International Journal of Emerging Trends in Health Sciences

Volume 2, Issue 1, (2018) 14-18

www.ijeths.eu

Segmentation and classification of breast cancer tumour

Gokalp Cinarer*, Bozok University, 66200 Yozgat, Turkey
Bulent Gursel Emiroglu, Kirikkale University, 71450 Kirikkale, Turkey
Ahmet Hasim Yurttakal, Bozok University, 66200 Yozgat, Turkey

Suggested Citation:
Cinarer, G., Emiroglu, B. G. & Yurttakal, A. H. (2018). Segmentation and classification of breast cancer tumour. *International Journal of Emerging Trends in Health Sciences. 2*(1), 014–018.

Received date October 22, 2017; revised date January 17, 2018; accepted date June 09, 2018.
Selection and peer review under responsibility of Prof. Dr. Nilgun Sarp, Uskudar University, Istanbul.

Abstract

Breast cancer is cancer that forms in the cells of the breasts. Breast cancer is the most common cancer diagnosed in women in the world. Breast cancer can occur in both men and women but it's far more common in women. Early detection of breast cancer tumours is crucial in the treatment. In this study, we presented computer-aided diagnosis (CAD) expectation-maximisation segmentation and co-occurrence texture features from wavelet approximation tumour image of each slice and evaluated the performance of SVM algorithm. We tested the model on 50 patients, among them, 25 are benign and 25 malign. The 80% of the images are allocated for training and 20% of images reserved for testing. The proposed model classified two patients correctly with a success rate of 80% in case of five-fold cross-validation.

Keywords: Breast cancer, computer-aided diagnosis (CAD), magnetic resonance imaging (MRI).

* ADDRESS FOR CORRESPONDENCE: **Gokalp Cinarer**, Bozok University, 66200 Yozgat, Turkey.
E-mail address: gokalp.cinarer@bozok.edu.tr / Tel.: +0506 533 5299

1. Introduction

The use of technology in the diagnosis and treatment of medical diseases is increasing day by day. One of the most used areas of this technology is computer-aided detection of breast cancer cells. X-ray, tomography, 3D imaging and ultrasound imaging techniques are used in breast cancer diagnosis together with developing technology. Among the deaths from cancer in the western world, after women's lung cancer, breast cancer comes second (Aydintug, 2004). Breast cancer is not only the most common malignancy in women throughout the world but also constitutes 29% of the estimated new cases of cancer in women, but it is also one of the major causes of death in all cancer types (26%) (Gupta & Jain, 1997).

To reduce unnecessary biopsies, recently, magnetic resonance imaging (MRI) is also used for the diagnosis of breast cancer (Wollins & Somerfield, 2008) since it has an excellent capability for soft tissue imaging. SVM is one of the algorithms used to diagnose breast cancer. We have used Support Vector Machines to diagnose these diseases. Some of the studies done in the literature using the SVM classification algorithm. Sampaio, Diniz, Silva, De Paiva and Gattass (2011) obtained 80% accuracy by extracting shape properties using the geostatic function. Wang, Shi and Heng (2009) obtained 91.4% accuracy using Gabor and GLCM features. Rejani and Selvi (2009) have obtained 88.75% accuracy using wavelet-based shape features. Moayedi, Azimifar, Boostani and Katebi (2010) have obtained 82.1% accuracy using contourlet-based features.

In this study, we propose an automated CAD to diagnose breast cancer tumour using MRI images. The proposed CAD consists of a series of procedures such that pre-processing, segmentation, feature extraction and classification.

2. Materials and methods

2.1. Pre-processing

Breast MRI images were subjected to a preprocessing step for noise removal and image segmentation. There are many ways to reduce noise. In this study, the median method was used to reduce noise. The median method is based on the adjacency of the pixels and median processing. With this method, the detection of tumour areas in the image is achieved without blurring the image and without losing the sharpness of the image. *n* median filtering, the size of the selected filtering patterns influences the output image, the use of large filtering patterns increases the amount of blurring in the image (Akar, 2006). Imaging Gaussian filtering is applied to make the image more smooth and selectable. Filtering performs operations such as sharpening on the image, extracting specific details, smoothing the image, edge sharpening or edge detection (Guvenc, 2008).

Two dimensional Gauss filter applied as

$$g(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

The σ (standard deviation) value for the variance is determined by the user and a fixed value is taken for the whole pixel in the image. Large selection of the variance value increases noise removal and at the same time causes blurring in the image and destruction of the edges. A small selection of the variance value reduces noise removal and increases the protection of edges and detail (Gonzales, 2002). Top line and bottom line transformations have been applied in order to obtain the brightness and contrast values at the top level in the study. Due to these features, top hat- and bottom hat-transformations help to separate the anomalies from other areas on the image (Gedik, 2013).

2.2. Segmentation

EM algorithm is used for segmentation of MRI images. The EM algorithm is an effective procedure for calculating the maximum likelihood estimate. Each iteration of the EM algorithm consists of two processes: expectation (E-step) and maximum probability (M-step).

In E-step, the missing data are estimated when the current estimate of observed data and model parameters is taken into consideration. At step M, under the assumption that missing information is known, the probability function is maximised.

2.3. Feature extraction

In this study, the attributes of MRI images were calculated using GLCM. GLCM is based on $P(i, j|d, \theta)$ estimate of the second-order composite state probability density function. The distance between these matrix pixels is *d* and the angle θ indicates the probability of passing from gray level *i* to gray level *j* (Demirhan & Guler, 2010).

The GLCM functions characterise the texture of an image by calculating how often pairs of the pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. The process is known as GLCM based texture analysis and gives information on the disposition of the structures and their relations with the environment (Long, Zhang & Feng, 2003; Pathak & Barooah, 2013).

Features	Description
Energy	It is the measure of the homogeneity of the image. As the homogeneity of the image increases, this value grows.
Entropy	The image gives a measure of complexity. Complex textures have high entropy.
Contrast	It is a measure of image contrast.
Homogeneity	A measure of similarity in different regions of the image.
Correlation	A measure of the linearity of the image. Linear structures in the direction of 'θ' lead to large correlation values in this direction.
Mean	The average gray level intensity within the ROI.
Standard deviation	The amount of variation or dispersion from the mean value

2.4. Classification

SVM is a useful method for building a classifier. It aims to create a decision boundary between two classes that enable the prediction of labels from one or more feature vectors (Noble, 2006). This decision boundary, known as the hyperplane, is orientated in such a way that it is as far as possible from the closest data points from each of the classes. These closest points are called support vectors (Huang et al., 2018).

Given a labelled training dataset:

$$(x_1, y_1),\ \ldots,\ (x_n, y_n), x_i \in R^d \text{and} y_i \in (-1, +1)$$

where x_i is a feature vector representation and y_i the class label (negative or positive) of a training compound *i*. The optimal hyperplane can then be defined as:

$$wx^T + b = 0$$

where *w* is the weight vector, *x* is the input feature vector and *b* is the bias. The *w* and *b* would satisfy the following inequalities for all elements of the training set:

$$wx_i^T + b \geq +1 \text{ } if \text{ } y_i = 1$$
$$wx_i^T + b \leq -1 \text{ } if \text{ } y_i = -1$$

The objective of training an SVM model is to find the *w* and *b* so that the hyperplane separates the data and maximises the margin $1/||w||^2$ (Huang et al., 2018).

3. Conclusion and discussion

Proposed CAD is implemented in the Python environment. Before the pre-processing procedure, the original image, intersection of MRI, at Figure 1a is cropped at the circled region at the top left to obtain the original ROI in Figure 1b). During the pre-processing procedure, the original ROI is applied to a median filter for reducing the noise, and then Gaussian Filter for smoothing the image. After that, top-hat and bottom-hat operations are performed to the resulting image. Later, the final image is segmented via clustering technique shown in Figure 1c.

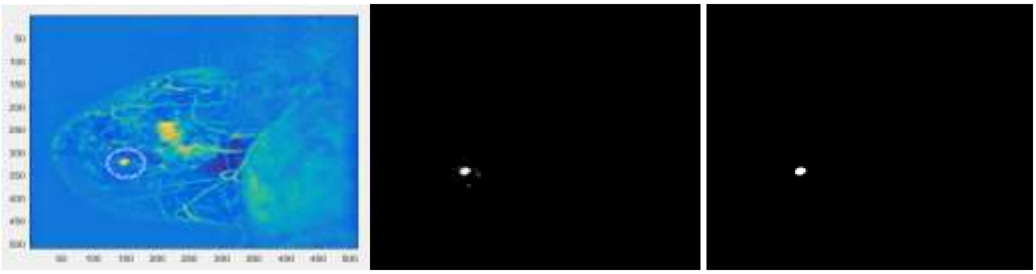

Figure 1. Original image (a) is cropped at a circled region at top-left to obtain ROI (b) and ROI is segmented and the result (c) is obtained

Table 1 presents statistics about the numerical features like mean, standard deviation, number of sample (count) or max–min values in the dataset.

Table 1. Statistics of features

	Contrast	Correlation	Energy	Homogeneity	Mean	Standard deviation	Entropy
Count	50	50	50	50	50	50	50
Mean	1.092645	0.084459	0.535788	0.842628	0.014111	0.156966	3.736366
Standard	0.297002	0.074666	0.079943	0.031449	0.007231	0.012799	0.353348
Min	0.587912	−0.057725	0.396633	0.785521	0.000912	0.133519	2.916530
Max	1.754464	0.253034	0.713214	0.907181	0.031628	0.177170	4.264678

SVM classification algorithm is fed with these features values to distinguish a benign and malign 50 tumour images. The 80% of the images were allocated for training and 20% of images reserved for testing. The Proposed CAD classified correctly with a success rate of 80% with no false-positive. The results are presented as confusion matrix in Table 2.

Table 2. SVM confusion matrix

		Predicted	
		Benign	**Malign**
Actual	Benign	4	2
	Malign	0	4

Some evaluating metrics such as Precision, Recall and F1 score for this fold are tabulated in Table 3. Throughout this paper, we presented a new user-independent time-saving CAD to diagnose breast cancer tumour using MRI images. The CAD's accuracy is 80.00% and has good accuracy among the existing ones in the literature with 50 MRI images.

Table 3. Classification performance

	Precision	Recall	F1-score	Support
Benign	1.00	0.67	0.80	6
Malign	0.67	1.00	0.80	4
Avg/total	0.87	0.80	0.80	10

It can be observed that these methods can be tested on different datasets to achieve the same success regardless of the dataset.

References

Akar, E. O. (2006). *Tarihi goruntulerin kalitesinin goruntu isleme teknikleri kullanilarak arttirilmasi*. Konya, Turkiye: Selcuk Universitesi Fen Bilimleri Enstitusu, Elektrik-Elektronik Muhendisligi A.B.D.

Aydintug, S. (2004). Meme kanserinde erken tani. *Sted, 13*(6), 226–229.

Baykara, M. (2009). Tibbi goruntulerde icerik tabanli goruntu erisim icin ozellik cikarim yontemleri.

Demirhan, A. & Guler, I. (2010). Ozorgutlemeli Harita Aglari Ve Gri Duzey Es Olusum Matrisleri Ile Goruntu Bolutleme. *Gazi Universitesi Muhendislik-Mimarlik Fakultesi Dergisi, 25*(2), 258–291.

Gedik, N. (2013). *Bilgisayar destekli teshis ile mamogram goruntulerinin siniflandirilmasi* (Ph.D. thesis). Karadeniz Technical University, Turkey.

Gonzales, R. C. (2002). *Woods, Digital Image Processing*. MA: Addison Wesley.

Gupta, A. & Jain, R. (1997). Visual information retrieval. *Communications of the ACM, 40*(5), 70–79.

Guvenc, U. (2008). *Uyarlanabilir Goruntu Filtre Tasarimi*. Ankara, Turkey: Gazi Universitesi Fen Bilimleri Enstitusu.

Haralick, R. M., Shanmugam, K. & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on System Man Sybernetics, SMC-3*(6), 610–621.

Huang, S., Cai, N., Pacheco, P. P., Narandes, S., Wang, Y. & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics, 15*(1), 41–51.

Long, F., Zhang, H. & Feng, D. D. (2003). Fundamentals of content based image retrieval.

Moayedi, F., Azimifar, Z., Boostani, R. & Katebi, S. (2010). Contourlet-based mammography mass classification using the SVM family. *Computers in Biology and Medicine, 40*(4), 373–383.

Noble, W. S. (2006). What is a support vector machine. *Nature Biotechnology, 24*(12), 1565–1557.

Pathak, B. & Barooah, D. (2013). Texture analysis based on the gray-level co-occurrence matrix considering possible orientations. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2*(9), 4206–4212.

Rejani, Y. & Selvi, S. T. (2009). Early detection of breast cancer using SVM classifier technique. *arXiv preprint arXiv:0912.2314*.

Sampaio, W. B., Diniz, E. M., Silva, A. C., De Paiva, A. C. & Gattass, M. (2011). Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Computers in Biology and Medicine, 41*(8), 653–664.

Siegel, R. L., Miller, K. D. & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians, 66*, 7–30.

Stricker, M. & Orengo, M. (1995, February). Similarity of color images. *SPIE Storage and Retrieval for Image and Video Databases III, 2185*, 381–392.

Texture analysis using the gray-level co-occurrence matrix (glcm). (2017). Retrieved from https://www.mathworks.com/help/images/gray-level-co-occurrence-matrix-glcm.html

Wollins, D. S. & Somerfield, M. R. (2008). Q and a: magnetic resonance imaging in the detection and evaluation of breast cancer. *Journal of Oncology Practice, 4*(1), 18–23.

Wang, D., Shi, L. & Heng, P. A. (2009). Automatic detection of breast cancers in mammograms using structured support vector machines. *Neurocomputing, 72*(13–15), 3296–3302.