VOL. 11, NO. 7, APRIL 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

BREAST CANCER DIAGNOSIS BASED ON FEATURE EXTRACTION BY HYBRID OF K-MEANS AND EXTREME LEARNING MACHINE ALGORITHMS

S. Chidambaranathan
Department of Computer Applications, St. Xavier's College (Autonomous), Tamilnadu, India
E-Mail: s.chidambaranathan14@gmail.com

ABSTRACT

Cancer is the most dreadful disease and breast cancer is the most commonly diagnosed disease. Automated disease diagnosis has gained substantial research interest these years. In this paper, a breast cancer detection algorithm that relies on different geometrical features of the image, k-means and Extreme Learning Algorithm (ELM) is proposed. The experimental results of the proposed algorithm are satisfactory in terms of detection accuracy and time complexity.

Keywords: breast cancer detection, ELM, K-means.

1. INTRODUCTION

Cancer is the most dreadful disease and breast cancer is the most commonly diagnosed disease. Automated disease diagnosis has gained substantial research interest these years. The advancement in information technology paves way for detecting cancer cells by means of feature comparison of normal and abnormal cells. Traditional method of cancer detection depends on mammograms, which are interpreted by radiologists and physicians.

In the year 1994, some radiologists interpreted 150 mammograms for analysing the effectiveness of tumour prediction. When the interpretations of all the radiologists are verified, it was found that the decisions of every radiologist are different though the mammogram is same [1]. Nowadays, many techniques are available for data collection and analysis. The data analysis techniques support the physician to come up with the effective diagnostic decision.

The detection accuracy of breast cancer detection can be boosted up with the incorporation of data mining techniques. Data mining techniques gain knowledge from the vast amount of data. The existing feature sets of breast cancer are differentiated into benign and malignant. The type of cancer can be predicted by means of the extracted features.

The appearance of breast cancer is not stable at their early stages. Hence, the physicians may not be able to locate the abnormalities. In such cases, this automated system helps the physicians to detect the abnormalities easily. A tumour detection algorithm has to identify the lesion and is needed to be accurate with reduced number of false negatives.

In this paper, an efficient cancer detection algorithm is proposed based on data mining techniques. While designing a cancer detection algorithm, accuracy and computational time are the key factors. In this work, the cancer detection is achieved by the hybrid algorithm of k-means and ELM.

The remainder of this paper is systematized as follows. Section 2 presents the review of literature. The proposed work is explained in section 3. The performance of the proposed algorithm is analysed in section 4. Finally, the concluding remarks are presented in section 5.

2. REVIEW OF LITERATURE

Mencattini A. *et al.* have proposed a new algorithm for mammographic image enhancement and denoising a mammographic image. Denoising is based on wavelet transform in this work [1]. Cao A.Z. *et al.* in [2] has investigated a Robust Information Clustering (RIC) algorithm, incorporating spatial information for breast mass detection in digitized mammograms. The detection system of this work employs RIC algorithm based on the raw Region of Interest (RoI) extracted from the global mammogram by two steps of adaptive thresholding.

The algorithm is claimed as robust in the sense that both the peak and valley of image intensity histogram are estimated and the pixels corresponding to valley in the histogram are clustered adaptively to the content of the image. Cascio D. *et al.* have proposed an algorithm for the detection of mass lesions in mammographic images. The algorithm follows the edge-based threshold operator strategy for segmenting the masses. The discriminating performance of the algorithm is verified by a supervised neural network [3].

In [5], Kom G. *et al.* have proposed an algorithm for detecting suspicious masses from mammographic images. In [6], te Brake G.M. *et al.* have proposed two different methods for segmenting the suspected regions.

Initially, the segmentations of masses were compared to the annotations made by the radiologist; secondly, a number of features were computed for all the segmented areas as normal and abnormal. Based on this, the regions are classified with the neural network.

Eltonsy N.H. *et al.* have proposed a technique in [8] for the automated detection of malignant masses in screening mammography. This technique is based on the presence of concentric layers surrounding a focal area with suspicious morphological characteristics and low relative incidence in the breast region. Mammographic areas with high concentration of concentric layers, with progressively lower average intensity are considered as suspicious deviations from normal parenchyma.

Singh S. *et al.* have proposed a novel set of metrics in [9], which measure the quality of the image enhancement of mammographic images in a computer-aided detection framework. This aimed at finding masses automatically using machine learning techniques. Saurabh Sharma et al. have presented a new algorithm for detecting suspicious lesions in mammograms by using adaptive thresholding [10].

Balakumaran T. *et al.* have come up with the algorithm to detect microcalcification in mammograms. This work proposes an algorithm that involves mammogram quality enhancement by using multiresolution analysis, which is based on the dyadic wavelet transform and microcalcification detection by fuzzy shell clustering [11]. In [12], Ted C. Wang et al. have proposed an algorithm that detects microcalcifications in digital mammograms. This is done by employing wavelet based sub-band image decomposition. The proposed method of this work is robust that it does not require the use of heuristics or the prior knowledge of the size and resolution of the mammogram.

Xhang X.P. have proposed a wavelet-packet multiscale image segmentation in [13]. This work is combined with a multiscale region based segmentation method and a new generic systematic scheme is generated. Using this scheme, suspicious tumour areas with exact boundaries are obtained on the basis of multiscale analysis in both grayscale and space. In [14], Xhang X.P *et al.* have developed an analytical model for the segmentation of targets. This is done by a novel multiresolution analysis in concert with a Bayes classifier, in order to identify the possible target areas. A method is developed in this work, which adaptively chooses the thresholds to segment targets from the background. This is done by a multiscale analysis of the image probability density function. Motivated by the above works, this paper proposes a hybrid algorithm of k-means and Extreme Learning Machine (ELM).

3. PROPOSED METHODOLOGY

The main features being considered by this work are area, circularity, correlation of pixel intensity, eccentricity and entropy of intensity distribution. These features are explained below.

3.1. Feature extraction

3.1.1. Area

A lesion can be defined as the total number of pixels within the affected area. Area is given by the total number of affected pixels.

3.1.2. Circularity

A rough representation of shape is provided by circularity and is given by

$$Circularity = \frac{Pm^2}{Ar} \quad (1)$$

Where, Pm^2 and Ar are the perimeter and area of the region respectively.

3.1.3. Correlation of pixel intensity

This concentrates on the correlation between the intensity of the pixels and is given by

$$cpi = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{M_c}\right)\left(\frac{y-\bar{y}}{M_n}\right) \quad (2)$$

Where, x and y are the index point of the centre pixel, $\bar{x}$ and $\bar{y}$ are the index co-ordinates of the neighbouring pixel. M_c is the mean of the pixel intensity with centre point value and M_n is the mean of the pixel intensity with neighbouring pixel value and n is the total number of pixels in the lesion detected image.

3.1.4. Eccentricity

Eccentricity determines the length of the affected boundary. A symmetric matrix namely 'symx' is defined. If ev_1 and ev_2 are the eigen values of symx matrix, then the values of semi-axis can be given by

$$E_1 = \sqrt{\left|\frac{ev_1}{2}\right|}, E_2 = \sqrt{\left|\frac{ev_2}{2}\right|} \quad (3)$$

and the eccentricity can be given by $ec = \frac{E_1}{E_2}$, if the value of ec is nearer to 1, then the affected area resembles circle, and if the value of ec is closer to 0, then the affected area may look like an ellipse.

3.1.5. Entropy of the intensity distribution

This feature focuses on the texture of the region, such as the roughness. This entropy can be computed by the below given formula.

$$Eid = -\sum_{m=1}^{4096} d_k \log(d_k) \quad (4)$$

Where d_k is the probability that the kth intensity lies in between I and I+δI

3.2. k-Means Algorithm

The extracted features are passed into the hybrid of k-means and ELM for accurate results. The k-means algorithm is responsible for clustering tumours based on the extracted features. It is computed by

$$min_{m_1,..m_k} \sum_{k=1}^{k} \sum_{i \in S_k} ||x^i - m_k||^2 \quad (5)$$

Where k is the cluster index, S_k is the kth cluster set, m_k is the central point in cluster S_k and k is the total number of clusters. The quality of the clusters is then measured by means of distance between the cluster member and the centroid of the cluster and the minimum distance between the clusters.

$$d_1 = \frac{\sum_{k=1}^{k} \sum_{i \in S_k} \sqrt{\sum_{j=1}^{F} (x_j^i - x_j^{m_k})^2}}{N} \quad (6)$$

$$d_2 = min \left[\sqrt{\sum_{j=1}^{F} (x_j^{m_{k1}} - x_j^{m_{k2}})^2} \right] \quad (7)$$

d_1 is the average distance between every member of the cluster with the centroid and d_2 is the minimum distance between two clusters.

Each cluster represents a specific tumour pattern. Each cluster centroid symbolizes the symbolic tumour of that cluster. After recognizing the malignant and benign tumour patterns, several symbolic tumours have been formed in both the malignant and benign data sets. The similarity between the untested tumour and the symbolic tumours plays an important role for diagnoses.

3.3. ELM

ELM [15]–[17] was originally proposed for the singlehidden-layer feedforward neural networks and was then extended to the generalized SLFNs where the hidden layer need not be neuron alike [18], [19]. In ELM, the hidden layer need not be tuned. The output function of ELM for generalized SLFNs (take one output node case as an example) is

$$f(x) = \sum_{i=1}^{L} a_i h_i(x) \quad (8)$$

Where a is the vector of output weights between the hidden layer of L nodes and the decision function of ELM is

$$f(x) = sign(\sum_{i=1}^{L} a_i h_i(x)) \quad (9)$$

ELM tends to reach not only the smallest training error but also the smallest norm of output weights.

Thus, the ELM effectively classifies between the normal and the abnormal cases with greater detection accuracy in lesser amount of time.

4. EXPERIMENTAL ANALYSIS

This work is tested over MIAS Mini Mammographic Database with 322 images and Matlab is employed for simulating this work [20]. The proposed work shows its excellence over k-SVM, ACO-SVM and PSO-SVM. The experimental results are shown in this section.

All these measures can be taken into account only if True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are in hand. The above mentioned parameters are computed in the following way.

4.1 True Positive (TP)

True Positive (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation (10).

$$TP = \frac{Number\ of\ Correctly\ classified\ images}{Total\ number\ of\ images} \times 100 \quad (10)$$

4.2 True Negative (TN)

True Negative (TN) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation 11.

$$TN = \frac{Number\ of\ falsely\ classified\ images}{Total\ number\ of\ images} \times 100 \quad (11)$$

4.3 False Positive (FP)

False Positive (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation.

$$FP = \frac{Number\ of\ correctly\ classified\ images}{Total\ number\ of\ images} \times 100 \quad (12)$$

4.4 False Negative (FN)

False Negative (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation.

$$FN = \frac{Number\ of\ falsely\ classified\ images}{Total\ number\ of\ images} \times 100 \quad (13)$$

With the above mentioned parameters, the accuracy, jaccord distance, sensitivity and specificity are calculated and the results are shown in graphs. From the experimental results, it is evident that the proposed system works well than the others.

4.5 Accuracy value

The accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value and the corresponding graph is presented in Figure-1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

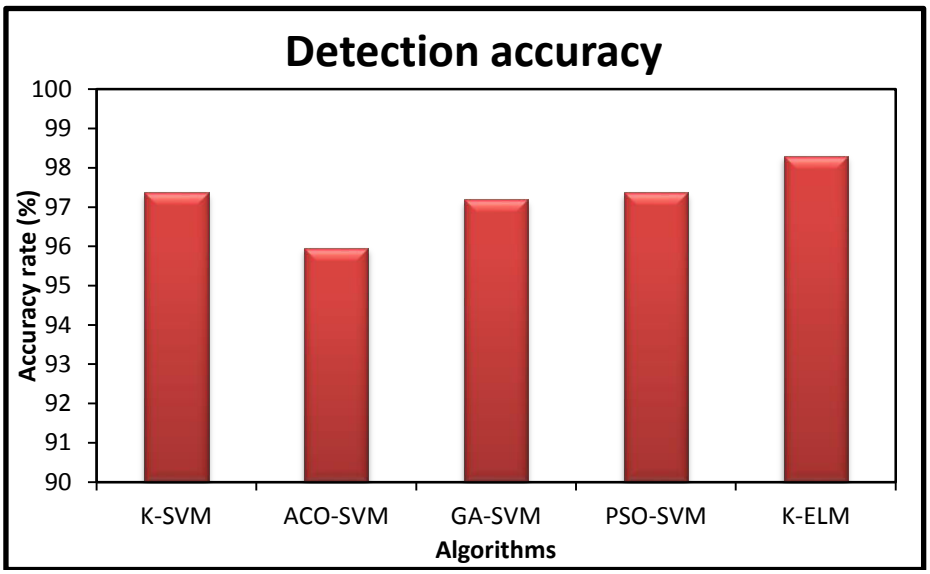

Figure-1. Detection accuracy analysis.

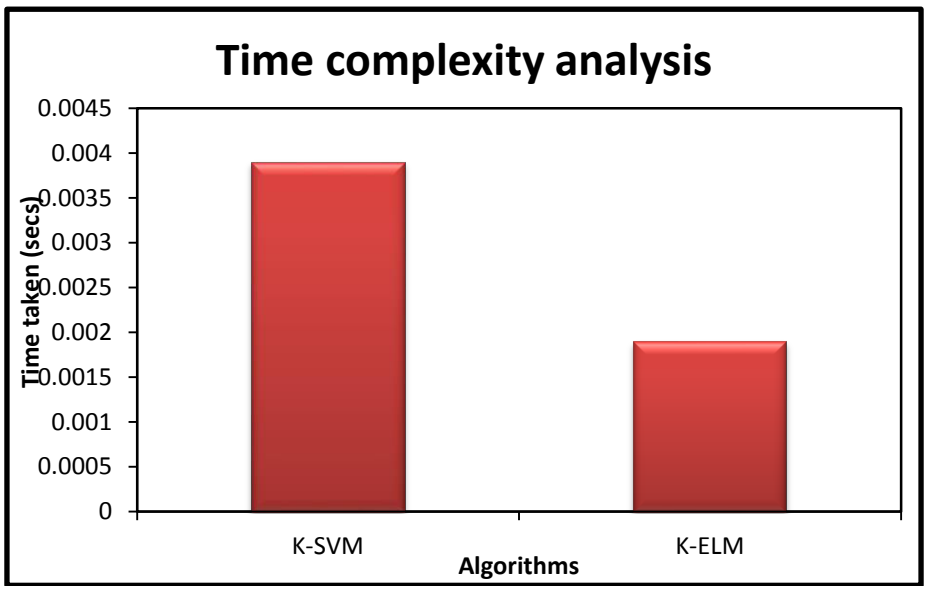

Figure-2. Time complexity analysis.

Thus, the proposed algorithm proves its performance by greater detection accuracy rate and the least time complexity. Thus, the proposed algorithm detects the breast cancer in a streak with good accuracy.

5. CONCLUSIONS

In this work, a system that automatically detects suspicious lesions is presented. This work hybrids the k-means and ELM algorithm. Initially, the features such as area, circularity, correlation of pixel intensity, eccentricity and entropy of intensity distribution are extracted. The extracted features are passed on to the hybrid of k-means and ELM. Finally, the image is classified with SVM as normal, benign or malignant in lesser period of time and with greater accuracy.

REFERENCES

[1] A. Mencattini, M. Salmeri, R. Lojacono, and F. Caselli. 2006. Mammographic Images Enhancement and Denoising for Microcalcification Detection Using Dyadic Wavelet Processing. IMTC 2006 - Instrumentation and Measurement Technology Conference Sorrento, Italy.

[2] A.Z. Cao, Q.Song, and X.L. Yang. 2008. Robust information clustering incorporating spatial information for breast mass detection in digitized mammograms. Comput. Vis. Image Understand. 109(1): 86-96.

[3] D.Cascio, F. Fauci, R. Magro, G. Raso, R. Bellotti, F. De Carlo, S. Tangaro, G. De Nunzio, M. Quarta, G. Forni, A. Lauria, M. E. Fantacci, A. Retico, G. L. Masala, P. Oliva, S. Bagnasco, S. C. Cheran and E. Lopez Torres. 2006. Mammogram Segmentation by Contour Searching and Mass Lesions Classification with Neural Network. IEEE Transactions on Nuclear Science. 53(5).

[4] D.H.Xu, A.Kurani, Furs t J.D. and D.S. Raicu. 2004. Run-length encoding for volumetric texture. The 4th IASTED International Conference on Visualization, Imaging, and Image Processing - VIIP 2004, Marbella, Spain.

[5] G. Kom, A. Tiedeu and M. Kom. 2007. Automated detection of masses in mammograms by local adaptive thresholding. Comput. Biol. Med. 37(1): 37-48.

[6] G.M. te Brake and N. Karssemeijer. 2001. Segmentation of suspicious densities in digital mammograms. Med. Phys. 28(2): 259-266.

[7] Mariam Biltwani, Nijad Al-Najdawi, Sara Tedmori. 2012. Mammogram enhancement and segmentation methods: Classification, Analysis and Evaluation. The 13th International Arab Conference on Information Technology. ACIT'2012 December 10-13, ISSN: 1812-0857.

[8] N.H. Eltonsy, G.D. Tourassi and A.S. Elmaghraby. 2007. A concentric morphology model for the detection of masses in mammography. IEEE Trans. Med. Imag. 26(6): 880-889.

[9] S. Singh and K.Bovis. 2005. An evaluation of contrast enhancement techniques for mammographic breast masses. IEEE Trans. Inf. Technol. Biomed. 9(1): 109-119.

[10] Saurabh Sharma, Ashish Oberoi. A new approach for Classification and Detection of Suspicious Lesions in Mammograms based on Adaptive Thresholding. International Journal of Computer Science and its Applications, ISSN 2250-3765, pp. 336-340.

[11] T. Balakumaran, Dr. ILA. Vennila and C. Gowri Shankar. 2010. Detection of Microcalcification in Mammograms Using Wavelet Transform and Fuzzy Shell Clustering. (IJCSIS) International Journal of Computer Science and Information Security. 7(1).

[12] Ted C. Wang and Nicolaos B. Karayiannis. 1998. Detection of Microcalcifications in Digital Mammograms Using Wavelets. IEEE Transactions on Medical Imaging. 17(4).

[13] X.P.Zhang. 2002. Multiscale tumor detection and segmentation in mammograms. In: Proc. IEEE Int. Symp. Biomed. Imag. pp. 213-216.

[14] X.P.Zhang and M.D.Desai. 2001. Segmentation of bright targets using wavelets and adaptive thresholding. IEEE Trans. Image Process. 10(7): 1020-1030.

[15] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew. 2004. Extreme learning machine: A new learning scheme of feedforward neural networks. In: Proc. IJCNN, Budapest, Hungary. 2: 985-990.

[16] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew. 2006. Extreme learning machine: Theory and applications. Neurocomputing. 70(1-3): 489-501.

[17] G.-B. Huang, L. Chen and C.-K. Siew. 2006. Universal approximation using incremental

constructive feedforward networks with random hidden nodes. IEEE Trans. Neural Netw. 17(4): 879-892.

[18] G.-B. Huang and L. Chen. 2007. Convex incremental extreme learning machine. Neurocomputing. 70(16-18): 3056-3062.

[19] G.-B. Huang and L. Chen. 2008. Enhanced random search based incremental extreme learning machine. Neurocomputing. 71(16-18): 3460-3468.

[20] http://www.breastcancerindia.net/.