Springer conference paper

Conference Paper · February 2021

CITATIONS
0

READS
25,984

1 author:

Kanyakumari Lakku
Andhra Loyola Institute of Engineering and Technology
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

A Novel Approach for Detection of Tumors in Mammographic Images Using Fourier Descriptors and KNN

L. Kanya Kumari[1(✉)] and B. N. Jagadesh[2]

[1] Department of Computer Science and Engineering, K L University, Guntur, Andhra Pradesh, India
kanyabtech@yahoo.com
[2] Department of Computer Science and Engineering, Srinivasa Institute of Engineering and Technology, Amalapuram, Andhra Pradesh, India
nagajagadesh@gmail.com

Abstract. The commonest disease in Indian women is breast cancer (BC). The Mammography Screening is the best solution to predict breast cancer in early stages. Though more research is going to predict BC in the early stages, the death rate is increasing year by year. So, we proposed a method for efficient classification to identify BC in the early stage. Our system uses efficient shape descriptors and is segmented by using Thresholding. Classification is done by using K-Nearest Neighbor (KNN). Our system classified the mammography images into Normal, Benign or Malignant.

Keywords: Breast cancer · Screening · Mammograms · Shape descriptors · Thresholding · K-Nearest neighbor

1 Introduction

Breast cancer is the second deadly diseases that occur in women, there is a chance for early detection of this cancer by the screening test. Screening is the best way to detect BC in the early stages as it is saving lives. Reasons behind getting BC are 1. BRCA1/BRCA2 gene 2. A strong family history 3. Radiation treatment. 4. A mutation in the TP53 or PTEN genes. The warning signs for BC are lumps, swelling, redness or darkening of the breast, change in the size of the breast and nipple discharge. Not all lumps lead to BC. Breast consists of small calcium deposits called micro calcifications which may lead to precancerous or early breast cancer [1].

According to IARC (International Agency for Research on Cancer)2,088,849 women are suffering from Breast Cancer(BC) [2]. The diagnosis of this cancer cell is a difficult task [3]. The screening of breast cancer can be done in many ways. They are digital mammography (like X-ray), ultrasound, CT-Scan, MRI and Tomosynthesis. One among them mostly using is digital mammography, and here the Physicians examine the mammographic images in order to find the possible occurrence of the breast cancer. These kinds of imaging techniques require different temperatures that correlate with various types of breast tissue pathology. But, the accurate prediction was

A. Kumar et al. (eds.), *ICDSMLA 2019*, Lecture Notes in Electrical Engineering 601,
https://doi.org/10.1007/978-981-15-1420-3_191

not possible because of human fatigue and habituation [4, 5]. Till now, there is no efficient method to control the occurrence of breast cancer.

In medical diagnosis and screening techniques, digital mammography is said to be the common technique that is employed in clinical practice, because of its low cost and accessibility [5]. There are four types of mammographic images. They are LMLO, RMLO, LCC, RCC (MLO: Medio Lateral Oblique Images, CC: Craniocaudal Images). In case of Mammography, a normal cell may look like a cancer cell and this is said to be false positive, this kind of misdiagnosis requires more tests and diagnostic procedures, so it might be more stressful for patients [6]. Further, the exposure to x-ray as an ionizing radiation might increase the risk of occurrence of breast cancer for pregnant women and women under 30 years old.

Moreover, Electronic Health Record (EHR) [7] is said to be an efficient collection of individual patients records populations including past records also. Health records might comprise a large range of data including general medical records, patient treatments, medical history, laboratory results and radiology images [8]. EHRs principally contains three types of data: (1) structured [9], which has particular format includes Demographic information, diagnosis, laboratory tests report, Patient insurance etc., (2) unstructured data [10], that includes images. (3) Semi-structured data [11], with rules or patterns collected from the large collection of data of a particular patient. The structured data is generally used for specific type of data that are easier to be identified and extracted quickly. But, each individual data element got from different sources does not provide information about the whole clinical context as it is got from the single source. The rich information gathered from different sources might help the radiologist in clear examination and diagnosis of the cancer cells using computerized techniques. Further, EHRs reduces the cost incurred for legacy systems, improves the quality of care, and mobility of records around the globe [12, 13].

2 Related Works

In [1] the dataset considered was MIAS (Mammogram Image Analysis Society) to classify mammogram images into benign or Malignant. This database consists of 322 sample images and is of 1024*1024 pixels. They have classified these images into 208 normal images, 63 in benign and 51 to malignant. Features are extracted from the images by using a discrete wavelet transform. They have classified the mammographic images by using Multi-Layer Perceptron and Radial Basis Function (RBF) techniques. They have concluded that RBF neural net gives better results than MLP.

In [14] used the Bayesian network (BN) modeling approach for filling the void and this BN structure was constructed using the K2 learning algorithm in a statistical computation way and then it assessed the acquired BN model in this research. The data used for this research was got from the clinical ultrasound dataset that was obtained from the Chinese local hospital and the UCI machine learning repository was got from the fine-needle aspiration cytology (FNAC). The research also formulated that, if the data obtained was in the form of ultrasound data, then the diagnosis was made with the help of cell shape as a feature for predicting the presence of cancer. Moreover, based on the blood signals a strong resistance index was suggested to guess the probability of

cancer. In the case of FNAC data, the bare nuclei were a differentiating factor of malignant and benign breast tumors and there was an independence contribution between the identical cell size and cell shape. The clinicians were able to make decisions on the prediction of cancer based on BN modeling approach and the features were evaluated based on certain important features that were identified by the model, in case of loosing of certain features in some unique features. The approach was valid to analyze the healthcare data and for modeling the diagnosis of the disease.

In 2015, the authors [15] generated the computer-aided diagnosis (CAD) system that was based on the features of the Breast Imaging Reporting and Data System (BI-RADS) for the breast ultrasound. The diagnosis and the testing of the computerized features that were related to ultrasound BI-RADs were done with the help of the database of 283 pathology-proven kind and malignant lesion. For different machine learning methods such as decision tree, artificial neural network, random forest and support vector machine, the "bottom-up" approach was followed to select the features for classification of performance. Further, in the database of 283 cases, the 10-fold cross-validation there was a larger area under the receiver operating characteristic (ROC) curve (AUC) was found as 0.84 having overall accuracy as 77.7% and it was obtained from the support vector machine.

In 2018, the research [16] was mainly formulated to predict the prognosis of breast cancer by using the Multimodal Deep Neural Network that was integrated along with the Multi-dimensional Data (MDNNMD). The architecture in this method was designed and the combination of multi-dimensional data was achieved in this paper. Moreover, when a comparison was carried out between the proposed method and the existing method having a single-dimensional data and also with the other existing approaches, it was found proposed prediction method worked well with higher improvement in performance.

In [17], The research aimed to find the accuracy of screening the breast cancer for the women, who were not in the general population screening program and this was done using the annual mammography with or without magnetic resonance imaging (MRI).Here, with only the women suffering from increased risk of BC, having no well-known gene mutation was used for this research, and six prospective screening trials were made based on An individual patient data (IPD) meta-analysis. In order to have accuracy in screening such as, with sensitivity, specificity and predictive values, a generalized linear mixed model was applied for the annual mammography having with or without MRI. In this research 2226 women (in median age: 41 years, having inter quartile range 35e47) and then with 7478 woman-years, having a BC rate of 12 (i.e., 95% confidence interval 9.3e14) in the 1000 woman-years. Moreover, the sensitivity of 55% (for standard error with a mean [SE] 7.0) and having a specificity of 94% (i.e., SE 1.3) was achieved in the Mammography screening. With MRI screening alone, a sensitivity of 89% (SE 4.6) and a specificity of 83% (SE 2.8) was obtained. Then, the sensitivity had increased up to 98% (SE 1.8, $P < 0.01$ when compared to mammography alone) and it too produced a low specificity of 79% and all this happened by introducing the MRI to mammography. Finally, it was concluded that with the women having a strong risk of BC, having no well-known gene mutation, they had high BC incidence both before and after age 50 and on the addition of MRI to mammography had increased the sensitivity and had decreased the specificity.

3 Dataset

We are using the MIAS Database (Mammogram Image Analysis Society) to classify mammogram images into Normal or Benign or Malignant. The original dataset is digitized to 50-micron pixel edge and it has been reduced to 200-micron pixel edge so that each image of size is 1024 pixels × 1024 pixels [18]. This database has 322 sample grey-scale images. We represented four sample images from MIAS dataset in Fig. 1.

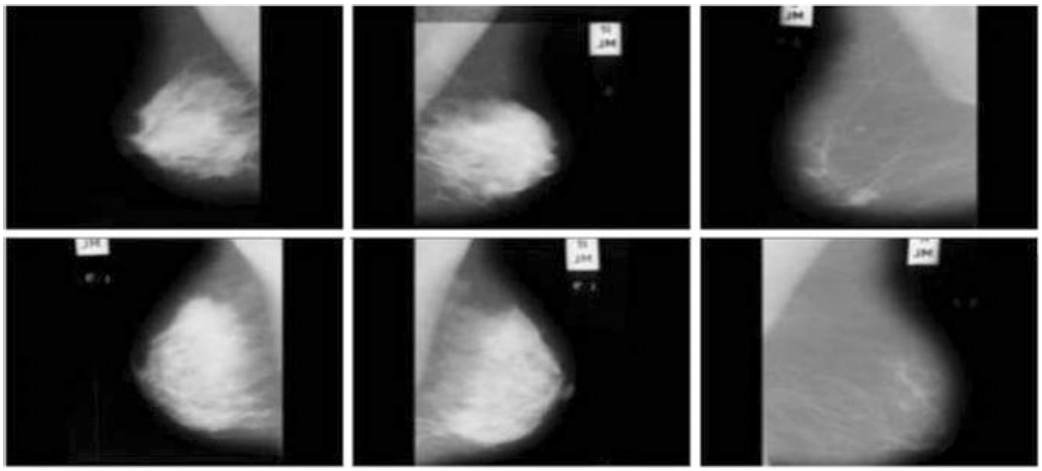


Fig. 1. Images from MIAS

4 Methodology

The details of the proposed methodology consist of various steps. They are

1. Shape descriptors for feature Extraction
2. K-Nearest Neighbor for classification.

4.1 Shape Descriptors Based Feature Extraction

In digital imaging, the characteristics are known by pixel, so a pixel is a small feature in an image. In medical imaging, each pixel value is playing an important role. There are number of approaches to obtain these pixel values. In this paper, we used shape descriptors to extract features from mammography images. Image descriptors give information about images which are homogeneous. In this, we are concentrating on shape-based descriptors. There is the number of techniques for standard descriptors viz., Moment Invariant Descriptor (MID), Curvature-Scale-Space-Descriptor (CSSD). Angular Radial Transform Descriptor (ARTD), Zernike Moment Descriptor (ZMD). Including these descriptors, many authors proposed a number of descriptors. Among all, used one technique MSC + GD (Modified Shape Context, Global descriptors) which was proposed in [19] is efficient than other techniques. The MSC uses centroid based angle calculation between any two points which is given by Eq. (1).

$$\theta(\mathrm{a},\mathrm{b}) = \tan^{-1}\left(\frac{y_2 - y_1}{1 + y_2 + y_1}\right) \tag{1}$$

where:

ɵ(a, b) = angle between a point (a, b)

y_2 = slope between the first and the second points

y_1 = slope between the first and the center points.

A Fourier transform is mostly used to apply transformations to recognize features in an image and Fourier coefficients are invariant to symmetry operations like scaling, translation and rotation. The important utility that optimizes Fourier transformation is the size of shape representation. Sampling is the main measure to generate a shape signature generation process. Among a number of sampling methods EAL (Equal Arc Length) gives better results [19]. The contour representation of an image is considered N-number of points. For a given contour signal 1-Dimensional Fourier transform is given by using fft() function in MAT lab. The 1-D FD (Fourier Descriptor) in Eq. (2).

$$\mathrm{FD_n} = \frac{1}{N}\sum_{x=0}^{N-1} s(x) \times e^{\left(\frac{-j2\pi nx}{N}\right)} \tag{2}$$

where,

$\mathrm{FD_n}$ = nth Fourier descriptor

$s(x)$ = 1−Dimensional contour signal

N = number of points of the contour

$$x = 0,\ 1,\ 2,\ \ldots \mathrm{N} - 1$$

By using Eq. (2) we calculated Fourier descriptors of size 'N' which are translation variant. This is translation variant because the radial distance is calculated based on centroid with respect to origin. The tumour cells are differentiated based on by using morphological operations [20] with structuring element se = 5. After morphological operations calcifications are visible as shown in the Fig. 2.

After identifying tumors statistical analysis is done. The mean and variance values are calculated by using Eqs. 3, 4 [21].

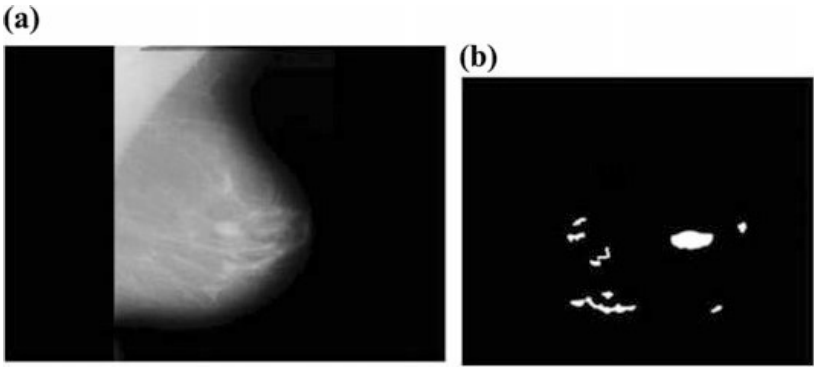


Fig. 2. **a** The image from MIAS025, **b** segmented image after morphological operations

$$\text{Mean}(a_n) = \frac{1}{m}\sum_{i=1}^{n} a_n \tag{3}$$

$$\text{Variance}(a_n) = \frac{1}{\text{m}-1}\sum_{n=1}^{m} (a_n - \text{mean}(a_n))^2 \tag{4}$$

4.2 K-Nearest Neighbor Classifier

K-Nearest Neighbor (KNN) is a method for classifying images which are Normal, Benign or Malignant. The input to KNN is set of feature vectors which are generated from shape descriptor called Fourier Transform. The proposed algorithm works as follows: 1. Load the Fourier descriptors to KNN 2. Find the differences among intensity values of the Fourier descriptor matrix. 3. The images which are having more differences are having the possibility of tumours. 4. Based on these differences we classified the images as Normal, Benign or Malignant. In our methodology of KNN the number of nearest neighbors used here is k = 3. The distances are calculated by using a formula Eq. (5).

$$\text{D}(\text{FD}_1, \text{FD}_2) = \frac{\sqrt{\sum_{i=1}^{m}(Xi - Yi) \times (Xi - Yi)}}{m} \tag{5}$$

where:

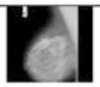
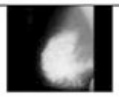
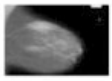
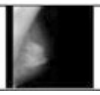
$\text{D}(\text{FD}_1, \text{FD}_2)$—Distance between the points

$X = \{ X_1, X_2, X_3 \ldots X_\text{n}\}$ and $Y = (\text{y}_1, \text{y}_2, \text{y}_3 \ldots \text{y}_\text{n})$

5 Results and Discussion

In this research, we suggested a method for detection of micro calcifications from mammograms which was explained above. The proposed methodology was applied in MIAS database which were grayscale images. The sample images were shown in Fig. 1. This dataset consists of 322 mammogram images of 161 patients. We applied the shape-based object recognition technique using Fourier descriptor to identify the tumors in mammogram images. The original size of images is 1024 × 1024. To increase the calculation speed, the original images are resized to re-size to 512 × 512. On these resized images the proposed methodology is applied to get shape descriptors and grayscale threshold considered was 0.7. The obtained features are translation invariant. The grayscale image is converted into binary image with threshold as mentioned above to convert the pixels greater than threshold to 1 and others as 0. Statistical measures were applied on Fourier descriptors. These are represented in Table 1 for some images taken from MIAS dataset. These values are given as input to the classifier called KNN. Here, we took k = 3 because the tumors may be Normal, Benign or Malignant. If we observe carefully the table, we observed that the measures are (in between 0 and 1) very low, low and high which represents Normal, Benign and Malignant. KNN classified images into Normal, Benign and Malignant.

Table 1. Statistical measures from FD

Image Number	Image	Tumors in the Image	Mean	Variance	Mass Type
Image019			0.1884	0.1529	Benign
Image 38			0.0773	0.0714	Normal
Image 105			0.3148	0.2686	Malignant
Image 150			0.125	0.1094	Benign
Image 214			0.0771	0.0712	Normal

6 Conclusion

In this paper, we concentrated on the most common disease in women is Breast cancer and presented the literature survey to predict BC. The classification techniques discussed are RBF, Naïve Bayes, ANN, SVM, Random Forest and different clustering algorithms. These are helpful to predict the cancer in early stages. But to analyze the data feature extraction is playing an important role. So, we used a novel approach to extract features based on shape descriptors. This will give more efficient and useful information which are hidden in the mammography images. Then we performed statistical measures like mean and variance. These measures are given as input to the KNN classifier. It classified the images into Normal, Benign and Malignant. Our proposed methodology is applied on MIAS database. So, from our methodology we can conclude that to detect tumors shape descriptors are more important. Finally we can say that the prediction of BC is still under development and we have to improve the accuracy of prediction rate.

References

1. Abirami C et al (2016) Performance analysis and detection of micro calcification in digital mammograms using wavelet features. In: *International conference on wireless communications, signal processing and networking (WiSPNET).* pp 2327–2331
2. https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf,26/11/2018
3. Rangayyan M, El-Faramawy NM, Desautels JEL, Alim OA (1997) Measures of acutance and shape for classification of breast tumours. IEEE Trans Med Imaging 16:799–810

4. Damiati S, Peacock M, Mhanna R, Sopstad S, Schuster B (2018) Bioinspired detection sensor based on functional nanostructures of S-proteins to target the folate receptors in breast cancer cells". Sen Actuat B Chem 267:224–230
5. Margolies LR, Salvatore M, Yip R, Tam K, Yankelevitz D (2018) The chest radiologist's role in invasive breast cancer detection. Clin Imaging 50:13–19
6. Rampun A, Morrow PJ, Scotney BW (2017) Fully automated breast boundary and pectoral muscle segmentation in mammograms. Artif Intell Med 79:28–41
7. del Carmen Legaz-García M, Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT (2016) A semantic web-based framework for the interoperability and exploitation of clinical models and EHR data. Knowl Based Syst 105:175–189
8. Forsyth AW, Barzilay R, Hughes KS, Lui D (2018) Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manage 55(6):1492–1499
9. Taggart J, Liaw S-T, Yu H (2015) Structured data quality reports to improve EHR data quality. Int J Med Informat 84(12):1094–1098
10. Vest JR, Grannis SJ, Haut DP, Halverson PK (2017) Using structured and unstructured data to identify patients' need for services that address the social determinants of health. Int J Med Informat 107:101–106
11. Amato F, De Pietro G, Esposito M, Mazzocca N (2015) An integrated framework for securing semi-structured health records. Knowl Based Syst 79:99–117
12. Varpio L, Rashotte J, Day K, King J (2015) The EHR and building the patient's story: a qualitative investigation of how EHR use obstructs a vital clinical activity". Int J Med Informat 84(12):1019–1028
13. Soguero-Ruiz C et al (2016) Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. IEEE J Biomed Health Inf 20(5):1404–1415
14. Liu S, Zeng J, Gong H, Yang H, Xuemei D (2018) Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach. Comput Biol Med 92:168–175
15. Shan J, Alam SK, Garra B, Zhang Y, Ahmed T (2016) Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. Ultrasound Med Biol 42(4):980–988
16. Sun D, Wang M, Li A (2018) A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. In: IEEE/ACM transactions on computational biology and bioinformatics
17. Phi X-A, Houssami N, Hooning MJ, Riedl CC (2017) Accuracy of screening women at familial risk of breast cancer without a known gene mutation: individual patient data meta-analysis. Eur J Cancer 85:31–38
18. Suckling J et al (2015) The mammographic image analysis society digital mammogram database excerpta medica. Int Congr Ser 1069:375–378
19. Madireddy RM et al (2014) A modified shape context method for shape based object retrieval. SpringerPlus 3:674
20. Sheba KU, Gladstone Raj S (2018) An approach for automatic lesion detection in mammograms. J Cog Eng 5
21. Shahin OR, Attiya G (2014) Classification of mammograms tumors using fourier analysis. Int J Comput Sci Netw Secur 110:14