Journal of Digital Imaging

Online Mammographic Images Database for Development and Comparison of CAD Schemes

Bruno Roberto Nepomuceno Matheus[1] and Homero Schiabel[1]

Considering the difficulties in finding good-quality images for the development and test of computer-aided diagnosis (CAD), this paper presents a public online mammographic images database free for all interested viewers and aimed to help develop and evaluate CAD schemes. The digitalization of the mammographic images is made with suitable contrast and spatial resolution for processing purposes. The broad recuperation system allows the user to search for different images, exams, or patient characteristics. Comparison with other databases currently available has shown that the presented database has a sufficient number of images, is of high quality, and is the only one to include a functional search system.

KEY WORDS: Mammography, computer-aided diagnosis, images database, image processing, CAD evaluation

INTRODUCTION

According to the World Health Organization, breast cancer is the second most common form of cancer in the world, with a prediction of over 1.5 million diagnoses in 2010 alone and causing more than half a million deaths a year.

So far, the best method for early detection and diagnosis of this disease is the mammographic exam.[1] Unfortunately, mammography requires great skill and care both in image acquisition and evaluation, consuming time and reducing the number of exams a radiologist can evaluate in a given period.

In attempting to speed up the process with gain in quality, several research centers have focused their efforts in computer-aided diagnosis (CAD) schemes for mammographic images, especially in the last few years with the advances of digital mammography. This effort is generating great advances in CAD technology worldwide.

These advances, however, have generated a growing need to test the CAD[2–5] schemes objectively and with sufficient quantity of reference images. These tests would confirm the CAD schemes' efficiency and quality, as required by the Food and Drug Administration,[6] allowing the scheme to be adopted by hospitals and clinics.

Obtaining and selecting those images can be quite a challenge to the developer of CAD schemes. To acquire a sufficient number and variety of images, the developer requires access to confidential files to hospitals and/or clinics with quality images and with trustworthy reports.

To avoid this difficulty, several research laboratories made up images databases for their internal use. Past experience[7,8] has shown that the construction of a proper database for CAD schemes[9] requires a great number of images in easy-to-use formats (by film digitization or by direct digital acquisition—as, for instance, from CR or FFDM systems). Together, the respective medical reports are required in order to sort them efficiently, allowing a fast recovery of the desired images/cases. The database must provide easy

[1]From the Depto. Eng. Electrica – EESC, Universidade de São Paulo, Av. Trabalhador São-carlense, 400, São Carlos, 13566-590 São Paulo, Brazil.

Correspondence to: Bruno Roberto Nepomuceno Matheus, Depto. Eng. Electrica – EESC, Universidade de São Paulo, Av. Trabalhador São-carlense, 400, São Carlos, 13566-590 São Paulo, Brazil; tel: +55-163-4137603; e-mail: bmatheus@sc.usp.br e-mail: bruno.matheus@gmail.com

Online publication 18 May 2010

doi: 10.1007/s10278-010-9297-2

access, especially for the search of images with specific characteristics.

Some papers have shown the existence of several public or private databases already in use by developers[10–13]; however, none of them seems to meet the requirements above.[14,15] Most databases lack a data retrieval system, have stored images of poor quality, or lack the required amount and variety of cases.

This project aims to fulfill the gap with the continuous development of a public mammographic image database using high-quality images with a great variety of diagnosis.

This database is available online at http://lapimo.sel.eesc.usp.br/bancoweb, requiring only a free subscription for full access.

METHODS

Using the data structure presented on 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. This model is represented in Figure 1, where the database configuration data are shown as darker rectangles.

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 accesses of the ‘user’ and also can insert new exams and new regions of interest.
- ‘administrator’, the highest level, has all the accesses above and can also edit, remove, and manage exams, images, and users.

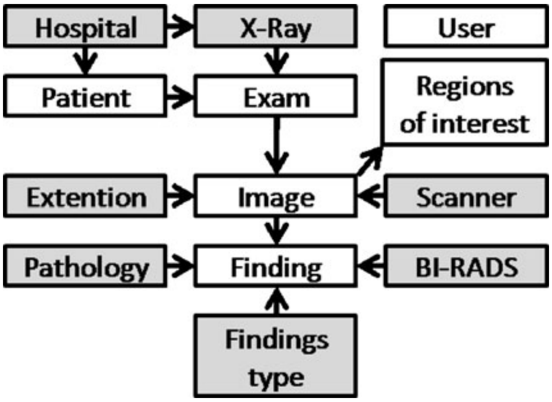


Fig 1. Relational model used for the “BancoWeb” database structure.

In order to access the database, the ‘user’ must register. Once registered, if required, the user may ask for ‘researcher’ status. The request is carefully considered by an ‘administrator’. Preferably, the requesting user must have radiology and mammography experience, must be able to contribute to the database with images or other information, and must also agree to only insert images following the protocols of the database for insertion and quality control. There are no specific protocols in place for this decision, each case being judged separately by the administration, guarantying that only respected and reliable professionals will be granted ‘researcher’ status. This assures the database integrity since only a few highly qualified experts can insert new cases. The ‘administrator’ level is restricted to the database development staff.

The access interface was developed to be easy and intuitive, while assuring data protection. Patient’s data inserted include, when available, 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. The availability of this information depends only on their provision by the hospital or clinic. It must be noted that no personal or identifiable information will be inserted in the database.

For a set of images to qualify for insertion, they must have its findings clearly specified and assured. In this case, anything in the image that the radiologist considered important enough to note in the medical report will be called findings, including benign calcifications, radio-opaque nodules, metallic artifacts, and others.

All findings, “normal” exams, and pathologies are confirmed by previous and later exams, including follow-up mammographies, ultrasound exams, and/or biopsy.

Data and image recovery is done by using a set of four modules, each one associated to another set of data: search by image and exam data, search by patient data, search by type of mammography equipment, and search by type of scanner used. The system is fully integrated, allowing search by any or all of the parameters in any combination of data sets. The modules division is only to facilitate the system use.

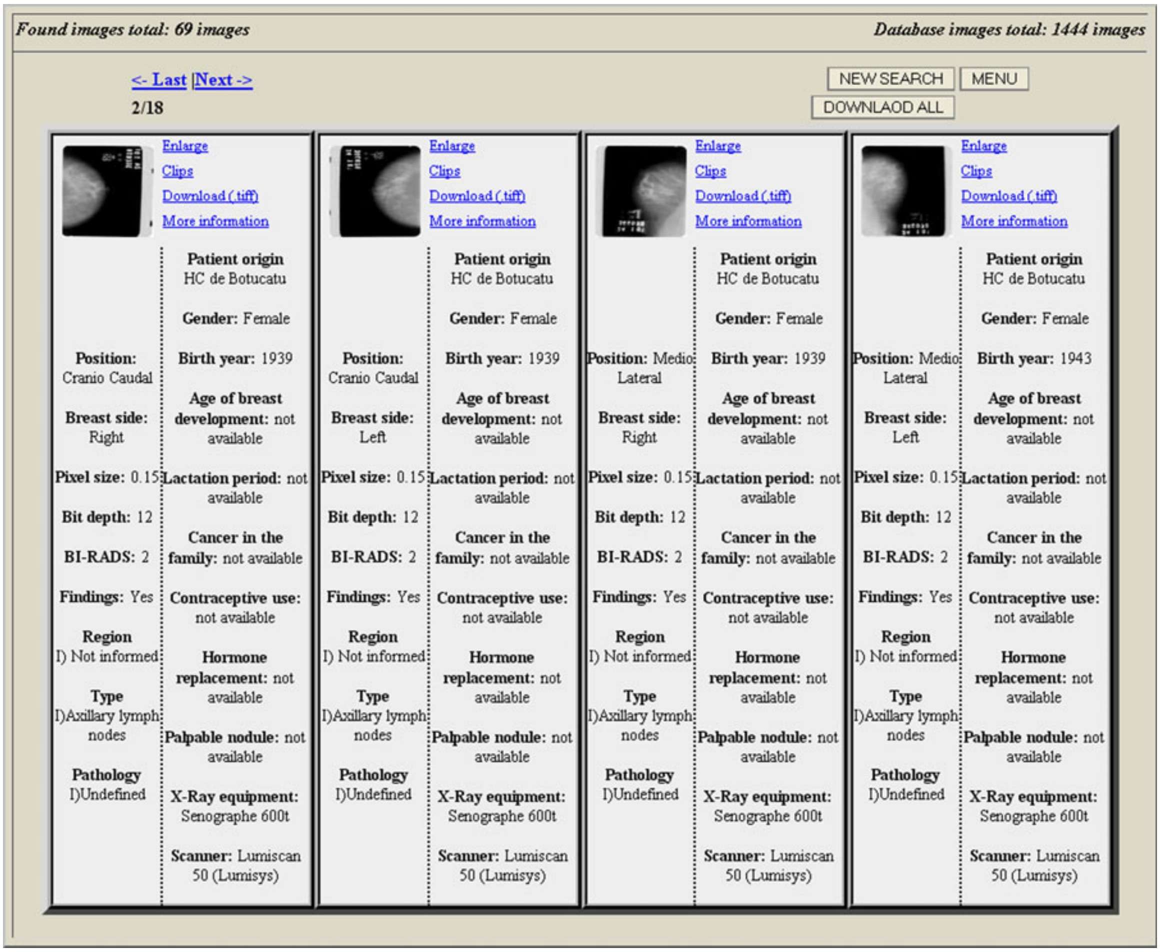
Search results are shown in sets of four images on each page. The most relevant data of the image, patient, and exam is shown in the screen, with an image thumbnail for reference (Fig. 2). For each image, there are also four options to the user: image download, image enlargement, more information recovery, and regions of interest.

The download option allows the user to download the complete image file without losses. These images are usually TIFF files, varying from 8 to 16 bits of contrast (each one represents about 8 MB in average). These are the images that should be used to test processing techniques. If the user wishes to download all the results of a search, it is possible by the use of the "*Download all*" button, presented above the set of results.

Enlargement option allows image visualization (in the screen) in its original size, but compressed with losses, in JPEG format (about 100 KB each image). Due to compression losses, these images are not recommended for processing but only for visual reference before the downloading of the original one.

In the "More information recovery" option, the user has access to more detailed information of the exam and patient, including complete medical report (without personal information), reasons for the exam, and complete set of images of that patient, ordered by exam.

The last available option, "Regions of Interest", links the user to another page to access the regions of interest selecting tools. Initially, a JPEG version



Found images total: 69 images

Database images total: 1444 images

Position: Cranio Caudal
Breast side: Right
Pixel size: 0.15
Bit depth: 12
BI-RADS: 2
Findings: Yes
Region I) Not informed
Type I)Axillary lymph nodes
Pathology I)Undefined

Patient origin HC de Botucatu
Gender: Female
Birth year: 1939
Age of breast development: not available
Lactation period: not available
Cancer in the family: not available
Contraceptive use: not available
Hormone replacement: not available
Palpable nodule: not available
X-Ray equipment: Senographe 600t
Scanner: Lumiscan 50 (Lumisys)

Position: Cranio Caudal
Breast side: Left
Pixel size: 0.15
Bit depth: 12
BI-RADS: 2
Findings: Yes
Region I) Not informed
Type I)Axillary lymph nodes
Pathology I)Undefined

Patient origin HC de Botucatu
Gender: Female
Birth year: 1939
Age of breast development: not available
Lactation period: not available
Cancer in the family: not available
Contraceptive use: not available
Hormone replacement: not available
Palpable nodule: not available
X-Ray equipment: Senographe 600t
Scanner: Lumiscan 50 (Lumisys)

Position: Medio Lateral
Breast side: Right
Pixel size: 0.15
Bit depth: 12
BI-RADS: 2
Findings: Yes
Region I) Not informed
Type I)Axillary lymph nodes
Pathology I)Undefined

Patient origin HC de Botucatu
Gender: Female
Birth year: 1939
Age of breast development: not available
Lactation period: not available
Cancer in the family: not available
Contraceptive use: not available
Hormone replacement: not available
Palpable nodule: not available
X-Ray equipment: Senographe 600t
Scanner: Lumiscan 50 (Lumisys)

Position: Medio Lateral
Breast side: Left
Pixel size: 0.15
Bit depth: 12
BI-RADS: 2
Findings: Yes
Region I) Not informed
Type I)Axillary lymph nodes
Pathology I)Undefined

Patient origin HC de Botucatu
Gender: Female
Birth year: 1943
Age of breast development: not available
Lactation period: not available
Cancer in the family: not available
Contraceptive use: not available
Hormone replacement: not available
Palpable nodule: not available
X-Ray equipment: Senographe 600t
Scanner: Lumiscan 50 (Lumisys)

Fig 2. Image search: search results.

of the image with all the regions of interest recorded in the database after selected by a specialist is displayed. Scrolling the mouse over the ROI information, the reason for its selection is displayed in an extra window. By clicking on the ROI, it can be downloaded in its original, uncompressed format.

Also in this stage a button called “Select Regions of Interest” can be used. This button gives access to a region selection tool, which allows one to mark visually any square region of the image. Once the region is selected, there is the option of downloading the selected ROI and/or inserting it the ROIs database (with the information on the reason of that specific selection).

To insert the ROIs in the database, the user must have ‘researcher’ status, allowing the user that inserted the image to select the regions. Any ‘researcher’ can select the ROIs of any image, so that the regions can be easily selected, even if by other than the original uploader. At the moment, however, very few images have their ROIs previously selected, since this a new tool in the database.

The selection tool was built by using HTML and JavaScript for user interaction and PHP and ImageMagick for actual processing at the server. It was 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 with several different image characteristics, e.g., BI-RADS categories, types of findings, and pathologies. This tool was developed to allow the researcher/developer to check and to evaluate the database.

A very important characteristic to evaluate the quality of a mammographic image is contrast resolution. Contrast resolution should be, at least, of 10 bits (1,024 gray levels). However, contrast resolution superior to 14 bits (16,384 levels) yields little differences in performance to most CAD schemes.[16,17] For this reason, the current database has focused, so far, in 12-bit contrast images (4,096 gray levels).

The default format chosen for the images was TIFF file, based on the amount of available tools for this format as well as due to its versatility, including the possibility of several different contrast levels. This format carries a header that informs one on the image size (lines × columns), level of contrast, compression scheme, and other information.[18] Despite the preference for TIFF images, the database is structured to store any image format.

When dealing with DICOM[1] standard, the information is partitioned and inserted in the database using the normal registration described above. In future developments, the database will allow the user to download the image set of an exam directly in DICOM format.

Currently, images available in our database are originated from two hospitals with two different types of mammography units (a Senographe 500t and a Senographe 600t). Those images were digitized by using two laser scanners—Lumiscan 50 and Lumiscan 75, Lumisys, Inc.—with 12 bits of contrast (gray scale) and 0.085 mm (about 30% of them) and 0.150 mm (70%) of spatial resolution. At this moment, the database does not have any images from Digital Mammography (FFDM), although the database is ready for insertion of such images. As more images are inserted in the database there will be a greater variety, including FFDM images, but at the moment we do not have accesses to FFDM images in great enough number.

For validation purposes, the database was extensively compared with other well-known and broadly used mammographic databases. The databases chosen for the comparison were DDSM (Digital Database for Screening Mammography),[10] MIAS (Mammographic Image Analysis Society Digital Mammogram Database),[11] CALMa (Computer Assisted Library for MAmmography)[12], and LLNL/UCSF database.[13]

The evaluation was performed considering the most recent publication of each database and their current tools available for all possible users.

RESULTS

As described above, the mammographic images database from LAPIMO, which is named from now on simply as *BancoWeb LAPIMO*, is already fully

[1] DICOM: *Digital Imaging Communications in Medicine* is composed of a set of images (usually in TIFF format) and additional information from the exam and apparatus used during image acquisition.

functional and can be accessed in Portuguese or English at http://lapimo.sel.eesc.usp.br/bancoweb.

Currently, this database has around 1,400 images—from around 320 patients. Other 5,000 images are also stored in our servers; they are being progressively transferred to the online database. Most of these images are screening mammography, with only a few being diagnostic work-up.

At this time, the database registers around 100 users. Although the system is ready for remote insertion from anywhere in the world, contacts are yet to be made with hospitals and clinics for direct insertion, which means that at the moment of writing all registered 'researchers' are local to the developing laboratory.

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. At this time, 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 to 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.

As previously mentioned, the database images are 12-bit files (contrast ranging from 0 to 3,500 or 3,800 levels, according to the characteristics provided by the scanners used during the digitization process) and their spatial resolutions are of 0.085 mm or of 0.150 mm, depending on the scanner used. The search system allows selection of the desired characteristics for image download.

These and other characteristics frequently vary between databases, making it difficult to compare CAD schemes from different sources. A possible solution for this constraint would be to test the CAD schemes over a common database.[14,15] Taking this point of view, such a database needs to allow any researcher/developer to access the data, providing a large set of images which statistically represent the most common population characteristics and structures of clinical interest. Thus, our database described here was developed to supply those characteristics, since its access is free and the images set is well representative of the mammography results usually found in the practice.

Table 1 shows the primary results of the comparison with other image databases.

Information regarding UCSF/LLNL and CALMa databases were obtained from the respective papers regarding their description or websites. For this reason, some information is missing (marked with dashes in Table 1).

Considering the amount of image files, the DDSM database surpasses all the other available databases since, even considering the internal data (i.e., for restricted use), the BancoWeb database is only about half the size of DDSM.

Contrast resolution, as mentioned above, should be between 10 and 14 bits.[16,17] In this regard, the MIAS database lacks in quality, offering only 8-bit images. It is important though to remember that

Table 1. Primary Results of the Comparison Between Available Databases. Dash Indicates Unavailable Information

	MIAS	DDSM	UCSF / LLNL	CALMa	Proposed database—BancoWeb
Origin	UK	USA	USA	Italy	Brazil
Amount of images	320	10,480	198	3,000	1,400
File access	Free	Free	Paid (US$ 100)	Closed	Free (requires registration)
Insertion access	Intern	Intern	Intern	Intern	Selected professionals
Contrast resolution (bits/pixel)	8	12 (80%) 16 (20%)	12	12	12
Spatial resolution (μm)	50	42 (19%) 43.5 (44%) 50 (37%)	35	85	75 (32%) 150 (68%)
Image file type	PGM	LJPEG (lossless JPEG)	–	–	TIFF
Search system	No	Yes, but not functional	–	–	Yes

the MIAS database was published in 1994, making it the oldest database among those under comparison here, and very little updates were provided since then. At the moment, the MIAS database is offline and the project seems to have ended.

Considering file types, the LJPEG format used by DDSM is a non-standard version of the free format, making it harder to work with since it requires specific libraries and/or software. Even softwares of large use in this field, as for instance MATLAB®, have no support for this version of LJPEG.

At the moment, the only database with a fully functional and broad search system is the BancoWeb, which we are presenting here. In its website, the DDSM claims to have a search engine, but it is not working and has not been updated since 2003.

Considering regions of interest, the DDSM database presents an overlay marking the region for each image with a finding. The overlay is composed by the coordinates of the region border of each detected structure. On the other hand, BancoWeb database also offers the regions of interest database and its selection tool previously described, although marking all the regions can take some extra time; unfortunately, as stated above, this tool is new and very few ROIS have been marked. This system, however, allows the user to download only the regions of interest, which is an advantage since it avoids, in some cases, the need to download the complete image. Unfortunately, as stated before, this is a new tool and very few images have ROIs selected.

Statistical profiles are also found only at the BancoWeb database, allowing a fast evaluation of content by use of graphics.

Finally, the proposed database is the only one that can allow remote upload of files. In other words, any authorized user ("researcher" or "administrator") can insert exams data (including images) in the base through the Internet. This will eventually allow specialists from different countries to insert directly such data. It is important to notice though that this requires a special authorization that only "administrators" can give.

CONCLUSION

To consistently test CAD schemes, a large amount and variety of images and to compare the schemes use of a single database would be preferable for a coherent evaluation of the results. In addition to the difficulty for accessing hospitals' and clinics' confidential files, this feature shows the need for a large public database with easy access and images of good quality.

The database proposed in this work is structured such as to allow free access by researchers and developers, requiring only online registration and offering them a broad search system and tools for selection of regions of interest and statistical profile.

Comparisons with other databases have shown that the BancoWeb database is the only one having a functional search system. The possibility of searching images with specific findings and/or pathologies helps the test of CAD schemes, allowing the developer to test programs with images containing the specific structure the CAD scheme was designed to detect.

Considering image quality, the BancoWeb database has images of 12 bit in gray scale contrast with spatial resolution between 0.075 mm and 0.150 mm, good enough for CAD analysis.[16,17]

The database still requires a greater variety of images of different spatial and contrast resolutions and the insertion of FFDM images to allow tests with other formats. The same is true about the ROIs database that is still not complete.

Furthermore, the database structure presented here for data insertion and recovery does not need to be restricted to X-ray mammographic images. By simply adding new tables to the model shown in Figure 1, the system could be used for ultrasound and MRI breast images, for example, or even images of other organs for CAD researches in different medical fields. These are indeed some of the possible future developments for this database.

Other future projects include the development of an online CAD system, integrating the database and a CAD scheme to automatically detect regions of interest of all images.

REFERENCES

1. IARC. World cancer report 2008. International Agency for Research on Cancer. Lyon

2. Qian W, et al: Computer assisted diagnosis for digital mammography. IEEE Eng Med Biol 14(5):561–569, 1995

3. Kallergi M: Computer-aided diagnosis of mammographic microcalcification clusters. Med Phys 31(2):314–326, 2004

4. Brzakovic D, Luo XM, Brzakovic P: An approach to automated detection of tumors in mammograms. IEEE Trans Med Imaging 9(3):233–241, 1990

5. Hadjiiski L, Chan H-P, Sahiner B, et al: Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. Radiology 233:255–265, 2004

6. Brown DG: The Evaluation of Computer-Aided Diagnosis Systems: An FDA Perspective. Proceedings of the 30th Applied Imagery Pattern Recognition Workshop (AIPR™01). [S.l.]: [s. n.]. 2001

7. Rangayyan RM, et al: A ROC evaluation of adaptive neighborhood contrast enhancement of digitized mammography. Digital Mammography 307–313. 1994

8. Edwards DC, et al: Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys 29:2861–2870, 2002

9. Nishikawa RM, Yarusso LM: Variations in measured performance of CAD schemes due to database composition and scoring protocol. SPIE 99: Med Imaging 3338:840–844, 1999

10. Heath M, et al: The digital database for screening mammography. Proceedings of the Fifth International Workshop on Digital Mammography, 2001, pp 212–218

11. Suckling J, et al: The mammographic image analysis society digital mammogram database. Exerpta Medica International Congress 1069:375–378, 1994

12. Amendolia SR, et al: The CALMA project. Nucl Instrum Methods Phys Res, Sect A, Accel Spectrom Detect Assoc Equip 461(1-3):428–429, 2001

13. Lawrence Livermore National Library/UCSF Digital Mammogram Database. Center for Health Care Technologies Livermore. Livermore, CA, USA

14. Nishikawa RM, et al: Effect of case selection of the performance of computer-aided detection schemes. Medical Physics 21:265–270, 1994

15. Nishikawa RM: Mammographic databases. Breast Dis 10(3,4):137–150, 1998

16. Nishikawa RM, et al: Performance of automated CAD schemes for the detection and classification of clustered microcalcifications. Digital mammography, Elsevier Science Publishers B.V., 1994, pp 13–20

17. Schiabel H, et al: Performance of a processing scheme for clustered microcalcifications detection with different images database. Digest of Papers of the 2000 World Congress on Medical Physics and Biomedical Engineering. Chicago: [s.n.]. 2000

18. Adobe developers association. TIFF**:** REvision 6.0 - Techinical specification. [S.l.]: [s.n.], 1992