for conducting most of the experiments described in this thesis is due to its sufficiently large volume, image quality, easy availability and diverse benchmarked cases. Most of the important international research work in this
area is conducted using MIAS dataset as evident in most of the dissertations. So MIAS
dataset have been used to prove the efficiency of my proposed methods, quantitative
assessment and to compare the results with other important dissertations.

The limitation of MIAS dataset is that it contains only one perspective view of mammogram
images, namely, MLO view. To conduct experiments on volume calculation MLO view alone
is insufficient and the other perspective view i.e. CC view is also required. So, LLNL/UCSF
dataset is used for volume calculations which contains images of both views for a particular
patient.

The MIAS dataset has categorised images into three categories based on the parenchymal
density as stated above. The internationally acceptable ACR BI-RADS scale that divides the
parenchymal density of mammogram into four categories has been used to measure the
density in this thesis. The DDSM categorises the images into four categories of ACR BI-
RADS. To conduct experimental analysis of density determination, volumes of normal
dataset from DDSM is used along with MIAS dataset.

4.7. Conclusion
The dataset has immense contribution towards development of effective and accurate
computer aided diagnostic system. Well organised and documented dataset are very rarely
available for academic research. In India, it is almost absent in public domain. The entire
research work conducted by me, the benchmarked dataset has been used to rigorously test
the proposed algorithms. A collection of digitized mammogram images, provided by the
radiologist with feature identification done by him have been used as test data. The sole
purpose of using this data is to make the system more robust by handling raw data directly
from the source. Since the use of data requires confidentiality so output of results could
not be presented in this thesis and cannot be included in the dataset mentioned above. In conclusion, thanks may be given to all organisations, who generously offer the well organised freely available dataset for academic research and betterment for health care facilities.