Chaos theory-based quantification of ROIs for mammogram classification

Jaroslaw Kurek [(1)], Bartosz Świderski[(1)], Sami Dhahbi[(2)], Michal Kruk[(1)], Walid Barhoumi[(2)], Grzegorz Wieczorek[(1)] , Ezzeddine Zagrouba[(2)]

[(1)]*The Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences*

166 Nowoursynowska Street, 02-787 Warsaw, Poland *jaroslaw_kurek@sggw.pl*, *bartosz_swiderski@sggw.pl*, *michal_kruk@sggw.pl*, *grzegorz_wieczorek@sggw.pl*

[(2)] *Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA) – RIADI*

Laboratory, ISI, 2 Street Abou Rayhane Bayrouni, 2080 Ariana, Tunisia.

sami_dhahbi@yahoo.fr, *walid.barhoumi@esti.rnu.tn, ezzeddine.zagrouba@fsm.rnu.tn*

ABSTRACT: The paper presents improved method for breast cancer diagnosis. The previous method has been presented in Dhahbi et al. 2015. The suggested improvements regards applying more diagnostic features such as wavelet packet decomposition, Hilbert matrix, fractal texture features, etc. Moreover we investigated several classifiers such as Random Forest, Support Vector Machine and Decision Tree. In this paper larger database (number of images/trials) has been used and reached better accuracy.

Key words: mammography, diagnostic feature, breast cancer diagnosis, wavelet transform

1 INTRODUCTION

Breast cancer is one of the common women-cancer worldwide. It does not matter which region we take into consideration: the developed or less developed world, statistic states that over 508 000 women died in 2011 due to breast cancer [1] and 522 000 in 2012 [2].

Breast cancer is arising from cells of the breast that develops locally in the breast and metastasizes to the lymph nodes and internal organs (e.g. lungs, liver, bones and brain). Breast cancer is the most common malignancy in women [1]. This represents approximately 23% of all cases of cancers in women and approximately 14% of deaths because of this [2]. It is estimated that each year breast cancer is diagnosed in 1.5 million women worldwide, and about 522,000 die because of this [2]. It is the most common cancer among people of highly civilized countries, such as USA, Canada, Australia and Western European countries. Least of breast cancer recorded in southern Asia and Africa. Breast cancer, which is the most common cancer in women, in men is fortunately rare.

To cope with this issue, screening program is applied. Usually for women this screening program made up of palpable breast examination by a doctor (like self-examination performed regularly carried out by the woman herself) and mammography (ionizing radiation to create mammography images). Such research is usually done in women over 50 years of age every three years, although many experts believe that it is worth it carried out every two years. In the USA, women in first-degree relatives of individuals who develop breast cancer are advised to use mammography every 2 years starting at age test 10 years younger than that in which the person became ill relative.

Screening mammograms can detect lumps or other breast abnormality at an early stage when they are not detectable by a woman or doctor, which increases very significantly the chance of cure [3]. A significant negative effect of the research is a phenomenon observed that the negative result of mammography (i.e. not confirming the presence of the disease) in many women creates a false sense of security and reluctance and negligence in systematic self-examination of the breast. This procedure is wrong, because in about 20-25 % cases, mammography cannot detect a cancer developing (these are called false negative results).

The Mammography screening is currently the best tool for nationwide mammogram screening programs reducing the mortality rate of breast cancer [4, 5]. Hence the improved methods of the breast cancer are usually based on the analysis of mammography images.

2 DATABASES

Total number of mammography images used in this paper is 10168 and it is a part of Digital Database for Screening Mammography (DDSM University of South Florida) which is currently the largest public mammogram database [6]. The dataset includes

2604 cases. Screening mammography typically involves taking two views of the breast, from above (Cranial-Caudal view, CC) and from an oblique or angled view (Medio-Lateral-Oblique, MLO). Hence when we take into consideration left and right breast, we have 4 images for every case in the used challenging dataset.

2.1 *Region of Interest (ROI) Cropping*

This is an important step to extract and focus on appropriate part of mammography image. Since the whole mammographic image comprises the pectoral muscle and the background with a lot of noise, features were computed on a limited ROI that contains the prospective abnormality [2]. Thus, cropping a region of interest removes the unwanted parts. Image cropping was performed manually based on the physician annotation provided in the dataset. We can describe ROI as rectangular space where in its center we can observe the lesion (Figure 1).

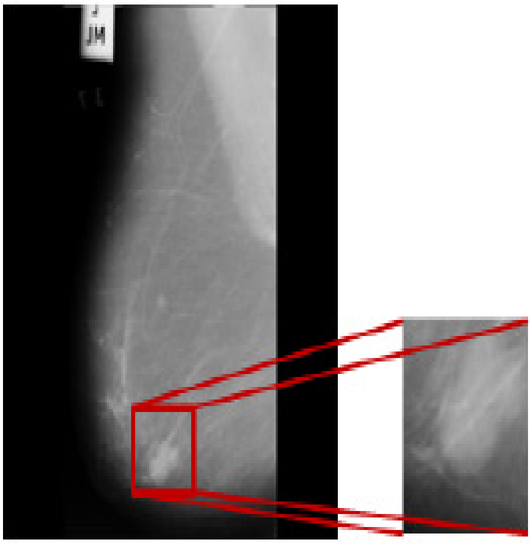

Figure 1 Example of region of interest (ROI) extraction.

2.2 *Trial subsets used in experiments*

Based on the above-described database, 10168 ROIs have been used in this work among the 10416 available ones [7]. Some number of all available ROIs has been rejected due to small quality of image. The dataset consists of three group of mammography result (Figure 2):

1. Normal tissue – 8254 ROIs
2. Benign lesion – 862 ROIs
3. Malignant lesion- 1052 ROIs

Examples of ROIs for every of three results are depicted on figure 2.

Due to the complexity of separating the three classes, we decided to merge ROI's belong to benign lesions and malignant ones. It means the issue is reduced to binary classification of normal tissue and abnormal tissue trials. Finally, we obtained:

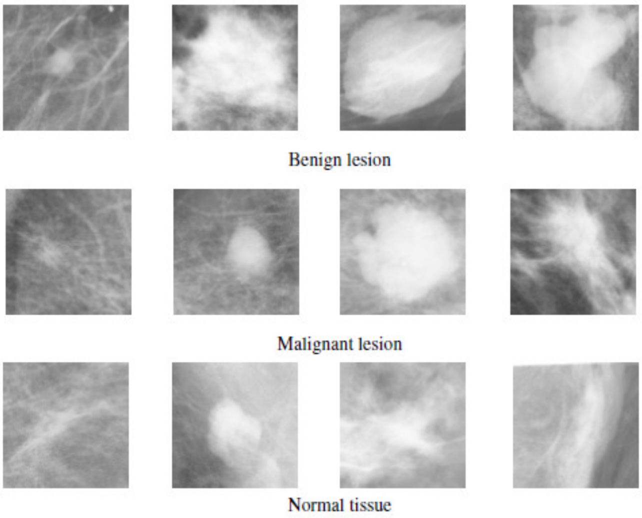

Figure 2 Sample of used ROIs from DDSM database.

1. Normal tissue – 8254 ROIs
2. Abnormal tissue – 1914 ROIs

3 FEATURES GENERATION

Based on the all extracted ROI's, we generated 122 potential diagnostic features for binary classification. Dataset of potential diagnostic features consist of the following group:

- features generated based on Hilbert's image representation (23 features).
- features generated based on segmentation-based fractal texture analysis (36 features).
- Kolmogorov-Smirnov descriptors (38 features).
- statistics from the gray-level co-occurrence matrix and other statistic (15 features).
- Maximum sub-region descriptors (10 features) using K-S statistic and Minkowsky approach.

3.1 *Hilbert's image representation*

Every image, which has regular representation in form of matrix of size n×m, has been transformed to one vector using our previously proposed Hilbert curve and recursive modeling. Building Hilbert curve of order 4 is depicted on Figure 3.

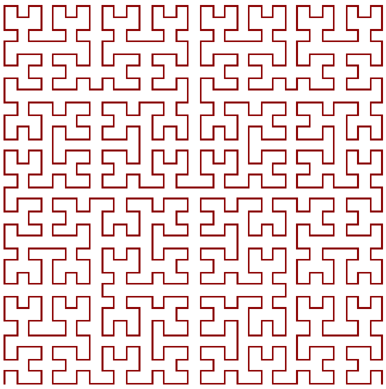

Figure 3 Example of Hilbert curve, which is a space filling curve that visits every pixel in a matrix (image).

3.2 *Trial subsets used in experiments.*

By means of Hilbert transformation of every image, we can obtain image vector representation for each image, which can be further treated as "time-series". Based on every image vector X representation, (where n denoted the vector length), we can calculate the following potential features:

- Mean:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Standard deviation:

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \mu)^2}$$

- Variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Skewness:

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\delta}\right)^3$$

- Kurtosis:

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\delta}\right)^4 - 3$$

- Root mean squared (RMS):

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}|x_i|}{n}}$$

- Crest factor (peak-to-rms ratio):

$$C = \frac{X_{peak}}{X_{RMS}}$$

- Wavelet energy of 4-level wavelet packet decomposition of db10 wavelet family:

After wavelet packet decomposition, the portion of energy for every terminal node has been calculated based on the following equation:

$$E = \sum_{k=1}^{N}\left|S_{jk}\right|^2$$

where, S_{jk} is appropriate coefficient discrete wavelet transformation.

3.3 *Segmentation-based fractal texture analysis*

This approach let us generate 36 texture features based on SFTA algorithm (Segmentation-based Fractal Texture Analysis) and returns 1×6 vector *D* extracted from the input grayscale image. When we apply 6th fractal order, we can obtain 36 texture features.

The SFTA algorithm decomposes images into various thresholded images using several sets of lower and upper threshold values. Implementation has been based on Costa approach [9]. Thresholded images are used to extract the fractal dimension. Since images with more jagged edges and prominent color differences tend to have higher fractal dimension, images with more uniform texture properties will have closer fractal dimension.

3.4 *Kolmogorov-Smirnov descriptors*

Kolmogorov-Smirnov descriptors reflect the change distribution of intensity of pixels placed in the rings of the increasing geometrical distances from the central point. The division of the images into coaxial rings is illustrated in Figure 6. It represents 4 concentric rings of equal number (56) pixels in each ring (equal number of pixels leads to more stable distribution of KS statistics) [10].

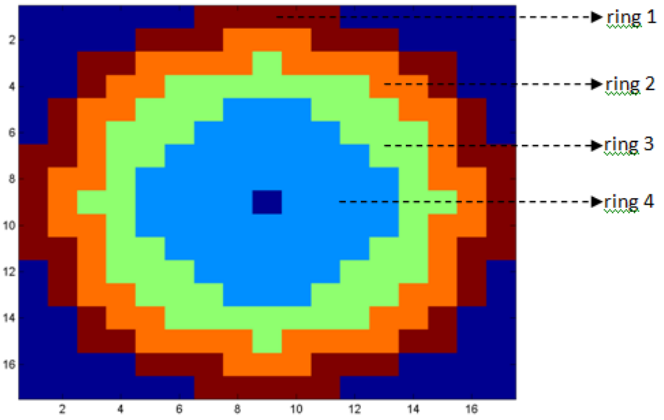

Figure 6 Illustration of 4 coaxial rings around the central pixel.

The central point is travelling along pixels uniformly distributed in the image. The results of the statistical analyses of these coaxial rings will be combined together by concatenating the pixel intensities corresponding to the same rings placed in equal distances, at different positions of the central pixel. Then cumulative Kolmogorov-Smirnov (KS) distances between the intensity of pixels x_i and x_j belonging to two different rings using KS test is then estimated. Based on above approach we can generate the following 7 features:

a) dKS_12 (mean KS statistics between ring no 1 and ring no 2).
b) dKS_13 (mean KS statistics between ring no 1 and ring no 3).
c) dKS_14 (mean KS statistics between ring no 1 and ring no 4).
d) the ratio dKS_13/ dKS_12.
e) the ratio dKS_14/ dKS_12.
f) the coefficient α0 of the approximation line $d_{KS} = \alpha_0 + \alpha_1 l + \varepsilon$.
g) the slope coefficient α1 of the approximation line $d_{KS} = \alpha_0 + \alpha_1 l + \varepsilon$.

Generally, we obtained 14 K-S features: 7 for mean and 7 for median approach. Rest of 24 similar features generated based on K-S approach has been described in [10].

3.5 *Statistics from the gray-level co-occurence matrix and other statistic*

Four features have been generated using gray-level co-occurrence matrix (GLCM) from a given image. GLCM calculates how often a pixel with gray-level (grayscale intensity) value *i* occurs horizontally adjacent to a pixel with the value *j*. Based on GLCM approach, the following statistic features were generated:

a) Contrast: measure of the intensity contrast between a pixel and its neighbour over the whole image.
b) Correlation: measure of how correlated a pixel is to its neighbour over the whole image.
c) Energy: sum of squared elements in the GLCM.
d) Homogeneity: value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Further 11 features are generated from general statistic measures of the treated image as one-dimensional vector. Hence, the following features were generated:

a) median
b) mean
c) standard deviation
d) kurtosis
e) min
f) max
g) high order cumulates and moments

3.6 *Maximum sub-region descriptors using K-S statistic and Minkowsy apporach*

The box-counting fractal dimension is measure characterizing the fractal complexity. It is the particular case of the Mandelbrot fractal dimension and is based on the notion of self-similarity of the structure at different scales. It measures how the length of the complex curve is changing when the measurement is performed with the increased accuracy [11].

To characterize any curve, this curve is covered with the set of regular squared areas of the size ε. Thus, we have to calculate the number of squared areas containing any part of the given curve. Number of founded areas is denoted as *N(ε)*.

$$d = \lim_{\epsilon \to \infty} \frac{\log(N(\epsilon))}{\log(\epsilon)}$$

In our approach, we have used 8 boxes. We suggested 122 various features based on which we can choose the best subset for binary classification.

4 FEATURES SELECTION

To choose appropriate diagnostic features, from all 122 potentially available, the best separating two class of tissue (normal and abnormal) sequential feature selection has been performed. This approach selects a subset of features from the data matrix *X* that best predicts the data in *y* by sequentially selecting features until there is no improvement in prediction [8]. There is possibility to apply chosen classifier (e.g. no-linear) for this selection method e.g SVM, random forest, decision tree. For every candidate feature subset, sequential feature selection performs 10-fold cross-validation by repeatedly calling function with different training subsets of *X* and *y*, XTRAIN and ytrain, and test subsets of X and y, XTEST and ytest. The result is logical vector indicating which features are finally chosen. After applying sequential features selection, we obtained 48 diagnostic features from the 122 available ones.

5 NUMERICAL EXPERIMENTS

The authors have applied three type of classifier: SVM, decision tree and random forest. For numerical experiments, the equal subset of normal and abnormal tissues trials have been chosen in randomize approach. The 10-fold stratified cross validation has been performed to compute the misclassification error for each classifier. Results are given in Table 1.

Table 1. Result of binary classification (normal vs. abnormal tissues) of mammography images.

Setting	Accuracy [%]	Standard deviation [%]
SVM	80%	5.1%
Decision tree	79 %	5.5%
Random forest	81%	4.5%

6 CONCLUSION

The obtained accuracy result (=81%) is more reliable than the one presented (=86.46%) in [2], because numerical tests are performed on 3828 total trials: 1914 trials of abnormal tissue and 1914 trials of normal tissue randomly chosen from 8254 available trials. In [2], accuracy is 86.46% but numerical experiments have been performed based on only 200 trials (ROIs): 100 trials normal and 100 trials abnormal. So, current method presented in this paper is more reliable based on the chosen selected features. The best result obtained for random forest classifier applying 200 trees in this algorithm.

7 REFERENCES

1. “Breast cancer: prevention and control”, WHO 2015.
2. S.Dhahbi, W. Barhoumia, E. Zagroubaa "Breast cancer diagnosis in digitized mammograms using curvelet moments", Comp. Biol. Med. 64 (2015) 79-90.
3. A. Jotwani, J. Gralow, Early detection of breast cancer, Mol. Diagnosis and Ther. 13(6) (2009) 349–357.
4. H. Nelson, K. Tyne, A. Naik, C. Bougatsos, B. Chan, P. Nygren, L. Humphrey, Screening for breast cancer: Systematic evidence review update for the U. S. preventive services task force, Ann. Intern. Med. 151(10) (2009) 727–W242.
5. S. Hofvind, G. Ursin, S. Tretli, S. Sebuodeg°ard, B. Moller, Breast cancer mortality in participants of the norwegian breast cancer screening program, Cancer 119 (2013) 3106–3112.
6. M. Heath, K. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, K. Chang, S. Munishkumaran, Current status of the digital database for screening mammography, in Digital Mammography, Springer Netherlands (1998) 457–460.
7. M. Jiang, S. Zhang, H. Li, N. Metaxas, Computer-aided diagnosis of mammographic masses using scalable image retrieval, IEEE Transactions on Biomedical Engineering 62(2) (2015) 783–792.
8. Matlab user manual, MathWorks, Natick, 2014
9. Costa, Alceu Ferraz, Humpire-Mamani, Gabriel, andTraina, Agma Juci Machado. An efficient algorithmfor fractal analysis of textures. 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images, 39–46, 2012.
10. Swiderski, S. Osowski, M. Kruk, J. Kurek, Texture characterization based on the Kolmogorov-Smirnov distance, Expert Systems with Applications 42(1) (2015) 503–509
11. Schroeder M., Fractals, Chaos, Power Laws, 6 Ed. New York: W.H. Freeman and Company, 2006.