BCDR D01 Readme

Version: 1.0

Date: 13-06-2013

Contact email: bcdr@inegi.up.pt

About this file

This file describes the BCDR-D01 dataset from the Breast Cancer Digital Repository (bcdr.inegi.up.pt).

About the dataset

BCDR-D01 (Digital Mammography dataset number 1) is the first dataset of BCDR for full-field digital mammography and is composed by 79 biopsy-proven lesions of 64 women, rendering 143 segmentations (average of 1.81 images per lesion) including clinical data and image-based descriptors. All lesions are nodules or a combination of nodules with other abnormalities. The raw images (craniocaudal and mediolateral oblique mammograms) associated to BCDR-D01 dataset are also available together with the coordinates of the lesion's contours and numerical anonymous identifiers for linking instances and lesions. BCDR-D01 is a binary class dataset due to the initial BI-RADS classification of the radiologist being replaced by the result of the biopsy (Benign vs. Malign finding). Missing values are represented by the text NaN (not a number). The attributes of the dataset are described at the end of the document.

Citing the dataset

If you publish any work using this or other dataset from BCDR, please cite the following articles:

BCDR - Breast Cancer Digital Repository (Released for public domain at April 4, 2012). http://bcdr.inegi.up.pt

Daniel Cardoso Moura, Miguel Angel Guevara López, Pedro Cunha, Naimy González de Posada, Raúl Ramos Pollan, Isabel Ramos, Joana Pinheiro Loureiro, Inês C. Moreira, Bruno M. Ferreira de Araújo, Teresa Cardoso Fernandes., **"Benchmarking Datasets for Breast Cancer Computer-Aided Diagnosis (CADx)."** Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science Volume 8258, 2013, pp 326-333. http://dx.doi.org/10.1007/978-3-642-41822-8_41, Print ISBN: 978-3-642-41821-1, Online ISBN: 978-3-642-41822-8.

Daniel C. Moura and Miguel A. Guevara López, "**An evaluation of image descriptors combined with clinical data for breast cancer diagnosis**" International Journal of Computer Assisted Radiology and Surgery, vol. 8, pp. 561-574, 2013/07/01 2013.
doi: 10.1007/s11548-013-0838-2

Raúl Ramos-Pollán, Miguel Angel Guevara López, Cesar Suárez-Ortega, Guillermo Díaz-Herrero, Jose Miguel Franco-Valiente, Manuel Rubio-del-Solar, Naimy González-de-Posada, Mario Augusto Pires Vaz, Joana Loureiro and Isabel Ramos. "Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis". J. Medical Systems 36(4): 2259-2269 (2012) (http://dx.doi.org/10.1007/s10916-011-9693-2)

History

13-06-2013	v1.0	First release of the dataset

Clinical and general data

Feature	Description
Age	The age of the patient at the time of the study
Breast Density	The density of the breast at the time of the study according to the BI-RADS standard
Mammography Nodule	The lesion contains a mass
Mammography Calcification	Calcifications were detected in the lesion
Mammography Microcalcification	Microcalcifications were detected in the lesion
Mammography Axillary Adenopathy	Axillary adenopathy detected
Mammography Architectural Distortion	Signs of architectural distortion
Mammography Stroma Distortion	Signs of stroma distortion
Classification	Classification of lesion given by the biopsy result.
Image View	Type of image view (1-RCC, 2-LCC, 3-RO, 4-LO).
Mammography Type	Presence of abnormality.

Set of intensity descriptors computed directly from the grey-levels of the pixels inside the lesion's contour identified by the radiologists

Feature	Description
Mean (i_mean)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, with n being the number of pixels inside the region delimited by the contour and x_i being the grey level intensity of the i[th] pixel inside the contour.
Standard Deviation (i_std)	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
Skewness (i_skewness)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$
Kurtosis (i_kurtosis)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}-3$
Minimum (i_min)	The minimum intensity value in the region surrounded by the contour
Maximum (i_max)	The maximum intensity value in the region surrounded by the contour

Set of texture descriptors computed from the Grey-level co-occurrence matrix related to the bounding box of lesion's contour identified by the radiologists.

Feature	Description
Energy (t_energ)	$\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$ with L being the number of grey-levels, and p being the grey-level co-occurrence matrix and, thus, $p(i,j)$ is the probability of pixels with grey-level i occur together to pixels with grey-level j.
Contrast (t_contr)	$\sum_{i}\sum_{j}(i-j)^2\, p(i,j)$
Correlation (t_corr)	

	$\frac{\sum_i \sum_j (ij)\, p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ with μ_x, μ_y, σ_x and σ_y being the means and standard deviations of p_x and p_y, the partial probability density functions.
Sum of Squares: Variance (t_sosvh)	$\sum_i \sum_j (i-\mu)^2\, p(i,j)$ with μ being the mean of $p(i,j)$ for all i and *j*.
Homogeneity (t_homo)	$\sum_i \sum_j \frac{1}{1+(i-j)^2}\, p(i,j)$
Sum Average (t_savgh)	$\sum_{i=2}^{2L} i p_{x+y}(i)$ with $p_{x+y}(i)$ being the probability of the co-occurrence matrix coordinates summing $i = x + y$
Sum Entropy (t_senth)	$se = -\sum_{i=2}^{2L} p_{x+y}(i) \log\left(p_{x+y}(i)\right)$
Sum Variance (t_svarh)	$\sum_{i=2}^{2L} (i-se)^2\, p_{x+y}(i)$
Entropy (t_entro)	$-\sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j) \log\left(p(i,j)\right)$
Difference Variance (t_dvarh)	$\sum_{i=0}^{L-1} i^2\, p_{x-y}(i)$ with $p_{x-y}(i)$ being the probability of the co-occurrence matrix coordinates subtracting $i = \lvert x - y \rvert$
Difference Entropy (t_denth)	$-\sum_{i=0}^{L-1} p_{x-y}(i) \log\left(p_{x-y}(i)\right)$
Information Measure of Correlation 1 (t_inf1h)	$\frac{-\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log\left(p(i,j)\right) + \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log\left(p_x(i)\,p_y(j)\right)}{\max\left(\sum_{i=1}^{L} p_x(i)\log\left(p_x(i)\right), \sum_{i=1}^{L} p_y(i)\log\left(p_y(i)\right)\right)}$
Information Measure of Correlation 2 (t_inf2h)	$\sqrt{1-\exp\left(2\left(\sum_{i=1}^{L}\sum_{j=1}^{L} p_x(i)\,p_y(j)\log\left(p_x(i)\,p_y(j)\right) - \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log\left(p(i,j)\right)\right)\right)}$

Set of shape and location descriptors of the lesion's contour identified by the radiologists.

Feature	**Description**
Area (s_area)	$area = \lvert O \rvert$ with *O* being the set of pixels that belong to the segmented lesion
Perimeter (s_perimeter)	$perimeter = length(E)$ with $E \subset O$ being the edge pixels
Center of mass (s_x_center_mass, s_y_center_mass)	Normalized coordinates of the center of mass of *O*
Circularity (s_circularity)	$4\pi \frac{area}{perimeter^2}$

Elongation (s_elongation)	$elongation = \frac{m}{M}$ with *m* being the minor axis and *M* the major axis of the ellipse that has the same normalized second central moments as the region surrounded by the contour
Form (s_form)	$\frac{perimeter \times elongation}{8 \times area}$
Solidity (s_solidity)	$\frac{area}{\lvert H \rvert}$ with H being the set of pixels that belong to the convex hull of the segmented region
Extent (s_extent)	$\frac{area}{\lvert B \rvert}$ with B being the set of pixels that belong to the bounding box of the segmented region