PAPER REF: 3004

BCDR: A BREAST CANCER DIGITAL REPOSITORY

Miguel A. Guevara López[1(*)], Naimy González de Posada[2], Daniel C. Moura[3], Raúl Ramos Pollán[4], José M. Franco Valiente[5], César Suárez Ortega[6], Manuel R. del Solar[7], Guillermo Díaz Herrero[8], Isabel M.A. Pereira Ramos[9], Joana Pinheiro Loureiro[10], Teresa Cardoso Fernandes[11], Bruno M. Ferreira de Araújo[12]

[1,2,3]Institute of Mechanical Engineering and Industrial Management (INEGI-FEUP), Faculty of Engineering, University of Porto, Porto, Portugal

[4,5,6,7 8]Extremadura Research Center for Advanced Technologies (CETA-CIEMAT), Trujillo, Spain

[9,10,11,12] Faculty of Medicine - Hospital of Sao Joao (FMUP-HSJ), University of Porto, Porto, Portugal

[(*)]*Email:* mguevaral@inegi.up.pt

ABSTRACT

This paper outlines the first Portuguese “Breast Cancer Digital Repository” (BCDR-FMR), a comprehensive annotated repository including digital content (digitized film mammography images) and associated metadata (clinical history, segmented lesions BI-RADS classified, image-based descriptors, biopsy proven, etc.). BCDR-FMR establish a novel reference to develop breast cancer computer-aided detection / diagnosis methods and for training medical students, formed physicians and other medical-related professionals.

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 and causing more than half a million deaths per year (B.R. Matheus, H. Schiabel et al., 2011). In the European Union it is responsible for one in every six deaths from cancer in women (I.C. Moreira, I. Amaral et al., 2012). In Portugal, each year, 4,500 new cases of breast cancer are diagnosed and 1,600 women are estimated to die from this disease (V. Veloso, 2009).

Double reading of mammograms (two radiologists read the same mammograms) has been advocated to reduce the proportion of missed cancers. But the workload and costs associated with double reading are high. Instead, Computer-Aided Diagnosis (CADx) systems can assist one single radiologist when reading mammograms providing support to their diagnosis (Ramos-Pollán, Guevara-Lopez et al., 2011a). Many research centers have focused their efforts in applications of CADx approaches for different modalities of breast imaging and associated metadata, being the correct patterns classification of breast cancer an important real-world medical problem. For this reason the use of Machine Learning Classifiers (MLC) in medical diagnosis is gradually increasing (A. Marcano-Cedeño, J. Quintanilla-Domínguez et al., 2011). MLC can explain complex relationships in the data and constitute the backbone of biomedical data analysis on high dimensional quantitative data provided by the state-of-the-art medical imaging and high-throughput biology technologies (Ramos-Pollán, Guevara-Lopez et al., 2011b).

Therefore, it is essential to produce breast cancer comprehensive annotated repositories properly assembling digital content and associated metadata (biopsy proven) that can be used as reference. This enables: (1) modeling and exploring MLC and (2) comparing the performance of breast cancer CADx methods.

This paper presents a Breast Cancer Digital Repository (BCDR-FMR) designed together with expert radiologists, following the ACR and BI-RADS standards (D'Orsi, C. J., et al., 2003). The BCDR-FMR is a repository representative of patients in northern Portugal. It is in continuous development and, at time of writing, it is composed of 1010 patient's cases including digitized film mammography images, clinical history, BI-RADS classification, biopsy proven, segmented lesions and selected pre-computed image-based descriptors. This research aims to the development of a comprehensive annotated breast cancer digital repository with two main objectives: (1) to establish a novel reference to develop breast cancer computer-aided detection / diagnosis methods and (2) to train medical students, formed physicians and other medical-related personnel involved in the diagnostic, treatment or research of breast cancer and associated technologies.

The remaining of the article is structured as follows: the Materials and Methods section, describes the data model and the main characteristics of the new developed breast cancer digital repository. In the Results and Discussion section are discussed the results obtained by the exploration and training of MLC carried on BCDR-FMR. Finally, conclusions and directions of future work are presented in the last section.

MATERIALS AND METHODS

The "Breast Cancer Digital Repository" (BCDR-FMR), the first Portuguese Breast Cancer database, was built with real patient's historical archives (complying with current privacy regulations as they are also used to teach regular and postgraduate medical students) supplied by the Faculty of Medicine – "Hospital São João", University of Porto, Portugal (FMUP). The BCDR-FMR data model is a subset of the DICOM medical file format (NEMA, 2010) customized by radiologists of the FMUP for storing and managing specific patient information related to digital mammography images (see data model in Fig. 1). The data model (Ramos-Pollán, Guevara-Lopez et al., 2010), (Ramos-Pollán, Guevara-Lopez et al., 2011a) supports each patient undergoing one or more studies, each study composed of one or more images (such as digitized film mammography images) and one or more lesions. Each image may have one or more segmentations (for different lesions) and each lesion can be associated to many segmentations, typically in mediolateral oblique (MLO) and craneocaudal (CC) images of the same breast. Moreover, each lesion can also be classified by several different evaluators, such as radiologists and machine learning classifiers (MLCs). Biopsy results are also available for each lesion, providing a ground-truth classification that can be used to evaluate the performance of radiologists and MLCs, as well as for training. Currently, for each segmented region, 24 features are automatically computed and stored making a features vector, which is representative of the image region statistics, shape and texture. Features vectors are associated to classifications to provide supervised datasets that may be used for training MLCs. In this work we consider only the Breast Image Reporting and Data System (BI-RADS) class family (D'Orsi, C. J., et al., 2003). BCDR-FMR allows also the storage of a variety of sets of experiments of classification runs, performed both by human experts and MLC, so that they become available for statistical analysis.

In order to evaluate the database for the task of classifying lesions as potentially malignant, 1439 instances of the database (corresponding to 1439 segmentations) were used to build a Linear Support Vector Machine classifier (SVM). Leave-one-out cross-validation was used to assess the performance of the classifier. Segmentations of lesions with BI-RADS score greater or equal than 4 were considered potentially malignant. The average area under the receiver operating characteristic curve (AUC) was the evaluation metric for this study.

RESULTS AND DISCUSSION

The BCDR-FMR has at present a total of 1010 cases including digital content (3703 digitized film mammography images) and associated metadata. Precise segmentations of identified lesions (manual contours made by medical specialists) are also provided. In total 795 lesions were segmented in MLO and CC images to produce a total of 1493 segmentations, in which are included: masses = 639, microcalcifications = 341 calcifications = 145, stromal distortions = 102, architectural distortions = 66, and axillary adenopathy = 2. Distribution of the lesions classification and patient age are presented in Fig. 2 and 3 respectively.

Currently, we have registered biopsy results for 276 lesions (163 benign and 113 malign), and this number is rapidly increasing. Lesion classification with SVM resulted in an average AUC of 0.85, demonstrating the suitability of the BCDR-FMR for this task. Furthermore, the performance of the classifier may be potentially increased since no parameter fine-tuning was performed, and no kernel was used to try capturing non-linear relations.

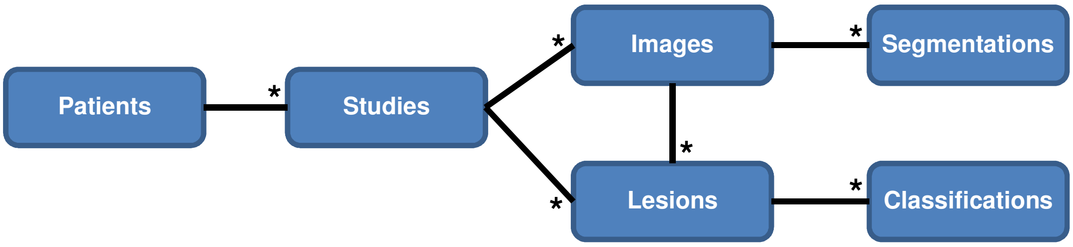

Fig.1. BCDR-FMR Data model

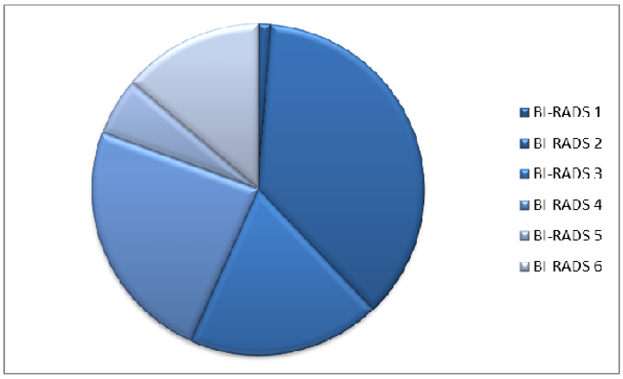

Fig.2. Distribution of BI-RADS classifications in BCDR-FMR

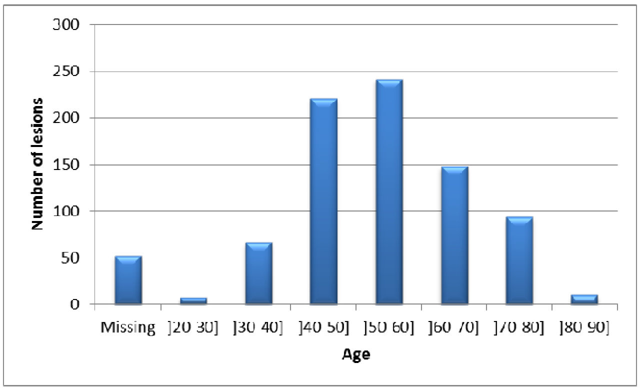

Fig.3. Age distribution in BCDR-FMR

CONCLUSIONS

The strengths of the actually presented BCDR-FMR, differentiate it from other reported databases, by including for each segmented lesion a set of pre-computed shape, intensity and texture attributes, which enhance and improve the lesion classification options. We consider that BCDR-FMR is a comprehensive reference for future works focused on the development of breast cancer computer-aided detection/diagnosis methods.

Furthermore, the BCDR-FMR has been validated by radiologists at the Faculty of Medicine – Hospital São João at University of Porto, Portugal (FMUP-HSJ) which makes it suitable in the process of training medical students, formed physicians and other medical-related professionals.

ACKNOWLEDGMENTS

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

NEMA. (2010). “Digital Imaging and Communications in Medicine”. Available: http://dicom.nema.org/ Last visited December 2011.

R. Ramos-Pollan, M. Guevara-Lopez et al., "Exploiting e-Infrastructures for medical image storage and analysis: A Grid application for mammography CAD" in The Seventh IASTED International Conference on Biomedical Engineering, Innsbruck, Austria, 2010.

R. Ramos-Pollan, M. Guevara-Lopez et al., “Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis” Journal of Medical Systems (2011a), pp. 1-11.

R. Ramos-Pollán, M. Guevara-López et al., "A Software Framework for Building Biomedical Machine Learning Classifiers through Grid Computing Resources", Journal of Medical Systems. (2011b) 1-13.

D'Orsi, C. J., et al., "Breast imaging reporting and data system: ACR BI-RADS-mammography", 4th Edition ed.: American College of Radiology, 2003.

B.R. Matheus, H. Schiabel, "Online Mammographic Images Database for Development and Comparison of CAD Schemes", Journal of Digital Imaging. 24 (2011) 500-506.

I.C. Moreira, I. Amaral et al., "INbreast: Toward a Full-field Digital Mammographic Database", Academic radiology. 19 (2012) 236-248.

V. Veloso, "Cancro da mama mata 5 mulheres por dia em Portugal", in: In: (Ed.) CiênciaHoje, Lisbon, Portugal, 2009.

A. Marcano-Cedeño, J. Quintanilla-Domínguez et al., "WBCD breast cancer database classification applying artificial metaplasticity neural network", Expert Systems with Applications. 38 (2011) 9573-9579.