KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 6 (2020) 320–324

www.elsevier.com/locate/icte

Breast cancer detection by leveraging Machine Learning

Anji Reddy Vaka[a], Badal Soni[a], Sudheer Reddy K.[b,*]

[a] *Department of CSE, National Institute of Technology, Silchar, India*
[b] *Researcher, Hyderabad, India*

Received 26 February 2020; received in revised form 5 April 2020; accepted 22 April 2020
Available online 7 May 2020

Abstract

India has witnessed 30% of the cases of breast cancer during the last few years and it is likely to increase. Breast cancer in India accounts that one woman is diagnosed every two minutes and every nine minutes, one woman dies. Early detection and diagnosis can save the lives of cancer patients. This paper presents a novel method to detect breast cancer by employing techniques of Machine Learning. The authors carried out an experimental analysis on a dataset to evaluate the performance. The proposed method has produced highly accurate and efficient results when compared to the existing methods.

Keywords: Machine Learning; Classification; Breast cancer; Deep learning

1. Introduction

Breast cancer (BC) is the malignant tumor that activates in the cells of the breast. A tumor has the potential to spread to other parts of the body [1,2]. BC is a universal disease that hammers the lives of women typically in the age group of 25–50. With the potential rise in the number of BC cases in India, the distress reaching is alarming. During the past five years, the survival rates of BC patients are about 90% in the USA and whereas in India the figure reports approximately 60% [3]. BC projection for India during 2020 suggests the number to go as high as two millions [4].

Specialist Doctors have identified hormonal, way of life and environmental factors that may increase an individual's odds of developing BC. Over 5%–6% of BC patients have linked to gene mutations that went through the ages of the family. Obesity, increasing age, postmenopausal hormonal imbalances are the other factors that cause BC.

As such, there is no prevention mechanism for BC, but early detection can significantly improve the outcome. Further, this can also considerably reduce the costs of the treatment. However, sometimes it is unusual to show cancer symptoms, so early detection is difficult. It is indispensable to employ mammograms and self-breast tests to detect any early irregularities before the tumor gets advanced [5].

The key objective of this paper is to propose a novice method to detect BC. This paper presents a detailed study of existing cancer detection models and presents the highly accurate and efficient results.

This paper is organized to have four sections. The literature and existing works are presented in Section 2. In Section 3, the proposed methodology has elaborated. The results and discussions are presented in Section 4. The results presented are proved to be accurate and efficient when compared to other models.

2. Literature review

This section presents literature survey. Relevant literature from multiple sources is referred for analysis of breast cancer detection. Further, authors reviewed various Datasets from regional and national cancer registries.

The authors have taken most popular BC detection methods namely; Naïve Bayes Classifier, Support Vector Machine (SVM) Classifier, Bi-clustering and Ada boost Techniques, R-CNN (Convolutional Neural Networks) Classifier, Bidirectional Recurrent Neural Networks (HA-BiRNN) [6–9]. These methods are described in this section.

SVM Classifier technique [6] is an amalgamation of RFE and SVM. RFE is a technique that operates by choosing

* Corresponding author.
E-mail address: sudheercse@gmail.com (Sudheer Reddy K.).
Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2020.04.009

Table 1
Performance analysis of most popular BC detection methods.

Methodology	Accuracy	Precision	Recall
Naïve Bayes classifier	95.61	95.65	93.61
SVM classifier	95.61	95.65	93.61
Bi-clustering and Ada boost techniques	95.75	95.72	96.26
RCNN classifier	91.3	91.3	89.3
Bidirectional Recurrent Neural Networks (HA-BiRNN)	82.50	80.09	79.03

dataset features depending on the least feature value in a recursive manner. Accordingly, SVM-RFE is operated by removing the inappropriate features (lowest weight feature) in all iterations.

AdaBoost is a most renowned ensemble technique and it is proficient of enhancing the accurateness of classification by combining several weak classifiers. The bi-cluster oriented classifiers can also be integrated with a strong ensemble classifier for superior generalization performance. During training, diverse weights are allocated and decisions are made depending on "weighted majority voting".

RNNs are the group of Neural Network (NN) that are deep in sequential dimension and were exploited widely in time-sequence modeling. In contrast to a traditional NN, RNNs are capable of processing the data points where the activation at every step is based on prior step.

CNN exploits the spatial data [3] amongst the image pixels and therefore, they depend on "discrete convolution". Accordingly, a gray scale image is presumed.

HA-BiRNN [9] comprises of two layers of encoder that are exploited for sentence encoder and word encoder, respectively. Along with this, sentence-level attention and word-level attention are also considered. Comparison of various BC detection methods is given in Table 1. The resultant values of accuracy, precision, and recall are listed.

2.1. Limitations of the existing methods

Naïve Bayes Classifier produces ill results when the training data is not represented [10]. The SVM classifier is unsuitable for large datasets and also not effective on high computer vision applications. When the data is imbalanced, Bi-clustering and Ada boost Techniques will lead to erroneous classification. RCNN takes more time to train the network. HA-BiRNN may produce wrong scores for BUS images [9]. With these limitations, the proposed methodology is introduced.

3. Methodology

However, the existing methods listed in Table 1 produce limited quality images and further have potential performance issues. Therefore a new method — Deep Neural Network with Support Value (DNNS) is introduced to produce better quality images and to fix other performance parameters. The authors propose a new algorithm or pseudocode along with mathematical formulas to evaluate the efficiency and performance.

Table 2
Summary of the dataset.

Magnification	Benign	Malignant	Training data	Testing data
40X	341	918	1471	402
100X	847	1149	1432	344
200X	811	1864	1549	473
400X	581	1498	1999	339
Total	2580	5429	6451	1558

3.1. Dataset

A well-annotated dataset is an essential requirement to produce a novel and robust method for the detection of BC. The collection of a dataset is highly difficult due to the unavailability of samples and confidentiality of the patients' demographic information. In this case, the well-annotated and large scale Dataset is taken from the M. G Cancer Hospital & Research Institute, Visakhapatnam, India. The dataset comprises 8009 histopathology image samples of over 683 patients of various magnification levels. The summary of the dataset is presented in Table 2. The given dataset has a set of histopathology images. These images are classified as benign and malignant tumors. The pre-processed images of the benign and malignant are applied to the proposed methodology to achieve the effective classification of BC cases. Further, the proposed method has produced highly accurate and efficient results when compared to the existing methods.

Data augmentation is employed to enlarge the dataset to lessen the problem of limited data size. Natural images are analyzed in a bottom-up approach. The majority of the Medical images are solved by a top-down approach. It is perceived that the augmentation techniques applied to natural images will not hold good for medical images. Due to this, the selection of the data augmentation approach on the dataset is complex. As the histopathology images have rotation; hence a rotation technique has been employed on both training and testing data. The rotation is carried out with 90°, 180°, and 270°.

Magnification alters the size of histology images that can enhance the quality for processing. As the histology images have numerous tissues, the analysis is bit complex at low magnification. It befits challenging for a system to learn distinct features from the said images with varied levels of magnification to make a differential diagnosis [5]. In the proposed system implementation, numerous phases of training are taken in aggregation with the preceding knowledge of the magnification factors as shown in Table 2.

The key objective of this proposed DNNS technique is to improve the efficiency and enhanced quality of images for

better prediction and diagnosis. In the proposed DNNS method the mathematical equations for calculation of histogram value, sigmoid function, and histo-sigmoid function have been revised and updated. The remaining equations are similar to standard BC detection algorithms [9].

The proposed DNNS methodology goes into three phases. In the Pre-processing phase, the input cytology images are pre-processed for the noise removal. This process has been done by using an effective filtering technique. In the second phase, the entropy, geometrical and textural features are extracted from the pre-processed images. The third phase segment the breast tumor from the extracted images. This was done by employing the Histo-sigmoid based fuzzy clustering.

3.2. Pre-processing

Pre-processing is the foremost activity in processing the images. Consider, ΔD is the breast cytology image database,

$$\Delta \mathrm{D} \in \{\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3, \ldots \mathrm{D}_n\}$$

where n is the number of images and D is the vector function.

Initially the input cytology images are pre-processed using an effective Gaussian filtering technique for the noise removal. Channels are utilized to expand splendor and complexity just as to change it up of surfaces, tones and embellishments to an image. Gaussian channel is characterized by,

$$G_D(x, y) = \frac{1}{2\pi {\delta_D}^2} \exp \frac{\left(-x^2 - y^2\right)}{2{\delta_D}^2} \tag{1}$$

where (x, y) is the current pixel of the image.

3.3. Feature extraction

The entropy, geometrical and textural features are extracted from the pre-processed images. Entropy (E) is the estimation of haphazardness that is utilized to describe the texture of the input image [11]. Shape features play an important role to distinguish the characteristics between normal and malignant cells. In textual features, each picture is partitioned into 'n' sub squares and quantized.

3.4. Histo-sigmoid fuzzy clustering

Histogram is an accurate representation of the distribution of numerical data and Fuzzy clustering is a method that permits one bit of information into extra two clusters.

The Histo-Sigmoid Fuzzy Clustering can be further elaborated in the following steps and experimented by using the various mathematical formulas.

Step-1: The histogram value of the image can be calculated by using Eq. (2).

$$H = \sum_{I=1}^{K} H_i \quad \text{Where H is the histogram function} \tag{2}$$

Step-2: The Sigmoid function can be calculated by using Eq. (3).

$$\delta = \frac{1}{1 + e^t} \tag{3}$$

Step-3: Substitute the histo-sigmoid values into the fuzzy clustering algorithm.

The fuzzy clustering contains the finite collection of elements, x={x_1, x_2, … x_n} and the cluster portioning of d = {d_1, d_2,… d_c}.

$\mathbf{U} = u_{ij} \in [0, 1]$, where I = (1, 2, … n) and j = (1, 2, … c)

$$F_C = \sum_{i=1}^{N} \sum_{j=1}^{D} u_{ij}^{C} \left\| x_i - d_j \right\|^2, 1 \leq c \angle \infty \tag{4}$$

where,

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - d_j\|}{\|x_i - d_k\|} \right)^{\frac{2}{m-1}}} \tag{5}$$

Step-4: Compute the histo-sigmoid function by applying the fuzzy clustering algorithm in Eq. (6).

$$u_{ij} = H + \frac{1}{\sum_{k=1}^{c} \left(\delta \frac{\|x_i - d_j\|}{\|x_i - d_k\|} \right)^{\frac{2}{m-1}}} \tag{6}$$

Input: Histo-sigmoid fuzzy clustered images
Output: Breast cancer analysis

1. **Begin:**
2. Initialize all weights and biases.
3. Set acquisition of knowledge rate $\Omega \in [0-1]$
4. For each input pattern E_k do
5. Compute difference of Output from actual output
6. For $i = 1$ to K do //with K number of feature modules
7. For layers =1 to $L'-1$ do
8. Compute the error based on the i-1 module i
9. End for
10. End for
11. Support value based normalization in equation (7)
12. After that Update weights of the Output
13. For $i = k$ to 1 do
14. For $i = 1$ to L' do
15. If module i is not a max-pooling layer then
16. Update weights and biases of the module i
17. End if
18. Update the thresholds of the module i for segments
19. End for
20. End for
21. **End**

Pseudocode of the proposed DNNS is presented below. The key functionality of the proposed pseudo code is to enable the support value to improve the range of the input images. The support value based normalization is calculated by using the equation in (7).

$$SN = \sup portvalue \times \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{7}$$

where, $Y_{\min}$ and $Y_{\max}$ are the minimum and maximum values in image Y, where SN is the support value based normalized image.

4. Results and discussions

The experimental results consummate for the proposed DNNS methodology and the comparison of classification with

Table 3
Performance measure.

Performance measure	Accuracy			
Segmentation	Image 1	Image 2	Image 3	Image 4
Proposed **(DNNS)**	0.967	0.957	0.955	0.929
FCM	0.893	0.923	0.944	0.843
Threshold	0.83	0.788	0.775	0.834

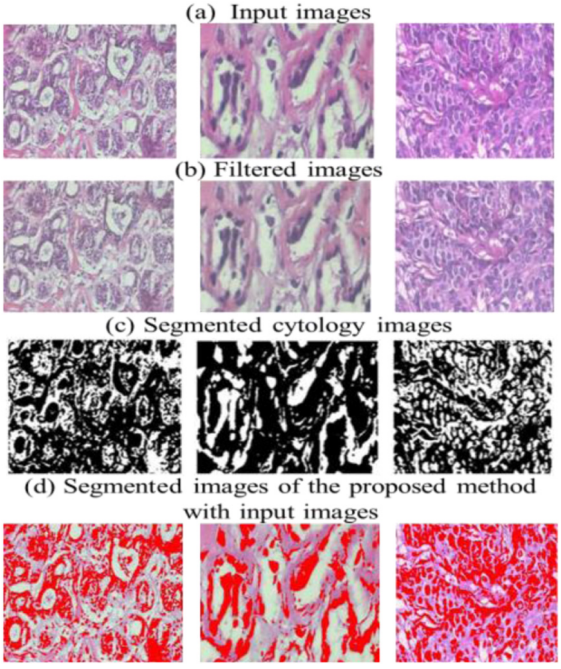

Fig. 1. Experimental results.

the existing SVM, Naive Bayesian, and Random Forest is analyzed in this section. The proposed DNNS classification is implemented in the operational stand of MATLAB. The experimental results are presented in Fig. 1. Fig. 1(a) shows the sample input images taken for segmentation process. Fig. 1(b) depicts the filtered images. Fig. 1(c) portrays the segmented cytology images. Fig. 1(d) illustrates the segmented image of the proposed method.

Performance measures such as Accuracy, Sensitivity, Precision, Recall, F-measure, Rank sum are calculated in the segmentation process and compared with the existing fuzzy C-Means (FCM) and Threshold. Comparison results of the proposed methodology with the existing FCM and Threshold are given in Table 3.

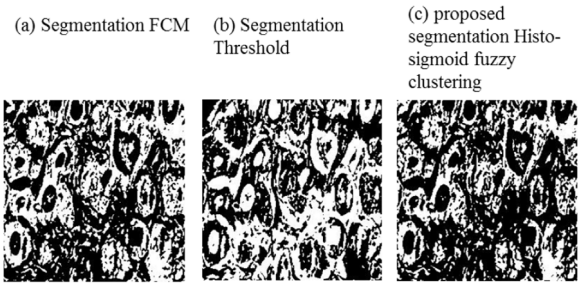

Fig. 2. Comparison of results (proposed method in (c) with the existing methods (a) and (b)).

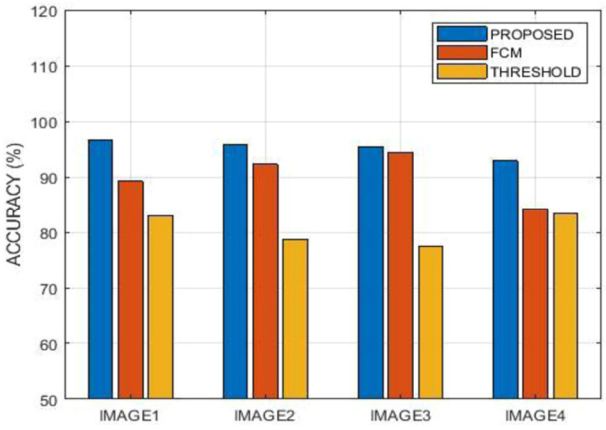

Fig. 3. Comparative results.

The performance measures are presented graphically in Fig. 2.

Comparison of performance of the DNNS method with the existing methods is presented Table 4.

Fig. 3 presents the comparative results of the proposed and existing.

5. Conclusions

The authors present the new method DNNS for detecting Breast Cancer. Unlike other methods, the proposed method is based on Support value on a deep neural network. To meet the better performance, efficiency, and quality of images, a normalization process has been employed. Experimental results proved that the proposed DNNS is quite better than the existing methods. It is ensured that the proposed algorithm is advantageous in both performance, efficiency and quality of images are crucial in the latest medical systems.

Table 4
Performance analysis of most popular BC detection methods.

Methodology	Accuracy	Precision	Recall
Naïve Bayes classifier	95.61	95.65	93.61
SVM classifier	95.61	95.65	93.61
Bi-clustering and Ada boost techniques	95.75	95.72	96.26
RCNN classifier	91.3	91.3	89.3
Bidirectional Recurrent Neural Networks (HA-BiRNN)	82.50	80.09	79.03
Proposed Methodology - Deep Neural Network with Support Value (DNNS)	**97.21**	**97.9**	**97.01**

CRediT authorship contribution statement

Anji Reddy Vaka: Data curation, Writing - original draft, Software, Validation. **Badal Soni:** Supervision, Visualization, Investigation. **Sudheer Reddy K.:** Conceptualization, Methodology, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to express deep sense of appreciation and heartfelt thanks to the Dean, Senior Doctors, Administrative Staff at Mahatma Gandhi Cancer Hospital and Research Institute, Visakhapatnam, Andhra Pradesh, India for their untiring support and extended help in sharing the dataset as given in Fig. 2.

References

[1] http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.
[2] Y.S. Hotko, Male breast cancer: clinical presentation, diagnosis, treatment, Exp. Oncol. 35 (4) (2013) 303–310.
[3] https://www.biospectrumindia.com/views/21/15300/statistical-analysis-of-breast-cancer-in-india.html.
[4] S. Malvia, S.A. Bagadi, U.S. Dubey, S. Saxena, Epidemiology of breast cancer in Indian women, Asia Pac. J. Clin. Oncol. 13 (4) (2017) 289–295.
[5] Shallu, Rajesh Mehra, Breast cancer histology images classification: Training from scratch or transfer learning?, ICT Express 4 (2018) 247–254.
[6] V. Anji Reddy, Badal Soni, Breast cancer identification and diagnosis techniques, in: Machine Learning for Intelligent Decision Making, Springer, 2020.
[7] Qiao Pan, Yuanyuan Zhang, Dehua Chen, Guangwei Xu, Character-Based Convolutional Grid Neural Network for Breast Cancer Classification, IEEE, 2017, p. 31.
[8] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, Joel J.P.C. Rodrigues, A novel deep learning based framework for the detection and classiþcation of breast cancer using transfer learning, in: Pattern Recognition Letters, Elsevier, 2019.
[9] Qinghua Huang, Yongdong Chen, Longzhong Liu, Dacheng Tao, Xuelong Li, On combining biclustering mining and adaboost for breast tumor classification, IEEE Trans. Knowl. Data Eng. 32 (4) (2020) 728–738.
[10] Shweta Kharya, SunitaSoni, Weighted naive bayes classifier: A predictive model for breast cancer detection, Int. J. Comput. Appl. 133 (9) (2016) 32–37.
[11] R.D.H. Devi, M.I. Devi, Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer, Int. J. Adv. Engg. Tech./Vol. VII/Issue II/April-June 93 (2016) 98.