Benchmarking Datasets for Breast Cancer Computer-Aided Diagnosis (CADx)

Daniel Cardoso Moura[1], Miguel Angel Guevara López[2,1], Pedro Cunha[1],
Naimy González de Posada[1], Raúl Ramos Pollan[3], Isabel Ramos[4],
Joana Pinheiro Loureiro[4], Inês C. Moreira[4], Bruno M. Ferreira de Araújo[4],
and Teresa Cardoso Fernandes[4]

[1] INEGI - Institute of Mechanical Engineering and Industrial Management,
Faculty of Engineering, University of Porto, Portugal
{dmoura,mguevaral,pcunha,nposada}@inegi.up.pt
[2] DETI - Department of Electronics, Telecommunications and Informatics,
University of Aveiro, Portugal
mguevaral@ua.pt
[3] Universidad Industrial de Santander, Bucaramanga, Colombia
rramosp@unal.edu.co
[4] FMUP-HSJ – Faculty of Medicine – Centro Hospitalar São João, University of Porto, Portugal
radiologia.hsj@mail.telepac.pt,
{joanaploureiro,ines.c.moreira}@gmail.com,
{bruno.araujo,teresafernandes}@hsjoao.minsaude.pt

Abstract. Designing reliable computer-aided diagnosis (CADx) systems based on data extracted from breast images and patient data to provide a second opinion to radiologists is still a challenging and yet unsolved problem. This paper proposes two benchmarking datasets (one of them representative of low resolution digitized Film Mammography images and the other one representative of high resolution Full Field Digital Mammography images) aimed to (1) modeling and exploring machine learning classifiers (MLC); (2) evaluating the impact of mammography image resolution on MLC; and (3) comparing the performance of breast cancer CADx methods. Also, we include a comparative study of four groups of image-based descriptors (intensity, texture, multi-scale texture and spatial distribution of the gradient), and combine them with patient's clinical data to classify masses. Finally, we demonstrate that this combination of clinical data and image descriptors is advantageous in most CADx scenarios.

Keywords: Breast cancer, image-based descriptors, clinical data, machine learning classifiers, computer-aided diagnosis (CADx), histograms of gradient divergence.

1 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 [1]. At present, there are no

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 326–333, 2013.

effective ways to prevent breast cancer, because its cause remains unknown. However, efficient diagnosis of breast cancer in its early stages can give women better chances of full recovery. Screening mammography is the primary imaging modality for early detection of breast cancer because it is the only method of breast imaging that consistently has been found to decrease breast cancer-related mortality [2].

Double reading of mammograms (two radiologists read the same mammogram) has been advocated to reduce the proportion of missed cancers. But the workload and costs associated with double reading are high. Therefore, many research institutions have focused their efforts in applications of Computer-Aided Diagnosis (CADx) approaches combining mammography image-based descriptors 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 breast cancer diagnosis is gradually increasing [3]. 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 [4].

While several produced mammography-based breast cancer databases (public or private) have been reported [1], [5-12], currently, the information included in these databases presents some undesirable issues: a) lesions are not exactly identified; b) are incomplete in terms of available features (pre-computed image-based descriptors and clinical data); c) have a reduced number of annotated patient's cases; and/or d) the database is private and cannot be used as reference. Altogether, these issues make it difficult producing golden standard datasets assembling properly extracted information of breast cancer lesions (biopsy proven) for assessing and comparing the performance of machine learning classifiers (MLC) and Breast Cancer CADx methods.

In preceding works, first we made an exploration of mammography-based MLC [13] and hereafter we made an evaluation of several groups of mammography image-based descriptors, clinical data, and combinations of both types of data for classifying microcalcifications, masses and all lesions together on two different Film mammography-based datasets [14]. As result, we obtained MLC with high performance and it was proposed a novel image-based descriptor that is especially designed for round-shape objects, such as masses, the Histograms of Gradient Divergence (HGD).

This paper proposes two benchmarking datasets (one of them representative of low resolution Film Mammography images and the other one representative of high resolution Full Field Digital Mammography (FFDM) images) aimed to: (1) modeling and exploring machine learning classifiers (MLC); (2) evaluating the impact of mammography image resolution on MLC; and (3) comparing the performance of breast cancer CADx developed methods. Also, it is included a comparative study of four groups of image-based descriptors (intensity, texture, multi-scale texture and spatial distribution of the gradient), and their combination with patient's clinical data to classify masses. The two benchmarking datasets used in this work are available for public domain at the Breast Cancer Digital Repository (BCDR – http://bcdr.inegi.up.pt) and it is the first experiment made on the FFDM-based dataset (BCDR-D01).

2 Materials and Methods

2.1 Benchmarking Datasets

The two benchmarking datasets proposed here were extracted from the Breast Cancer Digital Repository (BCDR). The BCDR is a wide-ranging annotated public repository composed by Breast Cancer patients' cases of the northern region of Portugal.

BCDR is subdivided in two different repositories: (1) a Film Mammography-based Repository (BCDR-FM) and (2) a Full Field Digital Mammography-based Repository (BCDR-DM). Both repositories were created with anonymous cases from medical archives (complying with current privacy regulations as they are also used to teach regular and postgraduate medical students) supplied by the Faculty of Medicine – Centro Hospitalar São João, at University of Porto (FMUP-HSJ). BCDR provides normal and annotated patients cases of breast cancer including mammography lesions outlines, anomalies observed by radiologists, pre-computed image-based descriptors as well as related clinical data.

The BCDR-FM is composed by 1010 (998 female and 12 male) patients cases (with ages between 20 and 90 years old), including 1125 studies, 3703 mediolateral oblique (MLO) and craniocaudal (CC) mammography incidences and 1044 identified lesions clinically described (820 already identified in MLO and/or CC views). With this, 1517 segmentations were manually made and BI-RADS classified by specialized radiologists. MLO and CC images are grey-level digitized mammograms with a resolution of 720 (width) by 1168 (height) pixels and a bit depth of 8 bits per pixel, saved in the TIFF format.

The BCDR-DM, still in construction, at the time of writing is composed by 724 (723 female and 1 male) Portuguese patients cases (with ages between 27 and 92 years old), including 1042 studies, 3612 MLO and/or CC mammography incidences and 452 lesions clinically described (already identified in MLO and CC views). With this, 818 segmentations were manually made and BI-RADS classified by specialized radiologists. The MLO and CC images are grey-level mammograms with a resolution of 3328 (width) by 4084 (height) or 2560 (width) by 3328 (height) pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in the TIFF format.

The **BCDR-F01** dataset is built from BCDR-FM and is formed by 200 lesions: 100 benign and 100 malignant (biopsy proven) and it is composed by a total of 358 features vectors (184 instances related to the 100 benign lesions and 174 instances related to the 100 malignant lesions).

The **BCDR-D01** dataset is built from BCDR-DM and is formed by 79 lesions: 49 benign and 30 malignant (biopsy proven) and it is composed by 143 features vectors (86 instances related to the 49 benign lesions and 57 instances related to the 30 malignant lesions).

Both datasets (currently, available for download at the BCDR website) are composed by instances of the same clinical, intensity, texture, multi-scale texture and spatial distribution of the gradient features. Clinical features include the patient age, breast density and a set of selected binary attributes for indicating abnormalities observed by radiologists, namely masses, microcalcifications, calcifications (other than microcalcifications), axillary adenopathies, architectural distortions, and stroma distortions. Thus, the clinical data for each instance of the datasets is formed by a total of 8 attributes per instance: 6 binary attributes related to observed abnormalities, an ordinal attribute for breast density, and a numerical attribute that contains the patient age at the time of the study. The same group of image-based features (intensity, texture, multi-scale texture and spatial distribution of the gradient) that we reported in [14] were utilized here, namely, Intensity statistics, Histogram measures, Invariant moments, Zernike moments, Haralick features, Grey-level run length (GLRL) analysis, grey-level differences matrix (GLDM), Gabor filter banks, Wavelets, Curvelets, Histograms of Oriented Gradient (HOG), and Histograms of Gradient Divergence (HGD). For the sake of brevity, the reader is addressed to [14] for a formal description of the descriptors and the range of parameters evaluated. All descriptors were computed from rectangular patches of the lesions that were generated by extracting the part of the mammogram within the bounding box of the outlines provided by both datasets.

2.2 Evaluation of the Benchmarking Datasets

For evaluating the datasets, and delivering baseline benchmarks for CADx, an experiment was conducted for classifying masses.

Classification was performed using several machine learning classifiers available on Weka version 3.6 [15], namely Support Vector Machines (SVM), Random Forests (RF), Logistic Model Trees (LMT), K Nearest Neighbours (KNN), and Naive Bayes (NB). For all classifiers with the exception of NB (which is parameterless), 3-fold cross-validation was performed on the training set for optimizing the classifiers parameters. Linear SVM was chosen for simplicity and speed with regulation parameter C ranging from 10^{-2} to 10^{3}. The number of trees of RF was optimized between 50 and 400, with each tree having $\log_2(A) + 1$ randomly selected attributes, where A is the number of attributes available in the current dataset. On LMT the number of boosting iterations was also optimized. Finally, the number of neighbours (K) of KNN varied from 1 to 20, and the contribution of each neighbour was always weighted by the distance to the instance being classified. For all classifiers, attribute range normalization [0..1] was performed as pre-processing with the minimum and maximum values of the attributes found in the training set and then applied to both train and test sets.

The evaluation measure used was the Area Under the Curve of the Receiver Operator Characteristic (AUC). Resampling without replacing was performed 50 times for each view (MLO and CC) resulting in 100 runs per experiment to provide different splits across training and test sets, with 80% of the cases randomly selected for training the classifier, and the remaining 20% used for test. The two views were trained and tested independently to prevent biasing results and finally the AUCs from both views were merged resulting in a total of 100 evaluations per experiment. When comparing descriptors, the best combination of parameters' values and classifier was used. Comparisons between the experiments were based on the median AUC of the 100 runs (mAUC) and were supported by Wilcoxon signed rank tests to determine whether differences have statistical evidence ($p < 0.05$). The experiment was done for both BCDR-F01 and BCDR-D01 datasets, for all descriptors including and excluding clinical data, and for all classifiers.

3 Results and Discusion

Results for the BCDR-F01 were previously published on [14] and are included here for reference and comparison with BCDR-D01. On the BCDR-F01 dataset (Fig.1), the standalone clinical data had a performance of mAUC=0.829, and was only outperformed by HGD (mAUC=0.860, p<0.001). When combining clinical data with image descriptors, four descriptors (Intensity Histograms, Zernike, GLRL and GLDM) did not show evidence of increasing the performance of standalone clinical data. All the remainder descriptors outperformed clinical data (p<0.028), with HGD being the best (mAUC=0.894). HGD performance was significantly superior to all the remainder (p<0.024). All descriptors significantly improved performance (p<0.001) when combined with clinical data.

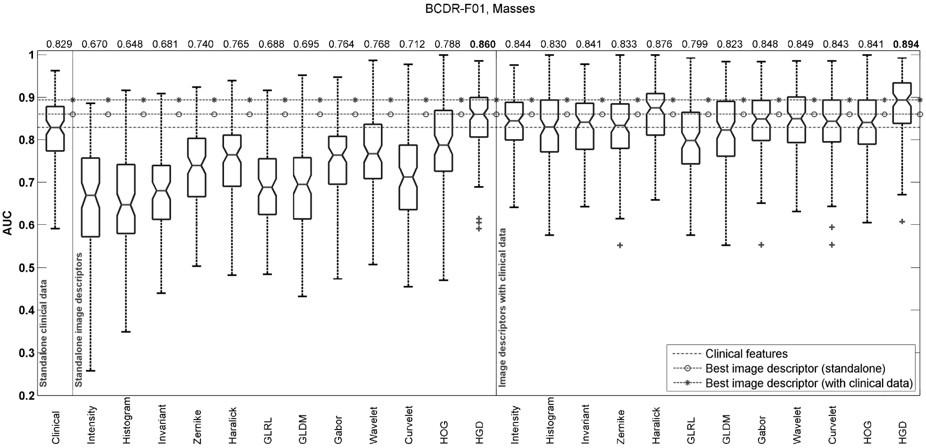

Fig. 1. Final rankings for all the evaluated descriptors on the BCDR-F01 dataset (image-based standalone and image-based combined with clinical data)

On the BCDR-D01 dataset (Fig.2), in contrast to the previous dataset, the standalone clinical data was outperformed by all descriptors (p<0.001). It was observed that this difference was mainly explained by an increase of performance of the image descriptors on the BCDR-D01 dataset, rather than a decrease of performance of the standalone clinical data. Haralick features scored the highest score with mAUC=0.938, and both Wavelets and HGD did not show statistical evidence of significant differences to Haralick features, with p=0.943 and p= 0.710 respectively. When clinical data is combined with the image descriports, Haralick features remains the descriptor with the highest score (mAUC=0.965), with Wavelets (p=0.241), Curvelets (p=0.889) and HGD (p=0.585) not showing statistical evidence of significant differences to Haralick features. All descriptors significantly improved performance (p<0.003) when combined with clinical data, with the exception of GLDM and HOG.

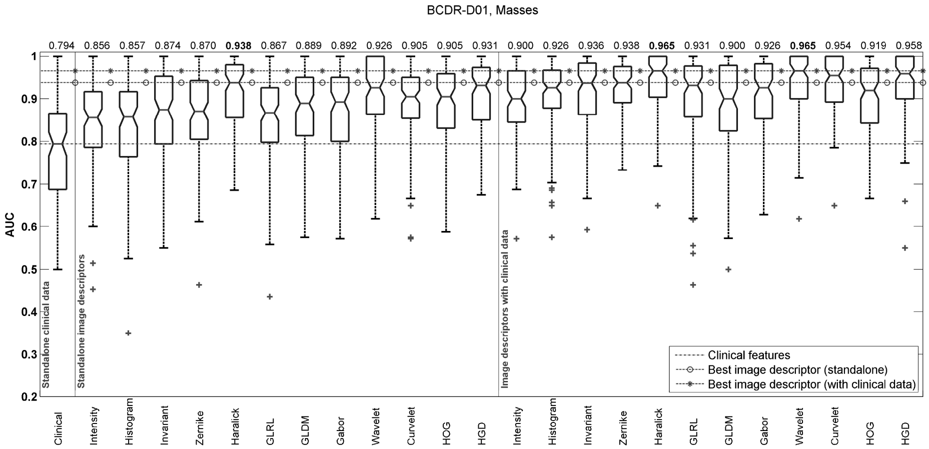

Fig. 2. Final rankings for all the evaluated descriptors on the BCDR-D01 dataset (image-based standalone and image-based combined with clinical data)

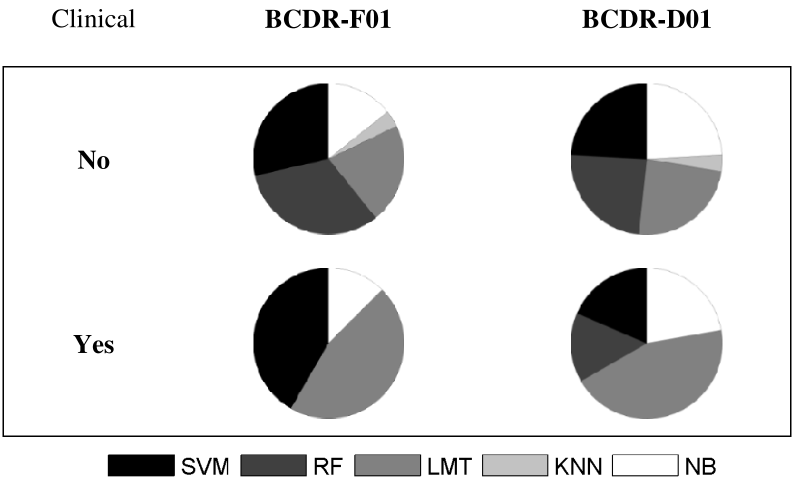

Fig. 3. Distribution of wins among classifiers on both datasets in the absence and presence of clinical data. A classifier scores a win for each time it enables to reach the best result for a given descriptor or when there is no evidence of significant differences to the best results (Wilcoxon test, p>0.05).

Regarding the performance, it can be observed (Fig. 3) that the performance of classifiers is more sensible to the absence or presence of clinical data, rather than to the dataset that was used BCDR-FM (Film) vs BCDR-DM (digital). KNN scored the worst results with performances ranging from 0% to 3% of the wins. When clinical data is not included in the datasets, the frequence of wins is equaly distributed by all the remainder classifiers on BCDR-D01 with 24% of wins, while on BCDR-F01 ranged from 21% (LMT) to 32% (RF). When clinical data is present, wins are dominated by LMT on both datasets (46% on BCDR-F01, and 44% on BCDR-D01), as well as SVM in the the film dataset with 42% of wins.

4 Conclusions

The main contributions observed in this work are: (1) Histogram of Gradient Divergence (HGD), a descriptor of shape through the gradient of the image that is naturally invariant to rotation and that was recently proposed in [14] was the only image descriptor scoring best or comparable to best on both datasets; (2) Haralick features despite being a texture descriptor and not a descriptor related to shape, scored best on BCDR-D01 and second on BCDR-F01, suggesting that texture information may be important for evaluating masses; (3) clinical information enabled to significantly increase the performance of image descriptors in 92% of the cases; (4) the relative performance of the classifiers is similar for the two datasets, then it is possible to expect that image resolution is not critical; and (5) the Breast Cancer Digital Repository (BCDR) demonstrated to be a suitable reference for exploring machine learning classifiers and breast cancer CADx methods.

Future work will be aimed at increasing the BCDR with new annotated patients cases and exploring the combination/selection of features from different groups of image-based descriptors for improving the performance of Breast Cancer CADx methods.

Acknowledgements. This work was initially supported by the IMED (Development of Algorithms for Medical Image Analysis) 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. This work was also partially supported by the Cloud Thinking project (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, “Mais Centro” program.

References

1. Matheus, B.R., Schiabel, H.: Online Mammographic Images Database for Development and Comparison of CAD Schemes. Journal of Digital Imaging 24(3), 500–506 (2011)
2. Christoyianni, I., Dermatas, E., Kokkinakis, G.: Fast detection of masses in computer-aided mammography. IEEE Signal Processing Magazine 17(1), 54–64 (2000)
3. Marcano-Cedeño, A., Quintanilla-Domínguez, J., Andina, D.: WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications 38(8), 9573–9579 (2011)
4. Ramos-Pollan, R., Guevara-Lopez, M.A., Oliveira, E.: A software framework for building biomedical machine learning classifiers through grid computing resources. J. Med. Syst. 36(4), 2245–2257 (2012)

5. Antoniou, Z.C., et al.: A web-accessible mammographic image database dedicated to combined training and evaluation of radiologists and machines. In: 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009 (2009)
6. de Oliveira, J., et al.: MammoSVD: A content-based image retrieval system using a reference database of mammographies. In: 22nd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2009 (2009)
7. Oliveira, J.E.E., et al.: Toward a standard reference database for computer-aided mammography. In: Proceeding 6915 Medical Imaging 2008: Computer-Aided Diagnosis (2008)
8. Heath, M., et al.: Current status of the digital database for screening mammography. In: Digital Mammography, pp. 457–460. Kluwer Academic Publishers (1998)
9. Markey, M.K., et al.: Self-organizing map for cluster analysis of a breast cancer database. Artificial Intelligence in Medicine 27(2), 113–127 (2003)
10. Nishikawa, R.M.: Mammographic databases. Breast Dis. 10(3-4), 137–150 (1998)
11. Suckling, J.: The Mammographic Image Analysis Society Digital Mammogram Database in Exerpta Medica. International Congress Series, vol. 1069, pp. 375–378 (1994)
12. Moreira, I.C., et al.: INbreast: Toward a Full-field Digital Mammographic Database. Academic Radiology 19(2), 236–248 (2012)
13. Ramos-Pollan, R., et al.: Discovering mammography-based machine learning classifiers for breast cancer diagnosis. J. Med. Syst. 36(4), 2259–2269 (2012)
14. Moura, D., Guevara López, M.: An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. International Journal of Computer Assisted Radiology and Surgery, 1–14 (2013)
15. Hall, M., et al.: The WEKA data mining software: an update. SIGKDD Explorations 11(1), 10–18 (2009)