Improving the breast cancer diagnosis using digital repositories

Cesar Suarez Ortega[1] cesar.suarez@ciemat.es, Jose M. Franco Valiente[1], Manuel Rubio del Solar[1], Guillermo Daz Herrero[1], Raul Ramos Pollan[1], Miguel A. Guevara Lopez[2], Naimy Gonzalez de Posada[2], Daniel C. Moura[2], Pedro Cunha[2], Isabel Ramos[3], and Joana Loureiro[3]

[1] Extremadura Research Center for Advanced Technologies (CETA-CIEMAT), Trujillo, Spain
[2] Institute of Mechanical Engineering and Industrial Management (INEGI-FEUP), Faculty of Engineering, University of Porto, Porto, Portugal
[3] Hospital of Sao Joao (HSJ-FMUP), Faculty of Medicine, University of Porto, Porto, Portugal

Abstract. Breast cancer is one of the cancer type most diagnosed. Its causes are unknown so there is not an effective way to prevent it, which increases the mortality rate. The early detection of breast cancer is the best practice to reduce this rate. The double reading of mammograms is a common practice to reduce the rate of missed cancer, but it has a hight cost. Computer Aided-diagnosis (CADx) Systems and Machine Learning Classifiers (MLCs) help to reduce these cost making automatic the second read of the mammograms.
The collaboration between CETA-CIEMAT, INEGI and FMUP-HSJ has generated a set of valuable resources to improve the breast cancer diagnosis process. We aim to achieve a reference repository for breast cancer diagnosis with BCDR, improving the existing implementations by storing a large number of annotated diagnosed cases reviewed by specialists, so researchers can have a reliable source of information for their researches. Using the BCDR data, the main MLCs algorithms are being tested in order to find the best configuration for obtaining accurate automatic diagnosis tool. MIWAD is a workstation which eases the specialist's job in their diagnoses. It is a rich client for BCDR, and offers a set of tools that ease the breast screening and the integration of any MLCs to its workflow. All these resources has been built on top of DRI, a software platform aimed to ease the creation and management of digital repositories over heterogeneous storage.
This work describes the advances made and the future work of the IMED project.

Keywords: Breast cancer; digital repository; clinical data; machine learning classifiers; computer-aided diagnosis

1 Introduction

Breast cancer is the second most common cancer in the world, according the World Health Organization. In 2010 there was about 1.5 million diagnoses, caus-

ing over half a million deaths per year [21]. Nowadays, the causes of breast cancer are unknown, so there is not any effective way to prevent it. However, the diagnosis on the early stages could lead to the patients to a full recovery, being the prevention the main mechanism that have doctors to reduce the mortality of breast cancer.

The double inform of mammograms is a trending practice to reduce the percentage of missed cancer [5], but it has associated high costs. The use of Computer-Aided Diagnosis (CADx) systems could reduce these costs [9], as they can provide an automatic second read to the mammograms [8]. These systems use trained Machine Learning Classifiers (MLC) [5] for the classification of mammograms which use image info and clinical data. The classifiers analyze big sets of data and images and being able to found relationships in them [10]. It is very important to have breast cancer wide-ranging annotated repositories in order to use them as reference for developing new MLCs and CADx systems.

This paper describes the work done by the collaboration between CETA-CIEMAT, INEGI, and FMUP-HSJ to improve the breast cancer diagnosis. This work can be divided in three different modules:

1. Breast Cancer Digital Repository (BCDR): A repository of annotated breast cancer cases.
2. Machine Learning Classifiers (MLCs) for breast cancer diagnosis.
3. Mammography Image Workstation for Analysis and Diagnosis (MIWAD): A specialized workstation prototype for the BCDR data management. It includes a CAD system which integrates the MLCs referenced previously.

2 DRI: Digital Repository Infrastructure

The Digital Repository Infrastructure (DRI) [18] is a software platform aimed at ease the creation and management of digital repositories. A REST API [2] is provided by the platform, so the repository data can be accessed using any programming language or framework. DRI can manage any kind of relational digital repository. In their definition stage, the repositories have to be described using a entity-relationship model defined by XML files. This model describes the digital content of the repository and all the metadata associated.

The main advantage of DRI is its module-based architecture, which allows to store digital content of the repository over heterogeneous storage and to define the data-model easily (using XML files, a well-known format). DRI can be configured to store the digital content in a FTP, a local file system or in Grid Storage Elements [1]. Moreover, developers can create new modules to support additional storage elements. DRI is used as the base of all the tools developed at the IMED project, so all the them can work over any kind of storage supported by DRI.

More details about DRI and its architecture can be found in [20].

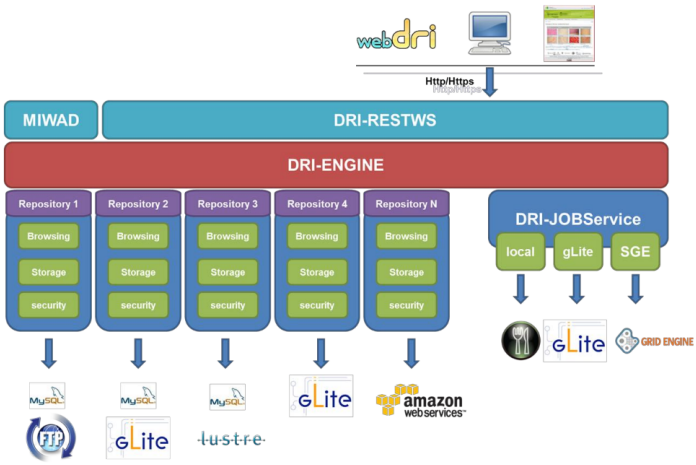

Fig. 1. DRI architecture

3 BCDR: Breast Cancer Digital Repository

As it was explained in the introduction, in order to develop MLCs for the diagnosis of Breast Cancer, a reference dataset is needed for training the classifiers. Currently, we cannot find any suitable repository to develop accurate MLCs, some of them have not enough cases and others are private, so the creation of the Breast Cancer Digital Repository (BCDR) was hosted by the IMED Project in March 2009. BCDR is a wide-ranging annotated repository of anonymous breast cancer studies and currently it is being developed by CETA-CIEMAT, INEGI and FMUP-HSJ. The BCDR has two main objectives:

1. To establish a reference to explore computer-based detection and diagnosis method.
2. To train medical students and other medical-related professionals using as reference the BCDR.

All the stored cases in BCDR has been supplied by FMUP-HSJ. They have been obtained from its real patient's historical archives used to train new specialists (complying with current privacy regulations). The data model used is a subset of the DICOM file format [4] customized by radiologist of the FMUP. This datamodel includes 63 features; 17 clinical features (i.e. age, breast density, lesion location, ...), 23 intensity, texture [7] and shape features [11] for the mediolateral oblique mammogram and the same features for the craniocaudal mammogram. The cases have been reviewed by specialists and are biopsy proven. All the made classifications are defined using the Breast Image Reporting and Data System (BI-RADS) class family [3].

All these collected classifications can be used to train new specialists. Their classifications can be compared to the ones stored in BCDR, which are definitive.

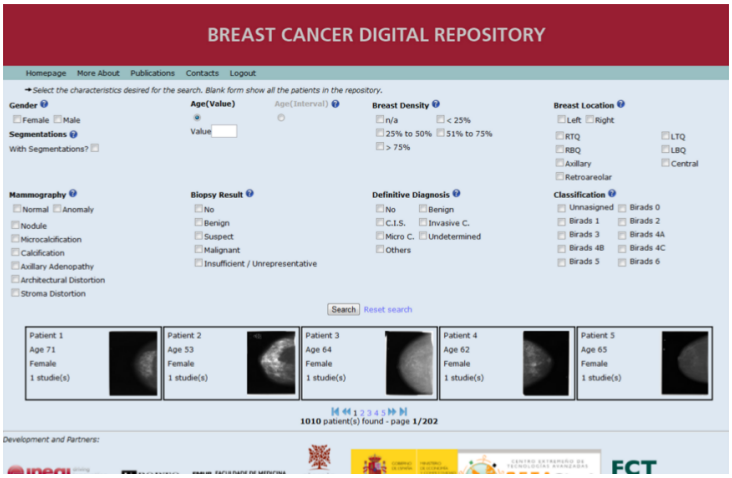

Fig. 2. BCDR web client

BCDR [16] is subdivided in two different repositories: the Film Mammography Repository (BCDR-FMR) and Full Field Digital Mammography Repository (BCDR-DMR). Both repositories use the same data model, but storing the mammograms in different formats. BCDR-FMR uses gray-level digitized TIFF images with a bit depth of 8 bits per pixel and a resolution of 720x1168. The images from BCDR-DMR have better quality: mammograms with a resolution between 3328x4084 and 2569x3328 and a bit depth of 14 bits per pixel. BCDR-DMR also includes ultrasound images with a resolution of 800x600 and a bit depth of 8 bits per pixel.

The BCDR-FMR, at the time of writing, has 1010 cases with their digital content (3073 digitized mammograms) and the associated metadata. Also, there is included the region of the manually segmented lesion by the specialist. There are 795 lesions segmented (masses, microcalcifications, calcifications, stromal distortions, architectural distortions and axillary adenopathies). All this lesions are classified using the BI-RADS class family. BCDR-DMR is composed by 600 cases and 2837 mammograms.

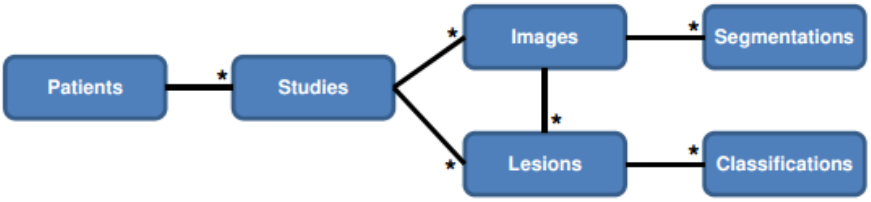

Fig. 3. BCDR data-model

BCDR-FMR is currently available at [19] and BCDR-DMR is under development but it is going to be released soon.

4 Machine Learning Classifiers (MLCs) for breast cancer diagnosis

In order to determine the efficiency of the datasets of BCDR, some MLCs were used for benchmarking purposes. The tests were made using datasets which includes 200 proven lesions of women and 358 segmented images. At the time of writing, three algorithms have been compared: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forests (RF). Our studies [17] reveal that the behavior of the classifiers depends of the amount of training data available. Some of them work better with smaller datasets, such SVM, and others like RF work better for larger datasets. At the time of writing, the best configuration of the MLCs have obtained results greater than the 85% AUC (ROC).

By using BCDR, we are working in the study of more algorithms with the objective of finding the best ones which allows creating more accurate MLCs for breast cancer diagnosis.

5 MIWAD: Mammography Image Workstation for Analysis and Diagnosis

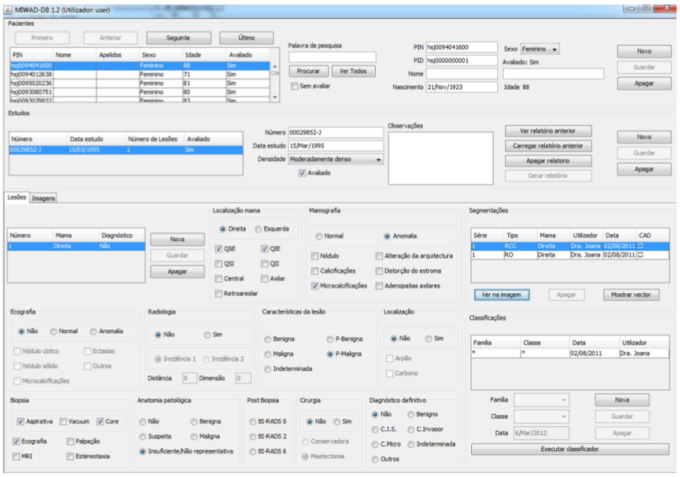

Fig. 4. MIWAD-DB screenshot

MIWAD [15] is a novel software suite for processing, analyzing and diagnosing mammograms, developed under the supervision of the specialists of HSJ. MI-

WAD consists of two interconnected desktop applications. The first one, called MIWAD-DB, allows the management of BCDR data. The second one, called MIWAD-CAD, processes, analyses and diagnoses mammography images by combining digital image processing, pattern recognition and MLCs. MIWAD tools have been created with the following purposes:

1. To feed the BCDR with mammograms and patient clinical data.
2. To build training datasets validated by specialists with patient information and features extracted from the segmented regions performed in the mammograms.
3. To test and improve developed MLCs.

MIWAD-DB implements all the functionalities related with the BCDR by using DRI to access the repository. By using this application, specialists can check the clinical data related to a patient, their studies, the detected lesions and its mammograms. This tool eases the addition of new data to BCDR and has a friendly user interface similar to the clinical reports used at FMUP-HSJ. MIWAD-CAD is connected to MIWAD-DB. It retrieves the all the information related to the studies done to the patient. On the other hand, MIWAD-CAD allows doctors to apply image filters to the mammograms to ease the lesion detection. The main feature of MIWAD-CAD is his auto-adjusting segmentation tool, which enables doctors to segment accurate region of interest (ROI) from mammograms with a few mouse clicks.

With the data provided in MIWAD-DB and the segmentations made in MIWAD-CAD, the specialists have enough information to classify the lesions of a study using BI-RADS family class. These classifications are stored at BCDR in order to enlarge the datasets using to train MLCs. Moreover, MIWAD provides a Java API to easily integrate third-party classifiers to the system as a new module. This API has been used to integrate the MLCs for breast cancer diagnosis described in the the previous section, so the doctors can start testing them in a few steps.

MIWAD can be a very valuable resource to train new radiologists. It can be used to compare easily the classifications made by novel specialists with the ones of the BCDR, which have been reviewed by trained specialists and biopsy proven. We are planning to integrate MIWAD in the formation of new specialists at FEUP.

The MIWAD applications has been implemented using Java, which means that they work correctly in Linux, Windows and MacOS.

6 Conclusions and future work

New datasets are being added to BCDR-FMR by the specialists of FMUP-HSJ using MIWAD applications. Also, BCDR-DMR is being developed, and we expect to have between 2000 and 3000 breast cancer patients cases in a short time. MIWAD is very valuable resource to enhance the collaboration among specialists, offering a user-friendly desktop application which feeds the BCDR

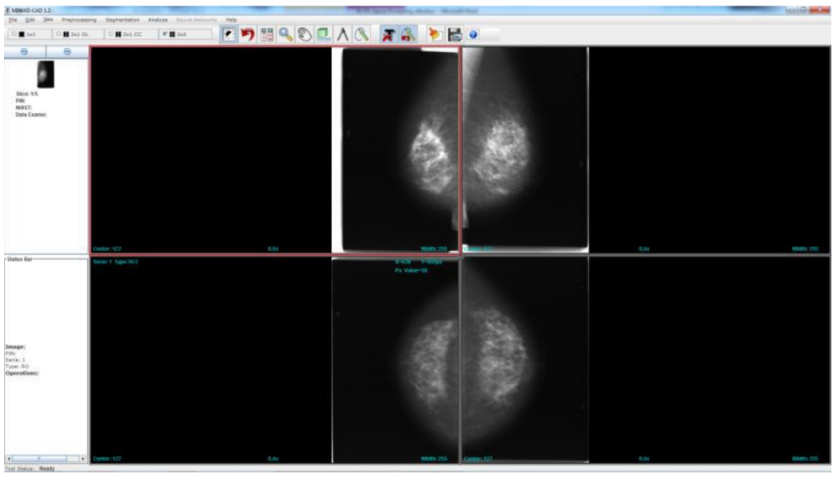

Fig. 5. MIWAD-CAD screenshot

with data reviewed by specialists. Also it provides a set of tools for assisting doctors (i.e. the auto-adjusting segmentation tool). It also features a transparent integration with any MLC. Now, we are working in a DICOM importer/exporter tool for MIWAD, so we can integrate our workstation with third-party systems (i.e. PACS), to add data new resources of data for BCDR and to offer integration with the MIWAD plugged MLCs

We believe that BCDR may become a standard repository for comparison of MLCs for breast cancer diagnosis due its rich feature set, the reliability of its data sources and the high volume of data storage. Also it is a very valuable tool to train new radiologist specialists.

7 Acknowledgments

This work is part of the IMED research collaboration project between CETA-CIEMAT(Spain), INEGI and FMUP-HSJ (Portugal). The three institutions express their gratitude for the supoort of the European Regional Development Fund.

Prof. Guevara acknowledges POPH – QREN – Tipologia 4.2 – Promotion of scientific employment funded by the ESF and MCTES, Portugal.

References

1. Foster I. and C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers, 1998.
2. R. Fielding and R. Taylor. Principled Design of the Modern Web Architecture. ACM Trans. Internet Technol. 2, 2 (May 2002), 115-150.

3. D'Orsi, C. J., et al. Breast Imaging Reporting and Data System: ACR BI-RADS mammography, 4th Edition ed.: American College of Radiology, 2003.
4. NEMA. (2010). Digital Imaging and Communications in Medicine. Available `http://dicom.nema.org/`
5. R. Ramos-Pollan, et al., Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis. Journal of Medical Systems (9 April 2011), pp. 1-11.
6. Hall M, et al. The WEKA Data Mining Software: An Update; SIGKDD Explorations, 2009, 11, p. 10-18.
7. Haralick RM et al. Textural Features for Image Classification. IEEE Transactions on Systems Man and Cybernetics. 1973, 2, p. 610-621.
8. Huo ZM, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms. Academic Radiology. 1998, 5, p. 155-168.
9. Joshua J, et al. Effectiveness of Computer-Aided Detection in Community Mammography Practice. NCI J Natl Cancer Inst. 2011, 103, p. 1152-1161.
10. Wang D, Shi L, Heng PA. Automatic detection of breast cancers in mammograms using structured support vector machines. Neurocomputing. 2009, 72, p. 3296-3302.
11. Sahiner B, Chan HP, Petrick N, Helvie MA, Hadjiiski LM. Improvement of mammographic mass characterization using spiculation measures and morphological features. Medical Physics. 2001, 28, p. 1455-1465.
12. B.R. Matheus, H. Schiabel, Online Mammographic Images Database for Development and Comparison of CAD Schemes, Journal of Digital Imaging. 24 (2011) 500-506.
13. I.C. Moreira, I. Amaral et al. INbreast: Toward a Full-dield Digital Mammographic Database, Academic radiology. 19 (2012) 236-248.
14. A. Marcano-Cedeo, J. Quintanilla-Dominguez et al. WBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Systems with Applications. 38 (2011) 9573-9579.
15. Jose M. Franco Valiente, Cesar Suarez Ortega et al., MIWAD: A Software Suite For Building CAD Methods. 15th International Conference on Experimental Mechanics (22-17 July 2012), Porto, Portugal
16. Miguel A. Guevara Lopez, Naimy Gonzalez de Posada et al., BCDR: A Breast Cancer Digital Repository. 15th International Conference on Experimental Mechanics (22-17 July 2012), Porto, Portugal
17. Daniel C. Moura, Miguel A. Guevara Lopez et al., Classifier Performance Vs Dataset Size: A Comparative Study For Breast Lesions Classification. 15th International Conference on Experimental Mechanics (22-17 July 2012), Porto, Portugal
18. A. Calanducci, F. Prieto, et al. 2008. Enabling Digital Repositories on the Grid. Proceedings of the Second International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP '08), 2008.
19. BCDR webpage: `http://bcdr.inegi.up.pt/`
20. DRI webpage: `http://dri.ceta-ciemat.es/`
21. V. Veloso, Cancro da mama mata 5 mulheres por dia em Portugal, In: (Ed.) CinciaHoje, Lisbon, Portugal, 2009