INbreast:

Toward a Full-field Digital Mammographic Database

Inês C. Moreira, MSc student, Igor Amaral, MSc, Inês Domingues, MSc,
António Cardoso, MD, Maria João Cardoso, PhD, Jaime S. Cardoso, PhD

Rationale and Objectives: Computer-aided detection and diagnosis (CAD) systems have been developed in the past two decades to assist radiologists in the detection and diagnosis of lesions seen on breast imaging exams, thus providing a second opinion. Mammographic databases play an important role in the development of algorithms aiming at the detection and diagnosis of mammary lesions. However, available databases often do not take into consideration all the requirements needed for research and study purposes. This article aims to present and detail a new mammographic database.

Materials and Methods: Images were acquired at a breast center located in a university hospital (Centro Hospitalar de S. João [CHSJ], Breast Centre, Porto) with the permission of the Portuguese National Committee of Data Protection and Hospital's Ethics Committee. MammoNovation Siemens full-field digital mammography, with a solid-state detector of amorphous selenium was used.

Results: The new database—INbreast—has a total of 115 cases (410 images) from which 90 cases are from women with both breasts affected (four images per case) and 25 cases are from mastectomy patients (two images per case). Several types of lesions (masses, calcifications, asymmetries, and distortions) were included. Accurate contours made by specialists are also provided in XML format.

Conclusion: The strengths of the actually presented database—INbreast—relies on the fact that it was built with full-field digital mammograms (in opposition to digitized mammograms), it presents a wide variability of cases, and is made publicly available together with precise annotations. We believe that this database can be a reference for future works centered or related to breast cancer imaging.

Key Words: Mammographic database; CAD; computer-aided detection; computer-aided diagnosis.

©AUR, 2011

According to the World Health Organization, breast cancer was responsible for approximately 519,000 deaths in 2004: 16% of all cancer incidence among women. In 2008, it was the most common form of cancer and cancer related death in women worldwide (1). In Portugal, 1500 women die every year from breast cancer, whereas in the European Union it is responsible for one in every six deaths from cancer in women (2). For this reason, early detection and diagnosis of breast cancer is essential to decrease its associ‐ated mortality rate. Therefore, mass screening is recommended by the medical community (2,3).

X–ray mammography is currently considered the best imaging method for breast cancer screening and the most effective tool for early detection of this disease (4). Screening

mammographic examinations are performed annually on asymptomatic women to detect early, clinically unsuspected lesions. The age at which mass screening mammography is generally recommended in the United States is 40 (5). In Europe, screening at 40 to 50 years old is still not consensual (6). However, in women with genetic mutations or significant family history of breast cancer, screening should start earlier, usually 10 years earlier than the age of diagnosis of the youngest relative (never before 25) (5).

Mammography comprehends the recording of two views for each breast: the craniocaudal (CC) view, which is a top to bottom view, and a mediolateral oblique (MLO) view, which is a side view (Fig 1) (6). The images can be acquired on x-ray film, such as a film-screen mammogram, or in digital format, such as with digital mammography (full-field digital mammography [FFDM] and computed radiography) (7).
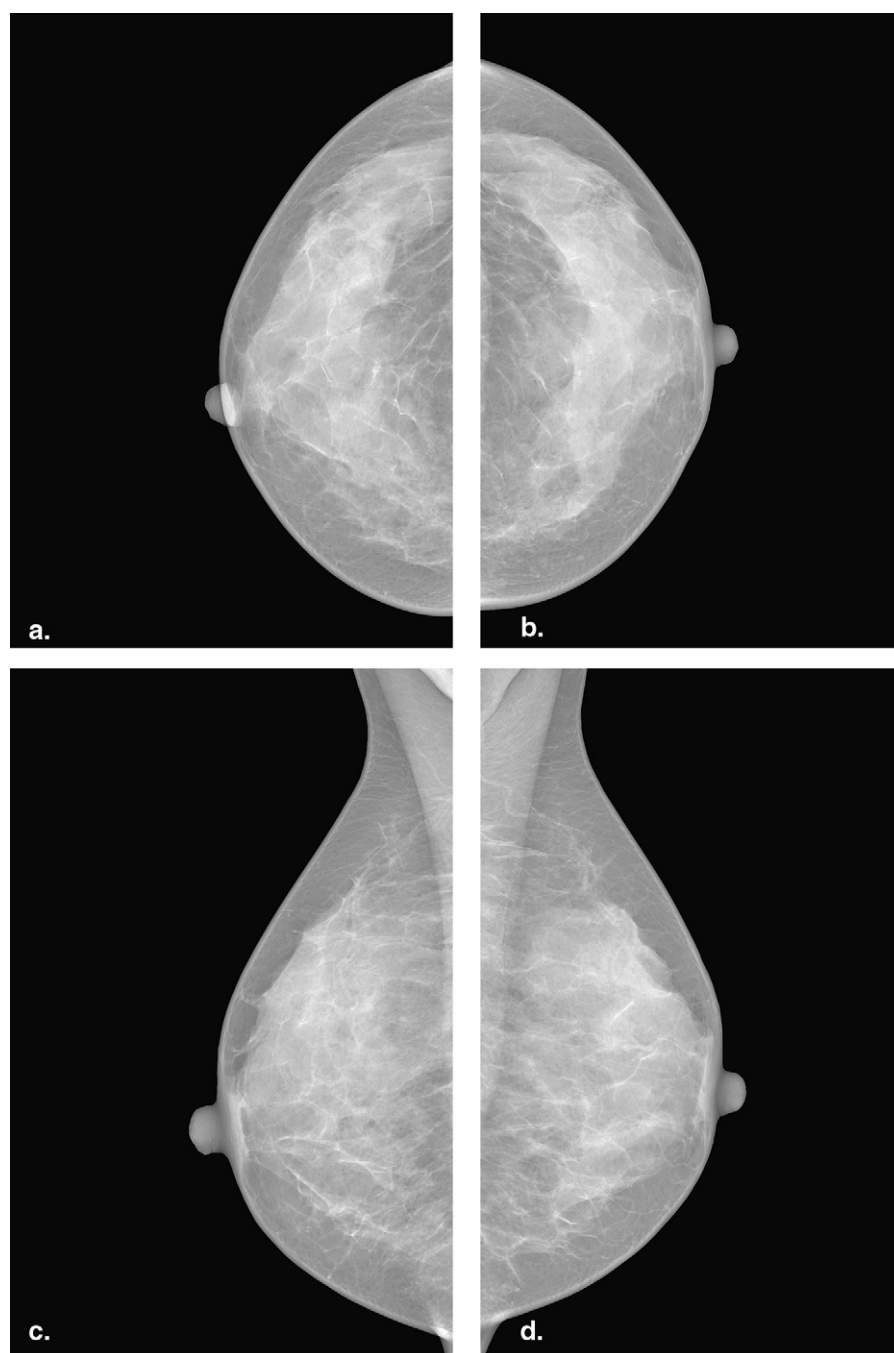
When radiologists examine mammograms, they look for specific abnormalities (8). The most common findings seen on mammography are masses, calcifications, architectural distortion of breast tissue, and asymmetries when comparing the two breasts and the two views. To standardize the termi‐nology of the mammographic report, the assessment of find‐ings and the recommendation of action to be taken, the American College of Radiology (ACR) has developed the Breast Imaging Reporting and Data System (BI-RADS) scale (9). Based on level of suspicion, the previously mentioned lesions can be placed into one of six BI-RADS categories: category 0, exam is not conclusive; category 1, no findings;

Figure 1. Mammogram examples: **(a)** craniocaudal (CC) view of the right breast; **(b)** CC view of the left breast; **(c)** mediolateral oblique (MLO) view of the right breast; **(d)** MLO view of the left breast.

category 2, benign findings; category 3, probably benign findings; category 4, suspicious findings; category 5, a high probability of malignancy; and category 6, proved cancer (Table 1). In case of categories 4 and 5, a biopsy is needed to exclude or confirm malignancy (10). Other important characteristic referred by the ACR is the breast composition tissue, related to the breast density shown in x-rays. There are four categories ranging from 1, for low density (fatty tissue), to 4, for very high density (dense tissue) (11).

Several studies (8,12,13) concluded that detection of suspicious findings by radiologists is a repetitive and fatiguing task, leading to a 10%–30% rate of undetected lesions. To decrease this rate, computer-aided detection and diagnosis (CAD) systems have been developed in the past two decades to assist the radiologists in the interpretation of the medical images (14,15). To design, test, and tune such computational systems, researchers demand a large number of mammograms (16). These datasets need to be digital, so if the images are acquired on x-ray film, they have to be digitized (15). Therefore, mammographic databases play an important role in the development of algorithms aiming at detecting and diagnosing lesions. They are also important to allow comparison of results from different studies (17–19). A different application is the use of database images

TABLE 1. Breast Imaging Reporting and Data System Assessment Categories

Category	Description
0	Needs additional imaging evaluation and/or prior mammograms for comparison
1	Negative
2	Benign finding(s)
3	Probably benign finding(s). Short-interval follow-up is suggested.
4	Suspicious anomaly. Biopsy should be considered.
5	Highly suggestive of malignancy. Appropriate action should be taken.
6	Biopsy proven malignancy

for teaching and training students in this specific medical field.

The common practice in the development of CAD algorithms has been the use of private sets of mammograms to design and evaluate the performance of the algorithms. This impairs the fair judgment of the quality of the individual work and the comparison of the accuracy of different methods because performance is database-dependent. Good results can have been obtained in databases with "easy" cases, whereas bad accuracies may have been achieved by using "difficult" databases (16,18,20). Public available databases could provide a common ground for researchers to develop, test, and compare their methods. However, to be effective, certain criteria should be met by the database.

This article aims to present a new mammographic research database originated at Centro Hospitalar de S. João (CHSJ) at Porto, Portugal, the INbreast database with the purpose of developing CAD methods and to overcome some limitations of existent databases. Details of the design of this database will be presented in this article as follows: requirements of digital mammographic databases, existent available databases, description of INbreast database, description of the findings in the database, proposal for a methodology for performance evaluation, discussion, and conclusion.

REQUIREMENTS FOR A DIGITAL MAMMOGRAPHIC DATABASE

According to previous studies (16,18), mammographic databases should take into consideration the following requirements.

Case Selection

The database should include various cases with images with normal breasts and all types of findings, and also all types of breast density. Normal images with structures that may be misleading (eg, superimposed tissue that looks like a mass) are important in order to make the classifiers more robust. The cases should be collected by a specialist experienced in mammography. Each case should contain four standard views,

unless it is a case from a patient with one breast only from previous mastectomy. Image acquisition should be adequate in terms of patient positioning, x-ray exposure, and with an absence of image blur due to patient motion.

Ground Truth

Biopsy proof for all cases should be available. Annotations should include the "ground truth" (GT) concerning cytology/histology for all cases, the location and boundaries of the lesion with the outline marking performed by an imaging specialist.

Associated Information

Clinical history such as age, family history, and previous biopsies can be useful for studying subpopulations of women (eg, women <50 years of age), and it may improve the performance of a CAD scheme by incorporating nonradiographic information. Additional information such as breast density (preferably given by a standard like ACR) and BI-RADS classification are also mandatory.

Organization of Database

A specific file format for digital mammograms does not exist. 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 (21). A division of the images on training and test sets should also be suggested. By doing so, different methods can be compared.

Distribution of Database

The database should be available, preferentially over the Internet. Continuous user support is also indispensable.

AVAILABLE DATABASES

There are several image databases, some public and some restricted to individual groups, which are used by researchers in the breast cancer area. However, these often do not meet all the requirements needed for a study (16–18,20,22–24).

The Mammographic Image Analysis Society Digital Mammogram Database

The Mammographic Image Analysis Society Digital Mammogram Database (MIAS) (25), despite being the oldest available database, is still widely used in literature. This database has been reduced in resolution and is reachable at http://peipa.essex.ac.uk/info/mias.html. Although images are still available, it is no longer supported.

MIAS consists of 161 cases, 322 digitized MLO images, with all types of findings, including benign and malign lesions, and also normal images. It has a high percentage of spiculated

TABLE 2. Most Used Databases in Literature

	MIAS (24)	DDSM (29)	BancoWeb (35)
Origin	UK	USA	Brazil
Year	1994	1999	2010
Number of cases	161	2620	320
Views	MLO	MLO and CC	MLO, CC, and other
Number of images	322	10,480	1400
Mode of image acquisition	Screen film	Screen film	Screen film
Image type file	PGM	LJPEG	TIFF
Resolution	8 bits/pixel	8 or 16 bits/pixel	12 bits/pixel
Lesion type	All kinds (with special concentration of spiculated masses)	All kinds	All kinds
Ground truth	Center and radius of a circle around the interest area	Pixel level boundary of the findings	ROI is available in a few images only
BI-RADS	No	Yes	Yes
Breast density	Yes (not ACR)	Yes (in ACR standard)	Yes (not ACR)
Clinical history	No	Age	Yes
Search system	No	Yes, but not functional	Yes
Access	Yes	Yes	Yes
Support	No	No	Yes

ACR, American College of Radiology; CC, craniocaudal; DDSM, Digital Database for Screening Mammography; LJPEG, lossless JPEG (Joint Photographic Experts Group); MIAS, Mammographic Image Analysis Society Digital Mammogram Database; MLO, mediolateral oblique; PGM, portable gray map; ROI, region of interest; TIFF, tagged image file format; UK, United Kingdom; USA, United States of America.

masses but, in Rangayyan's article (24), the author noticed that there was an unexpected elevated number of benign findings in relation to the malign ones. It contains breast density information, but not classified according to the ACR standards. However, because of the increasing usage of the ACR classification, it was decided to classify the set of mammograms according to that standard (26).

MIAS annotations consist in the center and radius of a circle around the area of interest. These types of annotations are not considered sufficient for some studies, as the one done by Oliver et al (20), where all circumscribed and spiculated lesions had to be manually segmented. Another drawback is the resolution to which the images have been digitized, which makes MIAS unsuitable for experiments on detection of microcalcifications (MCCs) (27). However, in previous work (28), the authors achieved a 100% detection rate of MCCs by applying two different detection methods. Llobet (29) considered that, in the case of calcifications, the GT region contains more healthy tissue than affected tissue. For this reason, calcifications were not included in his study.

The Digital Database for Screening Mammography

The most used database is the Digital Database for Screening Mammography (DDSM) (30), and it is accessible at http://marathon.csee.usf.edu/Mammography/Database.html, but is also no longer supported. It is the largest public database, with 2620 cases including two images from each breast (MLO and CC), for a total of 10,480 images, with all types of findings from normal images to images with benign and malign lesions. Some of the cases in this database were collected from the Nijmegen Database (31). The only patient

information included is the age, but it has breast density annotations (ACR) and BI-RADS annotations.

Image annotations include pixel level boundary of the findings. There are several articles whose authors got satisfactory results using this type of annotation (32–35). However, as noted in other studies (8,22,23), they are not adequate for the validation of segmentation algorithms because the precision is not good enough.

The BancoWeb LAPIMO Database

A more recent database is the BancoWeb LAPIMO Database (36). After registration, users can gain access and contribute to the database at http://lapimo.sel.eesc.usp.br/bancoweb/.

It has 320 cases, 1473 images with MLO, CC, and magnification views, with normal images, and images with benign and malign findings. Background patient information along with BI-RADS annotations is available. Annotations exist in only some of the images, in the form of a region of interest (ROI), but all have textual description of the findings. We did not find any published work related to this database, probably because it is a recent project. A summary of these databases can be found in Table 2.

Antoniou (19) refers to a web-accessible mammographic database, called MIRAcle DB, which is still in an experimental stage and became available online in the summer of 2011. This database, as well as others, is cited in the literature. However, as can be seen in Table 3, most of them are not available and consequently details could not be found.

There are also two grids that are a combination of images from multiple resources, which are available to the user as a single database, but can be stored in several servers.

TABLE 3. Other Databases Referred in Literature

	Nijmegen (37)	Trueta (38)	IRMA (39)	MIRAcle (19)	LLNL (39)	Málaga (38)	NDMA (40)
Origin	The Netherlands	Spain	Germany	Greece	USA	Spain	USA
Year	1998	2008	2008	2009	Unknown	Unknown	Unknown
Number of cases	21	89	Unknown	196	50	35	Unknown
Views	MLO and CC	MLO and CC	MLO and CC	Unknown	MLO and CC	MLO and CC	Unknown
Number of images	40	320	10,509	204	198	Unknown	1,000,000
Mode of image acquisition	Screen film	FFDM	Screen film	Unknown	Unknown	Unknown	Unknown
Image type file	Unknown	DICOM	Several	Unknown	ICS	Raw	Unknown
Resolution	12 bits/pixel	12 bits/pixel	Several	Unknown	12 bits/pixel	12 bits/pixel	Unknown
Lesion type	MCCs	All kind	All kind	Unknown	Calcifications	Masses	Unknown
Ground truth	Center and radius of a circle around the interest area	Center and radius of a circle around the interest area	Several	Region of Interest	Outline of calcifications	Pixel level annotations	Unknown
BI-RADS	Unknown	Yes	Yes	Yes	Unknown	Unknown	Unknown
Breast density	Unknown	ACR	ACR	No	Unknown	Unknown	Yes
Clinical history	Unknown	Unknown	Yes	No	Yes	Unknown	Unknown
Search system	No	Unknown	Unknown	YES	Unknown	Unknown	Unknown
Access	No	No	Yes	Summer 2011	Paid	Unknown	No
Support	No	Yes	Yes	Yes	Unknown	Unknown	Unknown

ACR, American College of Radiology; BI-RADS, Breast Imaging Reporting and Data System; CC, craniocaudal; DICOM, Digital Imaging and Communications in Medicine; FFDM, full-field digital mammography; ICS, Image Cytometry Standard; IRMA, Image Retrieval in Medical Applications; LLNL, Lawrence Livermore National Laboratory; MCC, microcalcification; MLO, medio-lateral oblique; NDMA, National Digital Medical Archive.
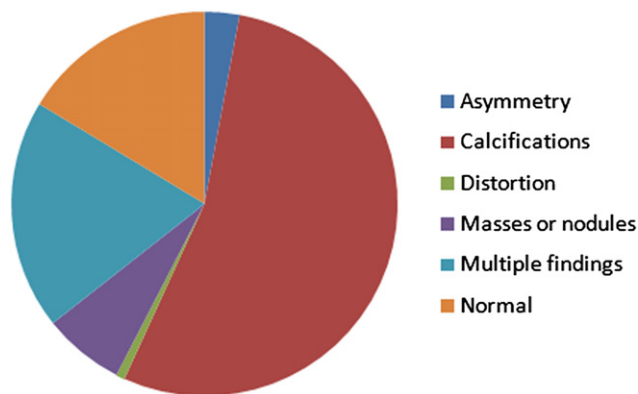
Figure 2. Chart describing the findings in the INbreast database.

Magic-5 (41) (previously known as GPCalma) is an Italian database built in 2002 and containing 967 cases with images in MLO, CC, and lateral views, making a total of 3369 images. The screen films were digitized with a resolution of 12 bits and saved in the DICOM format. Both masses and MCCs are present and the GT consists in the centre and radius of a circle around the interest area. Patient age is available but it has no BI-RADS categorization and the density classification is not on the ACR standard. Magic-5 limitations are related to the different environments were images were acquired, making them very heterogeneous. MammoGrid (42) is a collaboration between the United Kingdom, Italy, and Switzerland, with images being standardized using the Standard Mammogram Form representation and saved in the DICOM format. This grid has both screen films and FFDM images and annotation workstations are available in the participating hospitals. The main limitation of MammoGrid is that it is only available for associated institutions.

INBREAST DATABASE DESCRIPTION

The database was acquired at the Breast Centre in CHSJ, Porto, under permission of both the Hospital's Ethics Committee and the National Committee of Data Protection. The images were acquired between April 2008 and July 2010; the acquisition equipment was the MammoNovation Siemens FFDM, with a solid-state detector of amorphous selenium, pixel size of 70 μm (microns), and 14-bit contrast resolution. The image matrix was 3328×4084 or 2560×3328 pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). Images were saved in the DICOM format. All confidential medical information was removed from the DICOM file, according to Supplement 55 of the DICOM standard; the correspondence between images of the same patient is kept with a randomly generated patient identification.

INbreast has FFDM images from screening, diagnostic, and follow-up cases. Screening is made according to national and regional standards (5). Diagnostic is made when screening shows signs of anomaly. In follow-up images, cancer was previously detected and treated. A total of 115 cases were collected, from which 90 have two images (MLO and CC) of each breast and the remaining 25 cases are from women who had a mastectomy and two views of only one breast were included. This sums to a total of 410 images. Eight of the 91 cases with 2 images per breast also have images acquired in different timings (follow-up).

The database includes examples of normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, and images with multiple findings (Fig 2). According to BI-RADS, a mass is defined as a three-dimensional structure demonstrating convex outward borders, usually evident on two orthogonal views. Benign calcifications are usually larger than calcifications associated with malignancy, are usually coarser, and are often round with smooth margins and are much more easily seen. Calcifications associated with malignancy are usually very small. An architectural distortion is defined as a focal interruption of the normal mammographic pattern of lines (converging at the nipple), usually presenting as a star-shaped distortion, with no definite mass visible. An asymmetry lacks convex outward borders of a mass and it and can be represented in three ways: size asymmetry (difference in volume between the right and left breast), focal asymmetry (unilateral, localized area of parenchyma), and global asymmetry (difference in the amount of parenchyma between the right and left breast) (43). Concerning this distinction between asymmetries, this work does not take that into consideration.

The graphic in Figure 2 shows that there is a big prominence of calcifications on our database. This reflects the real population, where calcifications are the most common finding in mammography (44).

Images contain findings of six types: asymmetries, calcifications, distortion, masses, multiple findings (Fig 3), and normal (Fig 1).

The main characteristic of this work is the carefully associated GT annotation. Most of the databases, such as MIAS, only provide a circle around the area of interest. DDSM does have pixel-level contours but, as noticed in a previous study (22), they are not exact, which can impact accuracy measures by incorrectly assigning some target pixels to background and vice versa.

The annotations were made by a specialist in the field, and validated by a second specialist, between April 2010 and December 2010. When there was a disagreement between the experts, the case was discussed until a consensus was obtained. Annotations were made on OsiriX, an open-source picture archiving and communication system (PACS) workstation, running on a Macintosh platform. Each finding has a label that identifies the type of lesion. There are seven types of annotations: asymmetry (Fig 4), calcification (Fig 5a), cluster (of MCCs), mass (Fig 5b), distortion (Fig 6a), spiculated region (Fig 6b) and pectoral muscle (only in the MLO view; Fig 7). For the types, asymmetry (Fig 4), calcification (Fig 5a), mass (Fig 5b), distortion (Fig 6a) and pectoral muscle (Fig 7), a detailed contour of
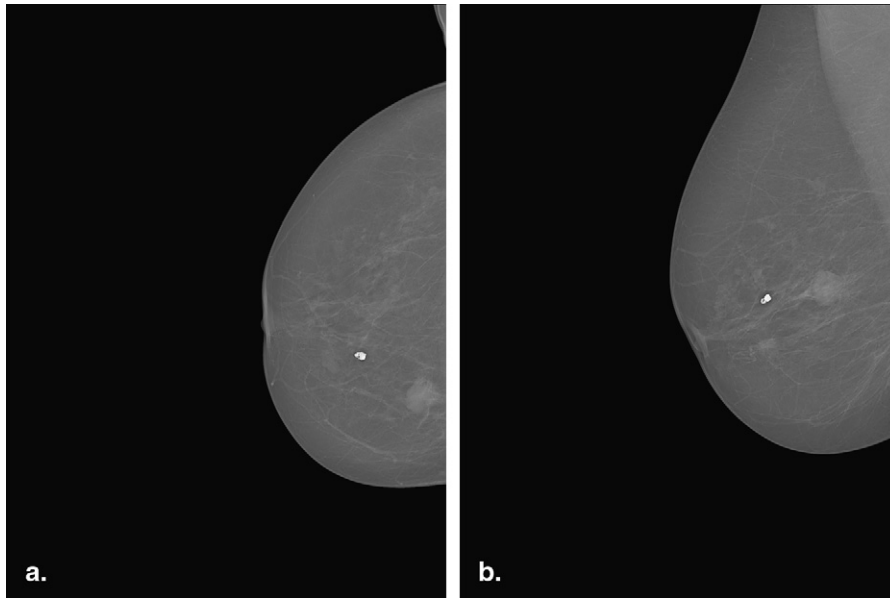
Figure 3. Database examples: multiple findings. **(a)** Craniocaudal view of the right breast; **(b)** mediolateral oblique view of the right breast.
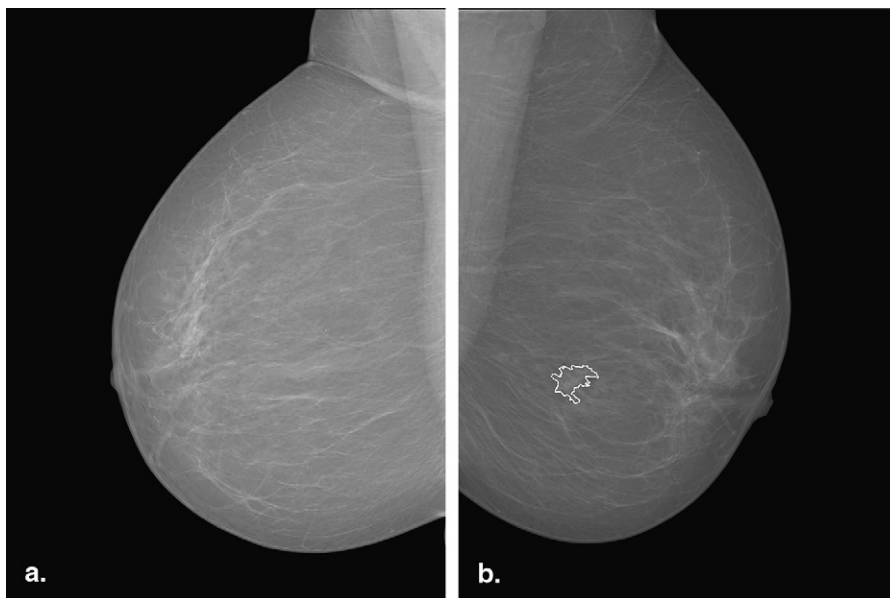
Figure 4. Annotation examples: asymmetry. **(a)** Mediolateral oblique (MLO) view of the right breast; **(b)** MLO view of the left breast.

the finding was made. An ellipse enclosing the entire cluster was adopted to annotate the clusters of MCCs (Fig 5a). When the mass is spiculated, besides a contour of the denser region, we added an ellipse enclosing all the spicules (Fig 6b).

The annotations were saved in XML format with the following structure.

- A standard header with the XML version and type of encoding information;
- The tag <key>NumberOfROIs</key> followed by an integer that indicates the number of annotations present in the image;
- For each ROI, there is a tag <key>Area</key> followed by the value of the area of the current ROI, the tag <key>

Center</key> followed by a the coordinates of the point in the centre of the ROI, the tag <key>Name</key> followed by the type of finding (mass, calcification, distortion, spiculated region) and some other general information's about the ROI;
- After the general information, for each ROI, a list of contour points is presented between the tags <array> and </array>.

Information regarding patient's age at the time of image acquisition, family history, ACR breast density annotation and BI-RADS classification is also provided (see Fig 8 for the distribution of BI-RADS classification on the database). A biopsy result for BI RADS 3, 4, 5, and 6 cases is also displayed whenever performed. The remaining cases were
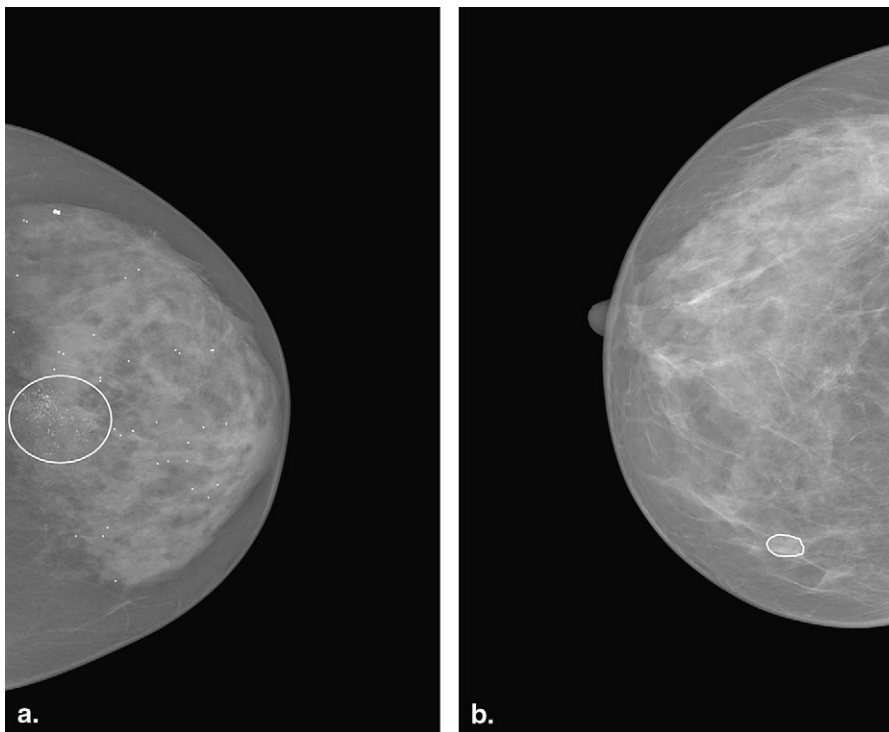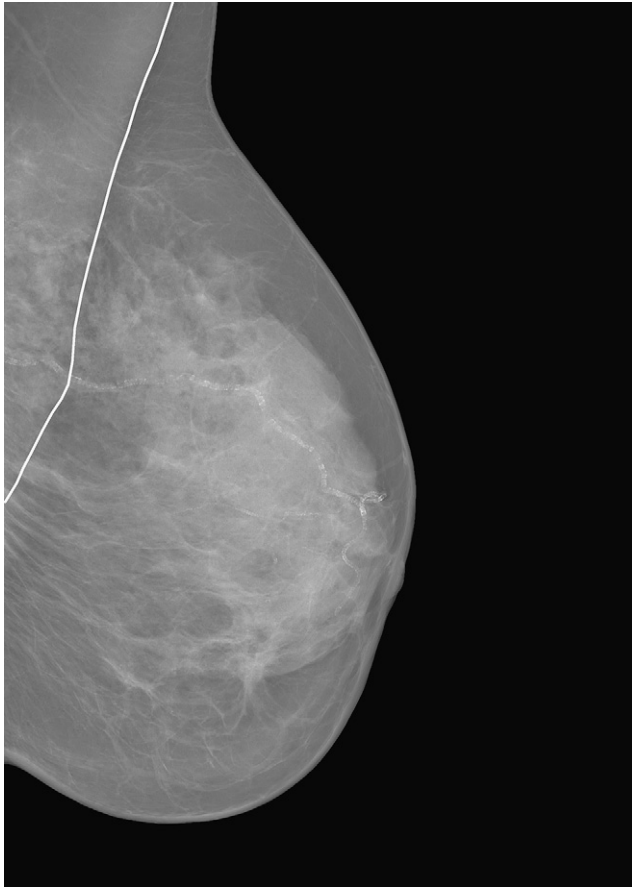
Figure 5. Annotation examples: **(a)** cluster; **(b)** masses.
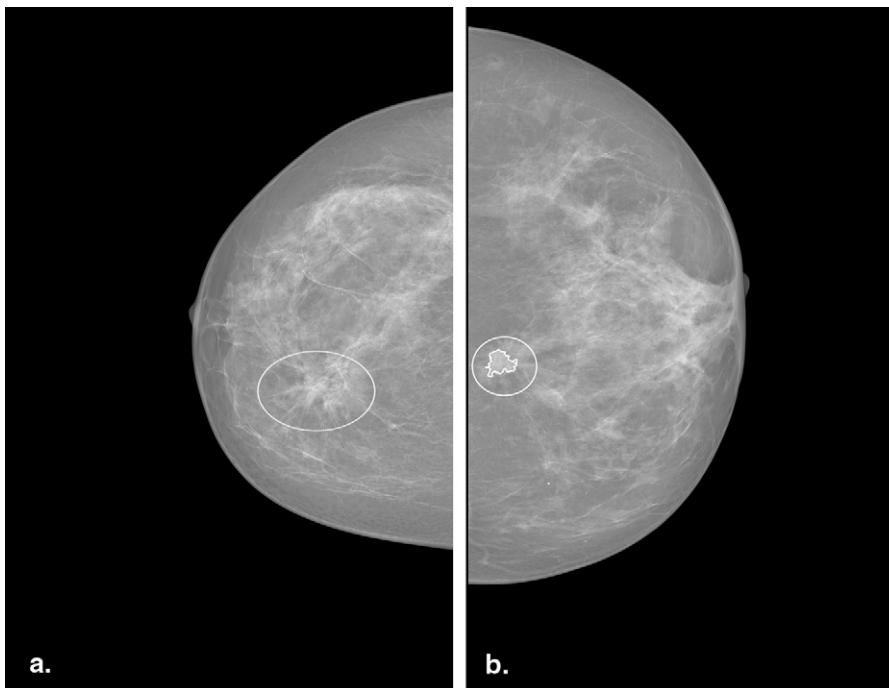
Figure 6. Annotation examples: **(a)** distortion; **(b)** spiculated region.

considered benign and therefore a biopsy was therefore not performed. Consequently, a biopsy was performed on 56 cases, of which 11 were found to be benign and the remaining 45 were malignant. The overall distribution of benign/malignant cases is shown in Figure 8.

The database is available at http://medicalresearch.inescporto.pt/breastresearch/GetINbreastDatabase.html. A division of the database into train and test sets is also suggested.

With the precise annotations in INbreast, future studies can be developed that cannot be performed with the currently available databases. Shape information is highly indicative of the malignancy of a mass (45) and therefore automatic shape assessment in the mammogram in often pursued. However, the coarse-grained annotation of current databases does not allow a proper validation of the discoveries. Also, MCC grouping and distribution is the mammogram is important

Figure 7. Annotation example: pectoral muscle.

a.

b.

Figure 8. Charts of **(a)** the Breast Imaging Reporting and Data System images distribution **(b)** benign/malignant cases distribution.

to the correct diagnosis. Again, the usual annotation of the MCCs with a single region enclosing all MCCs is insufficient for the development of automatic methods.
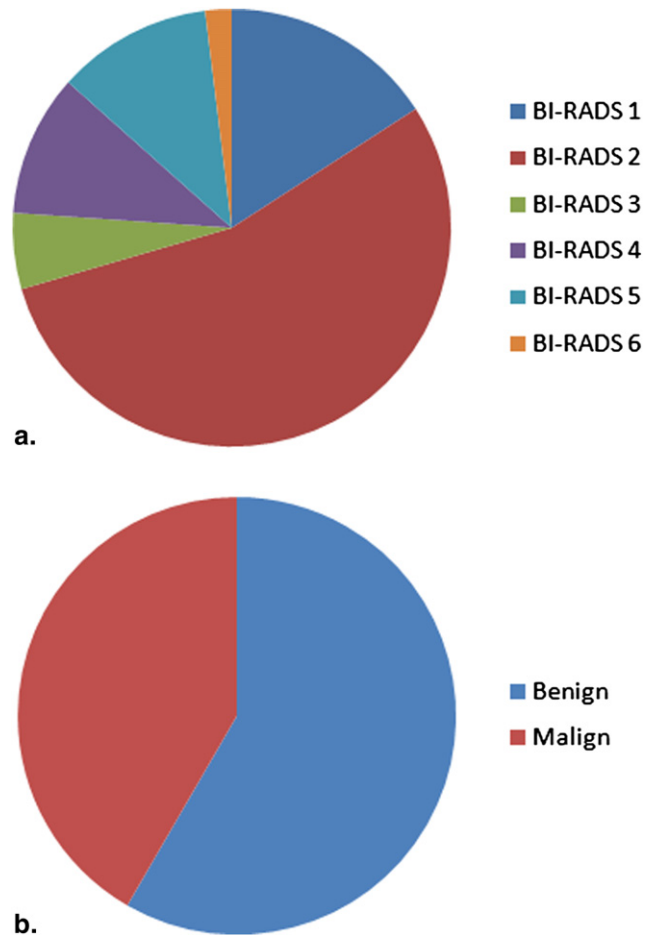
FINDINGS CHARACTERISTICS

One of the most important breast characteristic is density. Dense breasts are harder to analyze through mammography than nondense ones. For each image in our database, its density in ACR standard scale is available. A distribution of density for each BI-RADS class is presented in Figure 9.

There are a total of 116 masses among 107 images (≈ 1.1 masses per image). The number of masses (normalized by the total number of each class images) for each one of the BI-RADS classes is shown in Figure 10; the average mass size is $479 \, \text{mm}^2$ (with a standard deviation of $619 \, \text{mm}^2$); the smallest mass has $15 \, \text{mm}^2$ and the biggest has an area of $3689 \, \text{mm}^2$.

Localization distribution of Masses is depicted in Figure 11.

Concerning calcifications, they are present in 301 of the 410 images. The tag "cluster" was only used in 27 sets of calcifications, in 21 images (≈ 1.3 clusters per image). Of these 21 images, only 2 had no single calcifications annotation. A total of 6880 calcifications were thus individually identified in 299 images (≈ 23.0 calcifications per image). BI-RADS distribution is depicted in Figure 12.

Finally, the distribution of the patient age, also included in the database, is portrayed in Figure 13.

PERFORMANCE EVALUATION

In CAD research, the quality of the detection algorithm is usually reported with the miss detection rate, false-positive rate, or similar metrics. With masses the computation of such quantities is usually a straightforward process; because the number of objects is usually small, the correspondence between detected and manually annotated masses is usually clear. After adopting a distance notion between a detected and manually annotated mass (eg, a measure of overlap between the masses), a mass in considered correctly detected is its distance to an automatically detected mass is below a certain threshold. From there, the miss detection rate is just the number of undetected reference mass and the false-positive rate is just the number of automatically detected masses minus correctly classified masses.

This procedure is far from being easily extended to MCCs, with tens of objects per image. By computing the performance as the result of accumulating local errors, we will likely incur in many-to-one or one-to-many correspondences
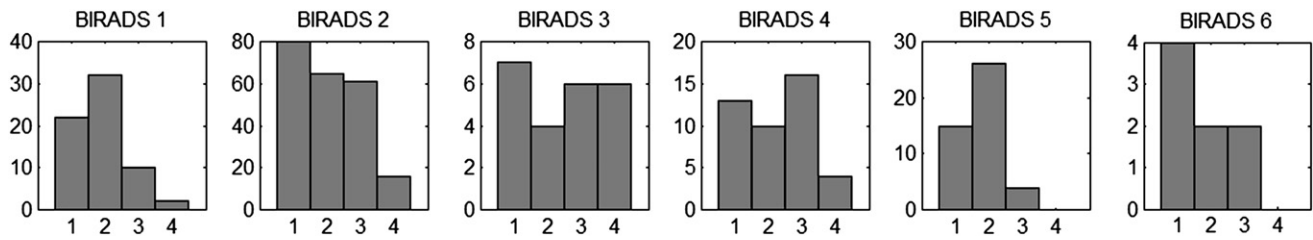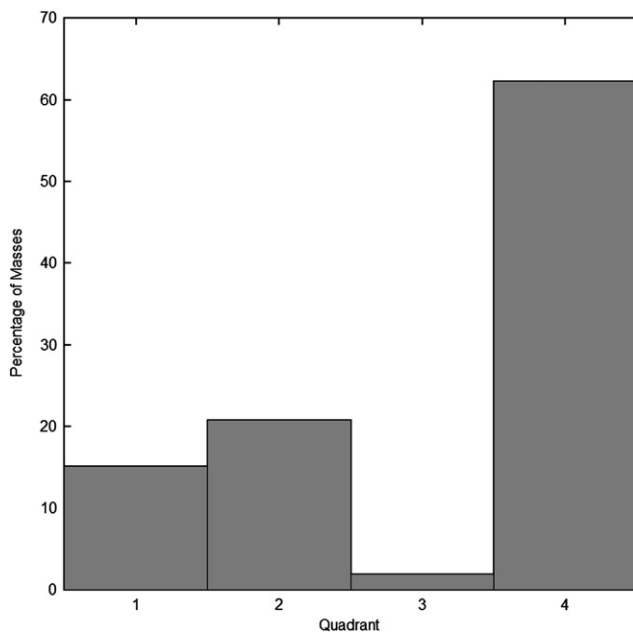
Figure 9. Distribution of density across the Breast Imaging Reporting and Data System scale.
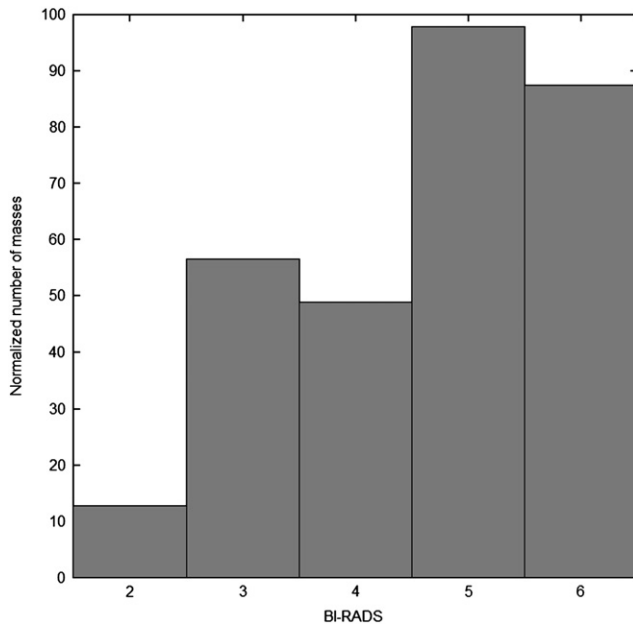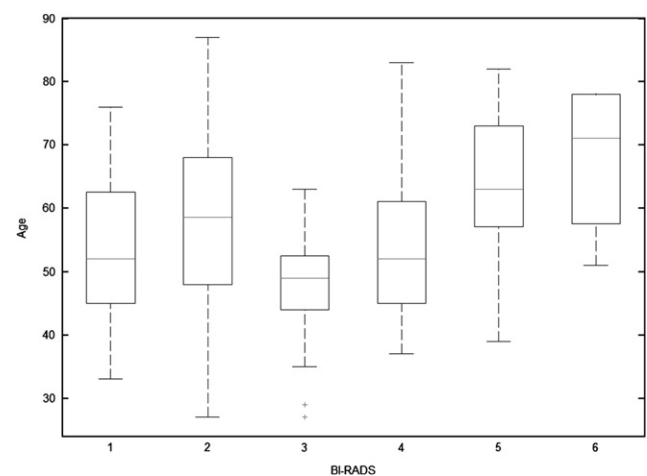

Figure 10. Normalized distribution of masses across the Breast Imaging Reporting and Data System scale.
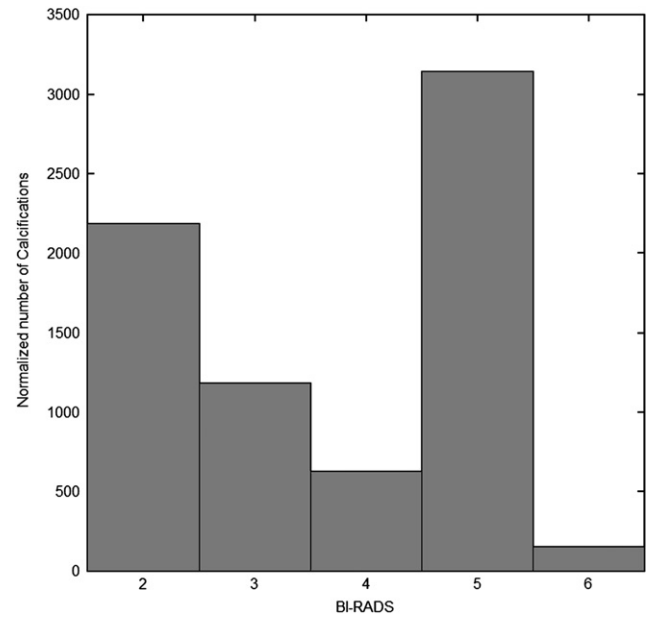
Figure 12. Normalized distribution of calcifications across the Breast Imaging Reporting and Data System scale.

Figure 11. Percentage of masses on each quadrant. UIQ, upper internal quadrant; LIQ, lower internal quadrant; LOQ, lower outer quadrant; UOQ, upper outer quadrant.

Figure 13. Distribution of age across the Breast Imaging Reporting and Data System scale.

between reference and automatically detected MCCs. Generally, we argue, as others before (46), that the most interesting measures arise when one defines the (dis)similarity as the result of optimizing a global function defined over all reference and detected objects simultaneously.
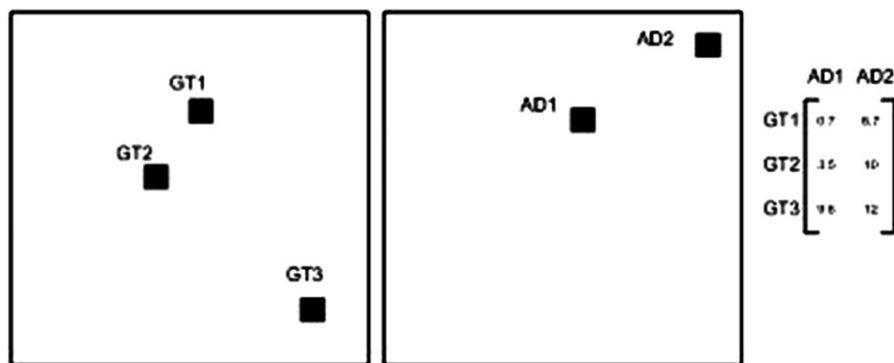
Figure 14. Toy example with three ground truth microcalcifications (MCCs) and two automatically MCCs. Using directly the Euclidean distance between the centroids, the MCCs results in the assignment problem results in the matrix at the right. The output would match ground truth (GT) 1 to automatic detection (AD)2 and GT2 to AD1. Saturating the distances to T2 = 5 would correctly match GT1 to AD1.

To evaluate the miss detection rate and false-positive rate, we propose to start by computing the distance between each reference MCC and each actually detected MCC; then we solve the matching problem on the resulting bipartite graph by minimizing the assignment cost (= distance). Only pairs with average error-distance below a certain threshold (T_1) are assumed correctly matched (the other pairs are assumed to originate from a false-positive staff line being matched to an undetected true staff line and are therefore unmatched). Now the two metrics result as the number of unmatched MCC (false positive) and unmatched reference MCC (missed to detect). It should be noted that these metrics only measure whether MCC are found, not how good the match is.

The detection threshold T_1 should reflect our imprecision acceptance in the detection process. A final remark is related to the notion of distance between two MCCs. Because MCCs are very small structures, it is a valid assumption to treat them as singular points, their centroids. The obvious choice of embracing the Euclidean distance between the centroids for the previously described global optimization can lead to unnatural results (see Fig 14).

The insight is that because high errors penalize a lot the global optimization, the final correspondence result tries to avoid such solutions. A workaround is to saturate the Euclidean distance:

$$d\,(\text{reference MCC, detected MCC})$$
$$= \min(\text{T2}; d\,[\text{reference MCC, detected MCC}])$$

where T_2 is a saturation value (an alternative approach would be to use a sigmoid function to compress the Euclidean distance). The motivation for this saturation process is that erring by T_2 (eg, 100) is the same as erring by any value above T_2 (eg, 700).

DISCUSSION

Having in attention the actual state-of-the-art on breast cancer research, FFDM databases are the natural step in the evolution of mammographic databases. As noted by Oliver and colleagues (20), there is no public available database made with digital mammograms. In this work, we address this gap by proposing a FFDM database with a wide variety of findings.

We do acknowledge that not all images in the database respect all quality assessment criteria, because the examination technique and patient related factors have some limitations (eg, previous surgery). Therefore, the database reflects a wide variability of cases and conveys the reality of the routine work of a radiographer.

Despite the fact that our database has a limited number of images, we strongly believe that it is more important to have imaging diversity, than a large number of similar images. Zheng et al (47) claim that, in the development of CAD systems, including difficult cases leads to better results than simply increasing the size of the database with easy masses. Nevertheless, increasing the size of the database will be a future phase of this research.

Annotation is a subjective, tedious, and extremely time-consuming task. Specialists are needed to perform the annotation, which can turn into an extremely difficult and costly task. That is probably the main reason why the currently available databases do not have accurate contours. In the present work, there was a big concern in making precise annotations. However, only two specialists were involved in the process. For that reason, the project will therefore integrate in the database additional specialists contributions to continuously improve the database and annotations quality.

CONCLUSIONS AND FUTURE WORK

We consider that this project has the potential to be a unique work in the field of mammographic databases. Notwithstanding the importance of the digitized databases, technological advances in image acquisition devices for radiology, together with the ubiquity of the computer, led to the development of the FFDM, where the digitalization-related loss of information is absent. Thus, the development of new databases that cover such technological advancements is a crucial step to develop future CADs.

With this database, we aim at increasing available resources in the breast imaging diagnostic field. This updated set of images can be used not only for research purposes, but also in medical practice, for instance, in a teaching environment. Within our team, we are interested in the development of a CAD system. Some MCCs and mass detection methods are currently being implemented

and tested, and classification methodologies are also under development.

Careful annotation is considered as an advantage over the currently available databases. This can motivate computer vision researchers to develop methodologies that take advantage of a better precision of the shape of the lesions to improve detection and/or malignancy classification algorithms. The high image diversity of this database will provide a challenge believed to be difficult to overcome but extremely useful to design more robust CADs. For these reasons, we believe that this database can be a reference for future works in the breast cancer imaging area.

The development of this database in an ongoing work; we plan to extend the number and variability of cases and at the same time, improving the quality of annotations, or even adding more case-related information, depending on the feedback provided by potential users. We are also starting to research the development of CAD systems benefiting from the unique characteristics of INbreast database.

REFERENCES

1. World cancer report. Lyon, France: International Agency for Research and Cancer, 2008.
2. Health statistics: atlas on mortality in the European Union. Luxembourg: Eurostat, 2009.
3. Fact sheet No. 297: Cancer. France: World Health Organization, 2009.
4. Misra S, Solomon NL, Moffat FL, et al. Screening criteria for breast cancer. Adv Surg 2010; 44:87–100.
5. Lee CH, Dershaw DD, Kopans D, et al. Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. J Am Coll Radiol 2010; 7:18–27.
6. European Cancer Observatory. Breast cancer screening programmes in the EU Member States in 2006. Available at: http://eu-cancer.iarc.fr/cancer-13-display-text-562-564.html,en.
7. Karellas A, Vedantham S, Suryanarayan S. Digital mammography: an emerging technology. Business Briefing: Future Directions in Imaging 2006; 2:5. Available at: http://www.touchbriefings.com/pdf/1677/Future_imaging_06_Con.pdf.
8. Sampat MP, Bovik AC, Whitman GJ, et al. A model-based framework for the detection of speculated masses on mammography. Med Phys 2008; 35:2110–2123.
9. Orsi CJ. The American College of Radiology mammography lexicon: an initial attempt to standardize terminology. AJR Am J Roentgenol 1996; 166:779–780.
10. Orel SG, Kay N, Reynolds C, et al. BI-RADS categorization as a predictor of malignancy. Radiology 1999; 211:845–850.
11. Mammography Accreditation Program: Submitting Clinical Images. Preston White Drive, Reston: American College of Radiology, 2009; 20191–4397.
12. Bird RE, Wallace TW, Yankaskas BE. Analysis of cancers missed at screening mammography. Radiology 1992; 184:612–617.
13. Kolb TM, Lichy J, Newhouse JH. Performance of screening mammography among women with and without a first-degree relative with breast cancer. Ann Intern Med 2000; 133:855–863.
14. Zheng B, Hardesty LA, Polle WR, et al. Mammography with computer-aided detection: reproducibility assessment - initial experience. Radiology 2003; 228:58–62.
15. Castellino RA. Computer aided detection (CAD): an overview. Cancer Imaging 2005; 5:17–19.
16. Nishikawa RM. Mammographic databases. Breast Dis 1998; 10:137–150.
17. Orellana CJ. Study of a mammographic CAD performance dependence on the considered mammogram set. Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society, Vancouver, British Columbia, Canada 2008; 4776–4779.
18. Nishikawa RM. Development of a common database for digital mammography research. Chicago, IL: University of Chicago, 1996.
19. Antoniou AC. A web-accessible mammographic image database dedicated to combined training and evaluation of radiologists and machines. 9th International Conference on Information Technology and Applications in Biomedicine, Cyprus, 2009.
20. Oliver A, Freixenet J, Martí J, et al. A review of automatic mass detection and segmentation in mammographic images. Medical Image Anal 2010; 14:87–110.
21. Bernarding J, Thiel A, Decker I, et al. Implementation of a dynamic platform-independent DICOM-server. Comput Methods Programs Biomed 2001; 65:71–78.
22. Singh S, Bovis K. An evaluation of contrast enhancement techniques for mammographic breast masses. IEEE Trans Inform Technol Biomed 2005; 9:109–119.
23. Song E, Xu S, Xu X, et al. Hybrid Segmentation of Mass in Mammograms Using Template Matching and Dynamic Programming. Academic Radiology 2010; vol. 17:1414–1424.
24. Rangayyan RM, Mudigonda N, Desautels J. Boundary modelling and shape analysis methods for classification of mammographic masses. Med Biol Eng Comp 2000; 38:487–496.
25. Suckling J. The Mammographic Image Analysis Society Digital Mammogram Database. In Exerpta Medica. International Congress Series 1069, York, England. 1994; 1069:375–378.
26. Oliver A, Lladó X, Pérez E, et al. A statistical approach for breast density segmentation. J Digital Imaging 2010; 23:527–537.
27. Domínguez AR, Nandi AK. Detection of masses in mammograms using enhanced multilevel-thresholding segmentation and region selection based on rank. In: IASTED International Conference on Biomedical Engineering. Austria: Innsbruck, 2007; 1:370–375.
28. Reljin B, Milosevic Z, Stojic T, et al. Computer aided system for segmentation and visualization of microcalcifications in digital mammograms. Folia Histochem Cytobiol 2009; 47:525–532.
29. Llobet R, Paredes R, Perez Cortes JC. Comparison of feature extraction methods for breast cancer detection. In Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). 2005; II:495.
30. Heath M, Bowyer K, Kopans D, et al. Current status of the Digital Database for Screening Mammography. In Digital Mammography 1998; 457–460.
31. Netsch T, Peitgen HO. Scale-space signatures for the detection of clustered microcalcifications in digital mammograms. IEEE Trans Med Imaging (MedImg) 1999; 18:774–786.
32. Joseph JYL, Gavrielides M, Markey M, et al. Computer-aided classification of breast microcalcification clusters: merging of features from image processing and radiologists. Medical Imaging Conference, San Diego, CA. Proc SPIE 2003; 5032:882–889. Available at: http://adsabs.harvard.edu/abs/2003SPIE.5032..882L.
33. Wang DF, Shi L, Heng PA. Automatic detection of breast cancers in mammograms using structured support vector machines. Neurocomputing 2009; 72:3296–3302.
34. Sakellaropoulos P, Costaridou LI, Panayiotakis G. A wavelet-based spatially adaptive method for mammographic contrast enhancement. Phys Med Biol 2003; 48:787–803.
35. Cheng HD, Cui M. Mass lesion detection with a FUZZY neural network. Patt Recognit 2004; 37:1189–1200.
36. Matheus BR, Schiabel H. Online mammographic images database for development and comparison of CAD schemes. J Digital Imaging 2010; 24:500–506.
37. Lo CS, Chung PC, Chang CI, et al. A computerized system for detection and segmentation of clustered microcalcifications. Joint Conference of International Computer Symposium 1996; 247–253.
38. Oliver A. Automatic mass segmentation in mammographic images. PhD thesis, Universitat de Girona (España) 2008.
39. Oliveira JEE, Gueld MO, de Araújo A, et al. Towards a standard reference database for computer-aided mammography. San Diego, CA: Med imaging Computer-aided diagnosis, 2008. 6915. Available at: http://spie.org/x648.html?product_id=770325.
40. Gardner WD. Breast cancer database provides faster access to patient records. InformationWeek, 2005. Available online at: http://www.informationweek.com/story/showArticle.jhtml;jsessionid=NBXENMVY3EP5AQSNDBGCKHSCJUMEKJVN?articleID=174400322. Accessed Jan. 18, 2011.
41. Tangaro S, Bellotti R, De Carlo F, et al. MAGIC-5: an Italian mammographic database of digitised images for research. Radiol Med 2008; 113:477–485.

42. Warren R, Solomonides T, del Frate C, et al. MammoGrid: a prototype distributed mammographic database for Europe. Clin Radiol 2007; 62: 1044–1051.

43. D'Orsi CJ, Mendelson EB, Ikeda DM, et al. Breast Imaging Reporting and Data System: ACR BI-RADS - Breast Imaging Atlas. Reston, VA: American College of Radiology; 2003. Available at: http://www.acr.org/secondarymainmenucategories/quality_safety/biradsatlas/biradsfaqs.aspx.

44. Danikas D, STheodorou SJV, Kokkalis G, et al. Mammographic findings following reduction mammoplasty. Aesthetic Plast Surg 2001; 25:283–285.

45. Sun L, Li L, Xu W, et al. A novel classification scheme for breast masses based on multi-view information fusion. In Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference, 2010; 1–4.

46. Liu G, Haralick RM. Optimal matching problem in detection and recognition performance evaluation. Patt Recog 2002; 35:2125–2139.

47. Zheng B, Wang X, Lederman D, et al. Computer-aided detection: the effect of training databases on detection of subtle breast masses. Acad Radiol 2010; 17:1401–1408.