61

Databases for Mammography

Maria Evelina Fantacci

CONTENTS
61.1 Introduction ................................................................. 1249
61.2 MIAS Database ............................................................... 1250
61.3 DDSM Database ............................................................. 1250
61.4 MAGIC-5 Database .......................................................... 1252
61.5 MammoGrid Project ......................................................... 1253
61.6 BancoWeb LAPIMO Database ............................................. 1253
61.7 BCDR Database ............................................................... 1254
61.8 INbreast Database ............................................................ 1254
61.9 LLNL/UCSF Database ....................................................... 1254
61.10 AMDI Database ............................................................. 1255
61.11 IRMA Project ................................................................. 1255
61.12 The DBMS ................................................................. 1255
61.13 Conclusions ................................................................. 1257
References ............................................................................. 1257

61.1 Introduction

A medical image dataset is the starting point for important epidemiological and statistical studies. In fact, it can be used to develop and test algorithms for computer-aided detection (CAD) systems, as the development of a CAD system is strictly related to the collection of a large dataset of selected images, for teaching and training medical students or as an archive of rare cases (Tangaro et al. 2008) (see Section IV, Chapter 59). This is true also for mammography, in particular after the worldwide spread of mammographic screening programs (see Section IV, Chapter 60). When the evaluation of a CAD algorithm begins with a retrospective evaluation of cancer cases (Heath et al. 2001), such preliminary evaluation is more time and cost effective than a prospective evaluation in a clinical setting. The task of obtaining the data for a retrospective CAD performance evaluation at a mammography center may be time consuming and expensive to achieve. When investigators of CAD methods utilize the resources in a common database, rather than using their own data, this expense may be decreased, and much more may be learned by a performance evaluation. In fact, provided the same data, the same performance measure and the same train and test methodologies are followed, results from different algorithms can be compared to find the relative strengths of each algorithm. This led to the development of new or combined approaches to the problem that yield superior performance (Brake and Karssemeijer 1998). The issues related to the production of a mammographic database to be used for research purposes were previously discussed, in 1993, by one of the firsts working group in a panel session at the IEEE Biomedical Image Processing and Biomedical Visualization Conference (Stern 1993).

In 1995, Osuch et al. (1995) proposed a mammography database for a national mammography audit and to monitor patients through a centralized system. Then, the Radiological Society of North America Medical Imaging Resource Center (RSNA MIRC) project (MIRC 2016) proposed a highly generic system to store and publish medical images, primarily for research and teaching support, that can federate a large number of remote databases and make them accessible as a single one.

Mammographic databases (Nishikawa 1998; Moreira et al. 2012) should take into consideration five fundamental requirements: case selection, ground truth, requirements of the digitizers (in the case of databases made with digitized films), organization of the database, distribution of the database. Case Selection: the database should include various cases comprising cases without findings, cases with all the possible types of findings and also all types of breast densities. Normal images with structures that may be misleading (e.g., superimposed tissue that looks like a mass) are important in order to make the classifiers more robust. Also, a specialist experienced in mammography should collect the cases and each case should contain the four standard views. It is considered that, for every 100 cases, approximately 200 images should contain a lesion. Ground Truth: biopsy proof for all cases should be available and, for cases in which a biopsy is not recommended, the mammography should have the same breast imaging reporting and data system (BI-RADS™) (Eberl et al. 2006) for at least 3 years. Annotations should include the
”ground truth” concerning the degree of malignancy, the location and the boundary of the lesion and a specialist should perform this outline. Associated Information: clinical history (e.g., age, family history, previous biopsies) can be useful to improve the performance of CADs. Requirements of the digitizer: one common approach is to digitalize at a very small pixel size, at 50 or 25 \( \mu \text{m} \). Organization of Database: medical images are usually saved in the DICOM (Digital Imaging and Communications in Medicine) format that gathers not only the image, but also some related metadata. A division of the images on training and test sets should also be suggested, in order to have comparable sets, and different methods can be compared. Distribution of Database: the database should be available, preferably over the World Wide Web. Continuous user support is also indispensable.

The first annotated mammographic databases publicly available for research purposes date back to the 1990s. Technological improvements in digitizing scanners made it possible to digitize radiographic films, with no significant loss of information, and the increasingly widespread dissemination of mammography screening programs and the advent of digital mammography gave an important boost to the realization of such databases. Moreover, the increasing availability of computer technologies (in particular grid and then cloud computing) has made possible the management and analysis of ever increasing amounts of such data. So, presently, many large datasets of digitized mammograms, among which are grid-based databases, are available on the Web or by contacting the developers.

### 61.2 MIAS Database

In the 1990s the Mammographic Image Analysis Society (MIAS) produced a Digital Mammographic Database (Suckling 1994), available for research purposes, containing 161 pairs of films, including both examples of abnormalities commonly encountered in screening and normal cases. The mammograms have been acquired in the framework of the UK National Breast Screening Program. The medio-lateral oblique view has been digitized by means of a microdensitometer, with a linear response in the optical density range of 0.0–3.2. The pixel size is 50 \( \mu \text{m} \) and the gray level resolution is 8-bit. The images were compressed, allowing the entire database of 322 digitized films to occupy less than 2 Gbytes. A copy of the database and a single-site license were available for research purposes for a small fee, covering duplication costs. The design criteria (mostly the limited number of mammograms coming from a single center and corresponding to only the medio-lateral oblique view) of this database were not ideal, due to both practical and financial constraints, but it has the merit to being the first research database of mammograms publicly and easily available. Therefore, it has been used to validate CAD systems since the late 1990s (Ibrahim et al. 1997). The current release, containing the original images in “Portable Gray Map” (PGM) format and associated truth data, is available at MIAS (2016).

### 61.3 DDSM Database

The Digital Database for Screening Mammography (DDSM) (Heath et al. 1998, 2001; USF 2016) is a database of digitized filmscreen mammograms with associated ground truth and other information. The purpose of this resource is to provide a large set of mammograms in a digital format that may be used by researchers to evaluate and compare the performance of CAD algorithms (Shi et al. 2008; Hapfelmeier and Horsch 2010). The database was completed in 1999. It contains 2620, four view, mammography screening exams. Since that time, the database has been enhanced through the addition of new software tools that simplify the extraction of image data to various file formats, allow simplified access to the data in the ground truth files and simplify the steps necessary to evaluate a CAD algorithm. The DDSM contains mammograms obtained from the Massachusetts General Hospital, the University of South Florida and Sandia National Laboratories. Additional cases were provided from Washington University School of Medicine. Additional collaborating institutions include Wake Forest University School of Medicine (Departments of Medical Engineering and Radiology), Sacred Heart Hospital and ISMD, Incorporated. The four standard views (medio-lateral oblique and craniocaudal) from each case were digitized by means of the scanners DBA M2100 ImageClear (42 \( \mu \text{m} \), 16 bits), Howtek 960 (43.5 \( \mu \text{m} \), 12 bits), Lumisyys 200 Laser (50 \( \mu \text{m} \), 12 bits) or Howtek MultiRad850 (43.5 \( \mu \text{m} \), 12 bits). The cases were all from mammography exams conducted between October 1988 and February 1999. The DDSM is organized into “cases” and “volumes.” A “case” is a collection of images and information corresponding to one mammography exam of one patient. A “volume” is a collection of cases collected together for purposes of ease of distribution. The volumes are available online. The README file explaining “everything” about the database is available. The cases were assigned to volumes according to the severity of the finding. Normal volumes contain mammograms from screening exams that were read as normal and had a normal screening exam 4 years later (±6 months). Benign without callback volumes contain exams that had an abnormality that was noteworthy, but did not require the patient to be recalled for any additional work-up. Benign volumes contain cases in which something suspicious was found and the patient was recalled for some additional work-up that resulted in a benign finding. Cancer volumes contain cases in which a histologically proven cancer was found. Every case in the DDSM contains the patient age, the screening exam date, the date on which the mammograms were digitized and the ACR (American College of Radiology) breast density that was specified by an expert radiologist. Cases in all volumes other than the normal volume contain pixel level ground truth markings of abnormalities. Each marking contains a subtlety value and a description that was specified by an expert mammography radiologist using the BI-RADS (Eberl et al. 2006) lexicon. The web page for the DDSM resource is at USF (2016) and some examples of available cases in the categories “normal,” “benign” and “cancer” are reported in Figures 61.1 and 61.2. Several software tools are linked to the main web page for the database. These tools simplify the most common tasks, such as case selection, data extraction and performance evaluation.

A search engine is available that uses a web interface to obtain a list of DDSM cases that meet search criteria defined by the volume type, ACR breast density, scanner, lesion pathology, assessment and subtlety, as well as by the BI-RADS keywords. A software package, which contains programs and “C” source
FIGURE 61.1 Examples of normal (left) and benign (right) images available in the DDSM database.

FIGURE 61.2 Examples of cancer cases available in the DDSM database.
code for decompressing images, converting them to 8 or 16 bits, rescaling them, re-mapping them to optical density, resizing them to other resolutions and for writing them in Portable Gray Map (PGM) or Tagged Image File Format (TIFF) files, is available through the database. This same package also provides a program for rendering an image of user selectable ground truth regions at a user specified resolution.

The DDSM has been extensively used by the research community. It is maintained at the University of South Florida for purposes of keeping it accessible on the web. Additional functionality for DDSM has been created by other research groups. The web services for the DDSM (Rose et al. 2006) have been developed to overcome two particular shortcomings: (1) the actual mam-mographic image data was encoded using a non-standard lossless variant of the JPEG image format; and (2) although detailed metadata was provided, it was not in a form that permits it to be searched, manipulated or reasoned over by standard tools. The web services realized will allow an individual to query for, and obtain, mammograms from the DDSM in a standard and well-supported image file format and to use the database within grid-based workflows, allowing digital mammography researchers to make use of distributed computing facilities.

### 61.4 MAGIC-5 Database

The MAGIC-5 database (Bagnasco et al. 2004, 2005; Tangaro et al. 2008) is a mammographic database of digitized images for research developed in the framework of INFN (Istituto Nazionale di Fisica Nucleare, Italy) experiments. The digitized images, and also a small number of digital ones, were collected, beginning in 1999, by a community of physicists, in collaboration with radiologists in several Italian hospitals (Bari, Udine, Palermo, Firenze, Torino, Napoli, Sassari) as a first step in developing and implementing a CAD system (Bellotti et al. 2006; Cascio et al. 2006). This consists of 3369 mammograms collected from 967 patients and classified according to lesion type and morphology, breast tissue and pathology type. A dedicated graphical user interface was developed to visualize and process mammograms to support the medical diagnosis directly on a high-resolution screen. Images were acquired in various mammographic centers using different mammographic screen/film systems and settings (all with molybdenum anode) in the framework of different applications, including both clinical routines carried out on asymptomatic women and screening programs addressing asymptomatic women. A workstation, composed of a personal computer (PC) running the Linux operating system, a film scanner and a dedicated disk, was installed at each site involved in the program. The parameters of the charge-coupled device (CCD) scanners used were! a pixel size of 85 µm and a 12-bit depth (4096 gray levels). All digitized mammograms are available in DICOM file format. Each exam is stored in a directory, which contains one DICOM file per image, including a collection of tags, values, types, length and value fields describing patient information, imaging procedure information and other image-related information. The diagnosis is provided according to classification criteria proposed by the American College of Radiology (Eberl et al. 2006) and, if required, is available as a text file. The image-related information includes image type (scanned/digital), study date, study identification (ID), series description, image laterality, view position, series number, pixel pitch and name and address of the institution where taken. When available, information about the device manufacturer, acquisition and calibration parameters (model name, tube voltage and current, anode features) is also provided. About 60% of patients in this database are more than 50 years old. Each exam contains one-to-six views. The repartition of the database in left/right breast images is 1835 (51%) and 1734 (49%), respectively, whereas for the craniocaudal/oblique/lateral views, it is 1601 (48%), 1456 (43%) and 312 (9%), respectively. Image size is 2067 × 2657 pixels, 85 µm of pitch (300 dots/inch). There are images from 306 (32%) patients who were defined as normal when there was no evidence of any lesion (confirmed by 3 years of radiological follow-up). The remaining images are from 661 (68%) “abnormal” patients: when a suspicious lesion was found by the radiologist in these images, it was classified as suspicious, benign or malignant. For all malignant lesions, cytological or histological results are also available. Detailed radiological annotations of abnormalities are included in the database as notes. The relative distribution of malignancy grade is 560 (35%) suspicious lesions, 468 (29%) benign lesions and 592 (37%) malignant lesions. Images that contain at least one mass or a cluster of microcalcifications, as diagnosed by an expert radiologist, are considered abnormal. There are 1062 images containing at least one region of interest (ROI) with a massive lesion and 304 images containing at least one ROI with microcalcifications. In total, there are 1296 (38%) abnormal images containing at least one lesion (massive or microcalcification or both) and 2073 (62%) normal images with no lesions. Each image can also contain more than one lesion, so the total number of ROIs is 1620 (1236 massive and 384 microcalcification). Each of these main classes of lesions (microcalcification clusters and massive lesions) is further classified according to the morphological characteristics of the lesion. The scheme of Lattanzio and Guerrieri (1998) has been adopted, which has been recognized as a satisfactory reference framework by an Italian panel of radiologists, with more than 20 years of experience in mammography, who identified and localized each lesion according to this classification. Each abnormal image comes with a description of the lesion. Mass location and size is defined by a radiologist-drawn circle, characterized by center coordinates (Xrad; Yrad) and radius (Rrad), which fully contains the mass. Mass radii range from 3.1–47.2 mm, with an average value of 11.7 mm, whereas the radius of the microcalcification clusters ranges from 1–72.8 mm, with an average value of 11.9 mm. Another important parameter to characterize the image is breast tissue type. Collaborating radiologists were asked to identify breast texture for a full-image characterization. A tissue classification recognized as a standard by many Italian radiologists (Lattanzio and Simonetti 2002) has been adopted: fibroadipose tissue indicates a fatty breast with little fibrous connective tissue (dense tissue percentage <25%); glandular tissue indicates the presence of prominent duct patterns (dense tissue percentage between 25%–75%); dense tissue indicates a dense breast parenchyma (dense tissue percentage >75%). The database presents some limitations, especially from an epidemiological point of view, mostly because images were collected in different clinical and screening conditions, so they do not represent a typical distribution of masses and microcalcifications in terms of the ratio
of benign to malignant cases and because they were collected from different centers and were acquired with different mammography units under different conditions, but it represents the largest Italian sample of mammograms. As in a screening program, data are collected from geographically remote sites. The growth of the database and the distributed nature of the collaboration highlighted a problem, however, as images were generally not replicated between remote sites. The approach used to solve the problem of remote access was to use techniques developed for grid computing (Bagnasco et al. 2004). Its integration in a grid computing environment also made the implementation of several remote-analysis-use cases possible, which have demonstrated functionalities potentially useful in screening programs and teletraining.

### 61.5 MammoGrid Project

The MammoGrid database (McClatchey et al. 2003, 2004; Amendolia et al. 2004) has been realized in the framework of the EU-funded MammoGrid project and contains digitized films from Udine (IT) and Cambridge (UK) and digital mammographies from Torino (Valdese, IT) hospitals, the pilot radiological institutions in this project. As grid computing promised to resolve many of the difficulties in facilitating medical image analysis to allow clinicians to collaborate without having to co-locate, the MammoGrid project has investigated the feasibility of developing a Grid-enabled European database of mammograms so that a set of important healthcare applications using this database have been enabled and the potential of the Grids has been harnessed to support co-working between healthcare professionals across the EU. Among the aims of having a Grids-enabled European-wide MammoGrid database were: to evaluate grid technologies and determine the requirements for Grid-compliance in a pan-European mammography database; to implement the MammoGrid database, using novel Grid-compliant and federated-database technologies that will provide improved access to distributed data and will allow rapid deployment of software packages to operate on locally stored information; to deploy enhanced versions of a standardization system (SMF, Standard Mammogram Form) that enables (Warren et al. 2007) comparison of mammograms in terms of intrinsic tissue properties independently of scanner settings; to develop software tools to automatically extract image information that can be used to perform quality controls on the acquisition process of participating centers (e.g., average brightness, contrast); to develop software tools to automatically extract tissue information that can be used to perform clinical studies (e.g., breast density, presence, number and location of microcalcifications [Reticò et al. 2006]) in order to increase the performance of breast cancer screening programs (Warren et al. 2007); to use the annotated information and the images in the database to benchmark the performance of the software previously cited; to exploit the MammoGrid database and the algorithms to propose initial pan-European quality controls on mammogram acquisition and ultimately to provide a benchmarking system to third party algorithms. The MammoGrid implementation was based on AliEn, a Grid framework developed by the ALICE Collaboration at CERN (CH). AliEn provided a virtual file catalogue in order to allow transparent access to distributed datasets and to provide top to bottom implementation of a lightweight Grid applicable to cases when handling of a large number of files is required.

The MammoGrid project proved that Grid infrastructures can be used for collaborative clinical analysis of database-resident, but geographically distributed medical images. Although the creation of a database was not the primary objective of this project, the database that was realized is significant also, because the images have been preprocessed by the SMF (Warren et al. 2007) tool. The SMF tool, which was implemented in the MammoGrid framework, is a fully automated, objective measurement tool to estimate the volume of glandular tissue in the breast from a mammogram which also compensates for variations in X-ray imaging techniques and is able to standardize images scanned on different film digitizers. It explicitly considers breast compression, exposure and tube voltage and incorporates a full physics model rather than using step-wedges in each image.

### 61.6 BancoWeb LAPIMO Database

The BancoWeb LAPIMO database (Nepomuceno Matheus and Schiabel 2011) is a public online mammographic images database of digitized films developed at São Paulo University (Brasil). It is free for all interested viewers and aims to help develop and evaluate CAD schemes (Patrocínio et al. 2004) which are available online at LAPIMO (2016), requiring only a free subscription for full access. The first release of the database (2011) had around 1400 images, from around 320 patients. Most of these images are screening mammography, with only a few being diagnostic work-up. All images are associated with their corresponding exam’s medical reports with several types of findings. The database also contains images corresponding to normal cases, that is, without any findings indicated on the reports. About 32% of the images stored in the database correspond to cases with some type of finding of clinical interest (microcalcifications and/or visible masses, for instance), which is a good representation of the mammography exam results among Brazilian women. Images without suspect findings of clinical interest are also important in CAD schemes evaluation, because they can be used for evaluating false-positive rates. Most of the exams stored were performed by women from 40–60 years old. This is a direct consequence of the medical recommendation to make such exams every 2 years, beyond 40 years old and annually above the age of 50. The database images are 12-bit files and their spatial resolutions are of 85 µm or of 150 µm, depending on the scanner used. The search system allows selection of the desired characteristics for image download. Using the data structure presented in the medical reports, the image origin data and the necessary information for a CAD (like type of file, contrast and spatial resolution and equipment used), a redundancy-free relational model was built. To avoid acquisition and maintenance costs and to facilitate future developers training, only open-source tools were used, specifically: Linux Debian operational system, PHP, HTML, JavaScript, and MySQL. For security reasons and to enable easy interaction with different professionals, the users were divided in three categories: “user” is the lowest level, having access to copy, search, and select images and regions of interest (ROIs) for download; “researcher,” the intermediary level, has all the access
of the “user” and also can insert new exams and new regions of interest; “administrator,” the highest level, has all the access levels above and can also edit, remove and manage exams, images and users. The access interface was developed to be easy and intuitive, while assuring data protection. Patient’s data inserted include gender, birth date, breast development age and contraceptive use, among other information useful to statistical analysis and even anamnesis. Exams data include, but is not restricted to, the medical report, breast density, date and reasons for the exam. All findings, “normal” exams and pathologies are confirmed by previous and later exams, including follow-up mammographies, ultrasound exams and/or biopsy. The system is fully integrated, allowing one to search by any or all of the parameters in any combination of datasets. The download option allows the user to download the complete image file without losses. These images are usually TIFF files, varying from 8–16 bits of contrast (each one represents about 8 MB, on average). The selection tool was built by using HTML and JavaScript for user interaction and PHP and ImageMagick for actual processing at the server. It is structured so that no plugin or download of any other tool is required. For statistical analysis purposes, a tool was made available for real-time analyses of the percentage profile of the database in relation to several different image characteristics, for example BI-RADS categories, types of findings and pathologies. Images available in the database originated from two hospitals with two different types of mammography units (a Senographe 500t and a Senographe 6000). Those images were digitized by using two laser scanners, Lumiscan 50 and Lumiscan 75 (Lumisys, Inc., Sunnyvale, California) with 12 bits of contrast (gray scale) and 0.085 mm (about 30% of them) and 0.150 mm (70%) of spatial resolution, respectively. The database is still increasing, also with FFDM (Full Field Digital Mammography) images.

### 6.1.7 BCDR Database
The BCDR (Breast Cancer Digital Repository) is a wide-ranging annotated public repository (Cardoso Moura et al. 2013; BCDR 2016) composed of Breast Cancer patients’ cases of the northern region of Portugal, and it represents the first Iberian wide-ranging annotated BCDR. The BCDR is subdivided into two different repositories: a Film Mammography based Repository (BCDR-FM) and a Full Field Digital Mammography based Repository (BCDR-DM). Both repositories were created with anonymous cases from medical archives (complying with current privacy regulations as they are also used to teach regular and postgraduate medical students) supplied by the Faculty of Medicine—Centro Hospitalar São João, at University of Porto (FMUP-HSJ). The BCDR provides normal and annotated patient cases of breast cancer, including mammography lesion outlines, anomalies observed by radiologists, pre-computed image-based descriptors, as well as related clinical data.

The BCDR-FM is composed of 1010 (998 female and 12 male) patient cases (aged between 20 and 90 years old), including 1125 studies, 3703 medio-lateral oblique (MLO) and craniocaudal (CC) mammography incidences and 1044 identified lesions clinically described (820 already identified in MLO and/or CC views). With this, 1517 segmentations were manually made and BI-RADS classified by specialized radiologists. MLO and CC images are gray-level digitized mammograms with a resolution of 720 (width) by 1168 (height) pixels and a bit depth of 8 bits per pixel, saved in TIFF format. The BCDR-DM, still in construction, is currently (BCDR 2016) composed of 724 (723 female and one male) Portuguese patient cases (aged between 27 and 92 years old), including 1042 studies, 3612 MLO and/or CC mammography incidences and 452 lesions clinically described (already identified in MLO and CC views). With this, 818 segmentations were manually made and BI-RADS classified by specialized radiologists. The MLO and CC images are gray-level mammograms with a resolution of 3328 (width) by 4084 (height) or 2560 (width) by 3328 (height) pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in TIFF format. The BCDR was released for public domain on April 18, 2012, it has been used to train Machine Learning Classifiers in CAD schemes (Suarez-Ortega and Franco-Valiente 2013) and it is still in development. Currently, four benchmarking datasets (two masses-based and two micro-calculifications and calcifications-based), representative of benign and malignant lesions (biopsy proven), comprising instances of clinical and image-based features, are available for free download to registered users.

#### 6.1.8 INbreast Database
The INbreast database, available at Inbreast (2016), is a mammographic database, with images acquired at a Breast Center, located in a University Hospital (Hospital de São João, Breast Centre, Porto, Portugal), recently used to validate a novel technique of mass characterization based on robust features-fusion (El Atlas et al. 2016). INbreast has a total of 115 cases (410 images), of which 90 cases are from women with both breasts (four images per case) and 25 cases are from mastectomy patients (two images per case). Several types of lesions (masses, calcifications, asymmetries and distortions) are included. Accurate contours made by specialists are also provided in XML format.

#### 6.1.9 LLNL/UCSF Database
Lawrence Livermore National Laboratories (LLNL), along with the University of California at San Francisco (UCSF) Radiology Department, have developed (USF 2016) a 12 volume CD Library of digitized mammograms featuring micro-calculifications, which has been used in mammography CAD (Liu et al. 2016). For each digitized film image, the extent of all calcification clusters and the contour and area of a few individual calcifications in each cluster are noted. Along with the “truth” images, a file with case history, radiologists comments and other information is provided. The library contains 198 films from 50 patients (four views per patient, but only two views from one mastectomy case), selected so as to span a range of cases of interest. These films were digitized to 35 microns. Each pixel was sampled to 12 bits of grayscale. As a result, each digitized mammogram results in an image that is about 50 megabytes in size, for a total of nearly 6 gigabytes for the entire library. The films were selected to present five normal, average, healthy cases
Databases for Mammography

(previous normal mammograms and no history of ultrasound, magnification views, biopsy, etc.), five normal but difficult cases (with either dense or fibrous breasts, implants or asymptomatic tissue), 20 cases of obviously benign microcalcifications (with at least 3 years of follow-up without change or developing cancer), 12 cases of suspicious, benign microcalcifications (all these benign cases had either a biopsy or a diagnostic mammogram plus at least 3 years of subsequent follow-up without change or developing cancer) and eight cases with a malignant cluster of microcalcifications, biopsy proven. The set is available via ftp link at USF (2016).

61.10 AMDI Database

The AMDI (Indexed Atlas of Digital Mammograms) is a system that integrates modules to permit the addition of new cases into the mammographic database by authorized radiologists and to assist research and education activities in breast cancer through a flexible and easy-to-use interface via the Web (AMDI 2016). The participating Institutions are the Hospital de Clínicas de Uberlandia (MG, Brazil) and the Instituto Victorio Valeri de Diagnósticos Médicos (SP, Brazil). The mammographic database was developed using the PostgreSQL with Image-Handling Extension (PostgreSQL-IE), which is an eXtension Relational DataBase Management System (XRDBMS) developed by the research group and available for downloading in the download section of AMDI (2016). The mammographic database was projected to include cases with all of the available mammographic views, radiological findings, diagnosis proven by biopsy, the patient’s clinical history and information regarding the lifestyle of the patient. Each exam of each case includes four views (two views of each breast: CC and MLO). To address the teaching and research aspects, the database links each mammogram with the contour of the breast, the boundary of the pectoral muscle (MLO views only), the contours of masses (if present), the regions of clusters of calcifications and the number of calcifications (if present) and the locations and details of any other features of interest. The contours of masses and regions of clusters of calcifications may be drawn interactively by an authorized expert radiologist, when he or she is including a new case in the database. The mammographic database also supports the inclusion of several mammographic exams of the same patient performed at different instances of time. This information can be used for temporal analysis of the breast, due to natural modifications that occur during the life of the woman or to analyze interval cancer.

The research system called SISPRIM—Sistema de Pesquisa para Recuperação de Imagens Mamográficas, that is Research System for Retrieval of Mammographic Images—integrated with AMDI, allows physicians and oncologists to study any possible statistical correlation between the incidence of breast cancer and the lifestyle of the patient. For this purpose, some of the important items of data supported by AMDI relate to information regarding the patient, including basic data about the clinical history of the patient as well as lifestyle, such as food habits, exercise, diet and the use of antidepressive medication, tobacco or alcohol. The SISPRIM research system also allows content-based image retrieval to assist the radiologist in breast cancer diagnosis. The e-learning system called INDIAM (INterpretation and Diagnosis of Mammograms) was designed to assist a medical student or resident in the interpretation of mammograms and diagnosis of breast cancer and is being developed using Web services (Giuliato et al. 2009). INDIAM makes available a tutorial that uses education techniques to guide the users (doctors, students or researchers) through concepts related to the diagnosis of breast cancer. It also makes available a module to simulate the analysis and diagnosis of breast cancer using cases retrieved from a mammographic database and another module for training the student in the interpretation of mammograms. The e-learning system is being developed utilizing an ontology for the interpretation of mammograms (OntoBreastCancer) that provides controlled and consistent vocabularies to describe concepts and relationships, thereby enabling knowledge sharing. The system makes available a user-friendly graphical Web interface that is configured according to the service being provided. AMDI provides a tool that enables the user to download cases from the mammographic database, so as to make the information available to authorized medical and research communities interested in breast cancer diagnosis. Figure 61.3 reports an overview of the AMDI system (AMDI 2016).

61.11 IRMA Project

IRMA (Image Retrieval in Medical Applications) is a cooperative project of the Department of Diagnostic Radiology, the Department of Medical Informatics, Division of Medical Image Processing and the Chair of Computer Science VI at the Aachen University of Technology (RWTH Aachen). The aim of the project (Deserno et al., 2011; IRMA 2016; MAMMOIMAGE 2016) is the development and implementation of high-level methods for content-based image retrieval with prototypical application to medico-diagnostic tasks on a radiologic image archive. Integrated to the IRMA project, an available mammography database (Oliveira et al. 2008) has been developed from the union of The Mammographic Image Analysis Society Digital Mammogram Database (MIAS), The Digital Database for Screening Mammography (DDSM), the Lawrence Livermore National Laboratory (LLNL) and routine images from the Rheinisch-Westfälische Technische Hochschule (RWTH), Aachen. Using the IRMA code, standardized coding of tissue type, tumor staging and lesion description was developed according to the ACR tissue codes and the ACR BI-RADS. The import can be done automatically using scripts for image download, file format conversion, file name, web page and information file browsing. Disregarding the resolution, this resulted in a total of 10,509 reference images and 6767 images are associated with an IRMA contour information feature file. In accordance with the respective license agreements, the database is freely available for research purposes and may be used for image-based evaluation campaigns such as the Cross Language Evaluation Forum (CLEF). It can be extended easily with further cases imported from a picture archiving and communication system (PACS).

61.12 The DBMS

The DBMS (Data Base Management System) (Adusei et al. 2010) role is to give data support to the CAD/CADx system, by providing
a comprehensive package of data including images, patient information, pathology, lesion types, and so on, and providing the capability to retrieve these data through criteria set by end users of the system. The database system is able to store and retrieve all images and other data needed by radiologists for cancer annotation and by the scientists for research purposes. It is designed and developed as a Knowledge Base system to support detection, diagnosis and research in mammography. The database design combined two standards, the Breast Imaging Reporting and Data System (BI-RADS) (Eberl et al. 2006) by ACR and the Facility Oncology Registry Data Standards (FORDS) by the Commission on Cancer standards. The current design is capable of providing information on cancer registry and radiology mammography imaging and reporting. It also has functionalities such as the provision of images and information for the scientists to develop CAD algorithms, allowing radiologists to retrieve and annotate images, and provides other information for statistical analysis too. The database is designed in a hierarchical architecture with three different levels. This makes the entities and relationships in the database flow from general to specific. The design model makes the database more efficient and more scalable. Microsoft Sequeal Server 2005 was chosen as the database management system for this project, because it meets the security and scalability requirements. The mammography database has gone through a number of phases in the requirement and analysis stage. In the first design stage, some features from publicly available databases, such as the University of South Florida Digital Database for Screening Mammography (DDSM), were adopted (Heath et al. 2001). It was redesigned after meetings with radiologists from the University of Mississippi Medical Center (UMMC). At this stage, a standard widely used by radiologists in cancer detection, BI-RADS, by the American College of Radiology, was adopted (Eberl et al. 2006). After the Facility Oncology Registry Data Standards by the Commission on Cancer (FORDS), the standard used by the cancer registry at UMMC, was introduced, the design of the mammography database was again fine-tuned. The final mammography database was then developed with the cancer registry standard FORDS combined with the radiologist standard BI-RADS, radiologist clinical cancer detection and staging and the specifications of the film scanner. The design is able to provide the capability for the database system to store and retrieve data for cancer registration, annotation, statistical analysis, research, teaching, and so on. The advantages of this design are: the design has been able to combine cancer registry data with the clinical radiology image and reporting categories (while in hospitals, these two systems are usually separated from each other); and the design provides the capability to list and categorize the images and pathology, according to lesion type. The database design architecture is a hierarchical one. The design encompasses different standards and data from different resources, with three levels being designed into the architecture. Interfaces exist between the different levels. At the top level is general information collected from the cancer registry. This part includes the information on patient, cancer identification,
TABLE 61.1
Characteristics of Some Mammographic Single Databases

<table>
<thead>
<tr>
<th>Name</th>
<th>Principal Organizations</th>
<th>Numbers,* Views</th>
<th>Classification</th>
<th>Pixel Size/Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIAS (film)</td>
<td>UK National Breast Screening Programme</td>
<td>322 images, 161 patients, 1 view</td>
<td>Normal, abnormal</td>
<td>50 (µm)</td>
</tr>
<tr>
<td>DDSM (film)</td>
<td>Massachusetts General Hospital, University of South Florida, Sandia National Laboratories</td>
<td>2620 screening exams, 4 views</td>
<td>Normal, benign, cancer</td>
<td>42, 43, 5, 50 (µm)</td>
</tr>
<tr>
<td>MAGIC-5 (film)</td>
<td>National Institute for Nuclear Physics, Bari, Udine, Palermo, Firenze, Turino, Napoli, and Sassiari hospitals</td>
<td>3369 images, 967 patients, 1–6 views</td>
<td>Normal, abnormal (benign, suspicious, malignant)</td>
<td>85 (µm)</td>
</tr>
<tr>
<td>LAPIMO (film)</td>
<td>San Paolo University</td>
<td>1400 images, 320 patients, 2 views</td>
<td>Normal, abnormal (BI-RADS)</td>
<td>85, 150 (µm)</td>
</tr>
<tr>
<td>BCDR (film)</td>
<td>Faculty of Medicine—Centro Hospital São João, University of Porto</td>
<td>3703 images, 1010 patients, 2 views</td>
<td>Normal, abnormal (BI-RADS)</td>
<td>720 × 1168</td>
</tr>
<tr>
<td>BCDR (digital)</td>
<td>Faculty of Medicine—Centro Hospital São João, University of Porto</td>
<td>3610 images, 724 patients, 2 views</td>
<td>Normal, abnormal (BI-RADS)</td>
<td>3328 × 4084, 2560 × 3328</td>
</tr>
<tr>
<td>Inbreast (digital)</td>
<td>Hospital de São João, Porto</td>
<td>410 images, 115 cases, 2 views</td>
<td>Normal, abnormal (BI-RADS)</td>
<td>3328 × 4084, 2560 × 3328</td>
</tr>
<tr>
<td>LLNL/UCSF (screen)</td>
<td>Lawrence Livermore National Laboratories, San Francisco Radiology Department of University of California</td>
<td>198 images, 50 patients, 2 views</td>
<td>Normal, benign, suspicious, malignant</td>
<td>35 (µm)</td>
</tr>
</tbody>
</table>

*The number of images is influenced by the presence of cases post-mastectomy and with different numbers of projections.

1257

Databases for Mammography

1257

comorbidities, staging and treatment, and so on. Since the information held in this part is general, it can be extended to fit other types of cancer diagnosis. At the middle level is the information on breast cancer, reporting, pathology and scanning. At this level, the information becomes more specific and narrowed down to each case of breast cancer, while it is still general compared to information held in the next level. At the lowest level in this hierarchy are the entities related to image and pathology of three lesion types, architecture distortion, calcification and mass. Each case has four scanned images. The information held at this level is specific to each image, such as laterality, view, shape, distribution, boundary of annotation, number of abnormalities, and so on. In total, in the database there are 21 tables and 293 attributes.

61.13 Conclusions

Many mammographic databases with different characteristics have been made available as a public research database to the international community over the past decades. Their principal characteristics (where present and as described in the literature) are summarized in Table 61.1.

The importance of such a database is remarkable, especially with regard the validation of CAD systems (Ibrahim et al. 1997; Patrocino et al. 2004; Bellotti et al. 2006; Cascio et al. 2006; Retico et al. 2006; Shi et al. 2008; Hapfelmeier and Horsch 2010; Deserno et al. 2011; Suarez-Ortega and Franco-Valiente 2013; Luqman Mahmood Mina and Nor Ashidi Mat Isa 2015; El Atlas et al. 2016; Liu et al. 2016) and image processing tools, but also for remote teaching and training and for epidemiological studies.

Recent advances in information technology have led to ever-changing and increasing systems, in which several centers located in different geographical places contribute to enlarging the database and share the resources (Heath et al. 2001; Eberl et al. 2006; Giulio et al. 2009; Adusei et al. 2010; Deserno et al. 2011; AMDI 2016; IRMA 2016; MAMMOIMAGE 2016).

REFERENCES


