International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Breast Cancer Detection using Gradient Boost Ensemble Decision Tree Classifier

S. Vahini Ezhilraman, Sujatha Srinivasan, G.Suseendran

Abstract*—Detection of any abnormalities in the human is a big challenge faced by many of the field experts. One such challenge is to detect the Breast Cancer. The prime mottobehind in making this paper is to detect the breast cancer with the help of breast images in an advanced and appropriate way. In this study, an attempt is made in such a way by applying the combination of various existing technics in the extracted breast images for getting better result in detecting the Breast Cancer. Consequently,feature extracting images are appliedusing Light gradient boosting ensemble decision tree classifier for identifying benign and malign features of an image. As a result, the normal and abnormal breast cancer image is detected by combining above applications. Besides, classification accuracy and minimize classification time metrics are also achieved more appropriately than the existing detectingtechnics.

Keywords*— Gaussian training loss , Breast Cancer detection, Kullback–Leibler divergence value ,Light Gradient Boost, Base classifiers, c4.5 decision tree, Steepest Descent Function

I. INTRODUCTION

In this study, the Light gradient boost machine learning technique is introduced for breast cancer detection. This method is developed to convert the weak learners into strong learners for getting high accuracy in detecting the breast cancer. But this weak learner method is used as a classifier, which does not provide high classification accuracy. At the same time, the strong learner is an ensemble of a weak classifier, which provides the accurate classification.

This ensemble boosting technique performs the tree-based classification. A Light gradient boost machine learning technique constructs the decision tree classifier as a weak learner in the form of vertical manner. Hence it is called as Leaf-wise Decision Tree Algorithm, which minimizes the amount of training loss while comparing with any another developed algorithm.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

S. Vahini Ezhilraman, Ph.D., Research Scholar, Department of Computer Science Vels Institute of Science,Technology &and Applications Advanced Studies (VISTAS), SRM Institute for Chennai-117

Sujatha Srinivasan, Associate Professor, Department of Computer Science Vels Institute of Science, Technology &and Applications Advanced Studies (VISTAS), SRM Institute for Chennai-117

G.Suseendran, Assistant Professor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology &and Applications Advanced Studies (VISTAS), SRM Institute for Chennai-117

II. RELATED WORK

A Max-Mean and Least-Variance method was introduced in [1] for breast tumor detection. The method does not handle a large number of breast images for accurate cancer detection.

An Artificial Neural Network was introduced in [2] for breast cancer detection with the extracted features. The neural network has three different layers for cancer detection hence it was failed to minimize the detection time.

A Particle Swarm Optimized Wavelet Neural Network (PSOWNN) was introduced in [3] for identifying the breast cancer from the digital mammograms. Though the classifier provides high accuracy, the performance of the detection time unsolved.

An increased risk of breast cancer detection in women was minimized in [4] with the less false-positive test in mammographic screening. The other parameters such as classification accuracy, false positive rate were not computed.

Though the system increases the breast lesion classification accuracy, the time was not minimized. In [5], a weighted K-means support vector machine (wKM-SVM) was introduced for increasing breast cancer detection. The weighting scheme does not provide the high accuracy in breast cancer detection. A convolutional neural network-discrete wavelet and curvelet transform were developed in [6] for breast cancer detection using mammograms images. But the accurate detection of breast cancer was not attained.

An ensemble empirical mode decomposition (EEMD) was developed in [7] to discover breast cancer using ultra-wideband (UWB) microwave images. The breast cancer detection process was not efficient. A selective ensemble method KNN, SVM, and Naive Bayes were introduced in [8] to detect the breast cancer using both ultrasound images and mammography images. This ensemble method failed to concentrate on the features in images. A fully automatic mass detection technique was introduced in [9] to attain lessfalse positive rate in the breast cancer detection with high sensitivity. The technique does not minimize the detection time. A CNN-based approach was developed in [10] for the classifying the histological breast cancer images with the extracted features. The approach does not use whole-slide breast histology images. The combination of discrete cosine transform and discrete wavelet transform features were introduced in [11] for classifying the mammograms images for breast cancer detection with high accuracy rate. But, the transformation technique takes a large amount of time for detecting breast cancer.

III. ENSEMBLE CLASSIFICATION

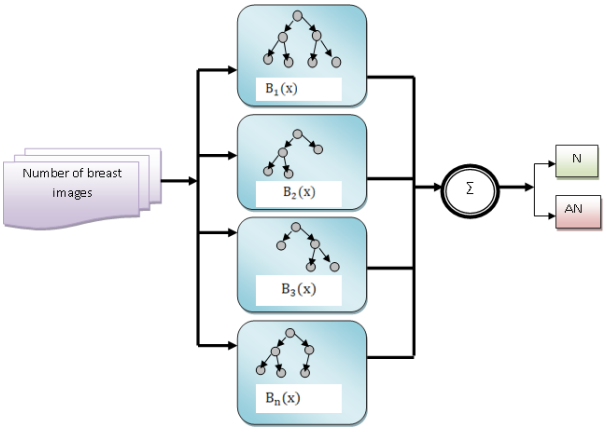


Fig .1. Ensemble classifications for breast cancer detection

Fig. 1.shows the ensemble classification process for accurate detection of breast cancer in less time. Let us consider the following training images $\{x_i, y_i\}$ where x represents the input breast images $bi_1, bi_2, bi_3, \ldots . bi_n$ and 'y' denotes a classification results. The set of base classifiers $\{B_1(x), B_2(x), B_3(x) \ldots . B_n(x)\}$ are constructed to train the input training images with the extracted features.

This ensemble classifier combines the results of all base classifier to make a strong one for minimizing the false positive rate. After that, the similar weights are initialized for all the base classifier. Then the Gaussian training loss is computed for each base classifier results based on the difference between actual and observed value. Based on the error value, the initial weights of all the classifiers are decreased or increased. Finally, the steepest descent function finds the best classifier with minimum training loss among the several base classifiers. This process increases the accurate classification and incorrect classification.

// Gaussian Light Gradient Boost Ensemble Decision Tree Classification Algorithm

Input: $bi_1, bi_2, bi_3, \ldots . bi_n$

Output: improve breast cancer detection accuracy

Begin

Collect $bi_1, bi_2, bi_3, \ldots . bi_n \in D$

Foreach bi_i

remove the noise

Extract the features

Construct base classifier $\{B_1(x), B_2(x), B_3(x) \ldots . B_n(x)\}$ with features

Compute Kullback–Leibler divergence $d\ [(p\ (f)|\ q\ (f))]$ splits the images into different classes $N\ or AB$

Combine all base classifier $y = \sum_{i=1}^{n} B_i(x)$

for each $\boldsymbol{B_i(x)}$

Initialize similar weight $\vartheta\ (t)$

Compute Gaussian training loss $\sigma\ [y, B_i(x)]$

Update the initial weights $\vartheta\ (i')$

If $[\vartheta\ (i-1)]$**then**

$B_i(x)$correctly classifies the images as $N\ or\ AN$

else

$B_i(x)$incorrectly classifies the images as $N\ or\ AN$

end if

Find classifier with lowest training loss $arg\ \min \sigma\ [y, B_i(x)]$

End for

Obtain strong classification results $\sum_{i=1}^{n} B_i(x)\ \vartheta\ (i')$

End for

end

The above algorithm clearly describes the classification based breast cancer detection. From the above dataset, the breast images are collected and noises are removed from those input images for an accurate classification. The features of the tissues from the images are extracted. After extracting the features, the light gradient boost machine learning ensemble classifier is applied for identifying cancer from the images. The ensemble classifier uses the c4.5 leaf wise decision tree as a base classifier for splitting the number of images into normal or abnormal based on the Kullback–Leibler divergence. After the classification, base classifier results are combined into a strong one by minimizing the training loss. The training loss is computed for each base classifier results. Based on the loss value, the initial weights are updated. Then the gradient decent function discovers best classifier with minimum training loss. This process reduces the false positive rate. As a result, the classification accuracy gets improved. The ensemble classification provides accurate results than the normal classifier. In general, there exists is several machine learning techniques have been presented for breast cancer detection. Recently, the size of the imageis increased and it becomes complex using traditional boosting algorithms to give faster results.

IV. SIMULATION METHOD

In this section, the simulation results and discussion of the proposed system using deep CNN [12] and SD-CNN [13] along with already existing methods are described using various parameters viz., Time Classification and Accurate Classification.

A. Sample calculation for classification accuracy

- **Proposed GLGBDTC:** Number of images correctly classified is 8 and the total number of images is 10. The classification accuracy is evaluated as given below.

$$CA = \frac{8}{10} * 100 = 80\%$$

- **Existing Deep CNN:** Number of images correctly classified is 6 and the total number of images is 10. The classification accuracy is evaluated as given below,

$$CA = \frac{6}{10} * 100 = 60\%$$

- **Existing SD-CNN:** Number of images correctly classified is 7 and image taken are ten. The classification accuracy is evaluated as given below.

$$CA = \frac{7}{10} * 100 = 70\%$$

The classification accuracy is measured by conducting the experiments using breast cancer image database with a number of images varied from 10 to 100. For each classification techniques, ten different results are described in table 1 with various input images are taken as input. The below table value clearly describes the classification accuracy using GLGBDTC technique is increased as compared to the conventional classification techniques namely deep CNN [12] and SD-CNN [13]. Let us consider the 10 images to calculate the classification accuracy. The GLGBDTC technique correctly classified 8 images from the 10 images and attaining 80% of accuracy. Whereas, the deep CNN [12] and SD-CNN [13] classified 6 and 7 images correctly from the 10 input images. Then the resultant classification accuracy of these two classification techniques is 60% and 70% respectively. The above said results are attained by the mathematical calculation. Similarly nine runs are performed and the attained results are plotted in the graphical representation.

TABLE I Classification Accuracy

Number of images	Classification accuracy (%)		
	GLGBDTC*	Deep CNN	SD-CNN
10	80	60	70
20	85	75	80
30	90	83	87
40	88	78	83
50	92	84	88
60	88	80	85
70	91	84	87
80	93	85	89
90	89	79	84
100	93	85	89

** see Table III, for GLGBDTC Result analysis*

B. Calculation For Classification Time:

- **Proposed GLGBDTC:** Total number of images is 10 and the time taken for classifying the single image is 1.9ms. Then the classification time is computed as follows,

$$CT = 10 * 1.9 = 19ms$$

- **Existing Deep CNN:** Total number of images is 10 and the time taken for classifying the single image is 3.1ms. Then the classification time is computed as follows,

$$CT = 10 * 3.1 = 31ms$$

- **Existing SD-CNN:** Total number of images is 10 and the time taken for classifying the single image is 2.6ms. Then the classification time is computed as follows,

$$CT = 10 * 2.6 = 26ms$$

TABLE. II Classification Time

Number of images	Classification time (ms)		
	GLGBDTC*	Deep CNN	SD-CNN
10	19	31	26
20	26	40	36
30	38	54	45
40	42	64	52
50	45	65	55
60	50	66	60
70	56	70	62
80	60	76	70
90	67	79	74
100	80	88	85

** see Table III, for GLGBDTC Result analysis*

Table 2 clearly shows the various simulation results of clustering time with three different classification techniques GLGBDTC technique and Deep CNN [12] and SD-CNN [13]. This is the major parameter to provide accurate treatments for the patient.

V. RESULT ANALYSIS

A. Performance Results of Classification Accuracy

Theclassification accuracy is computed as the ratios of a number of images are correctly classified as normal or abnormal to the total number of breast images taken from the database. The classification accuracy is mathematically computed as follows,

CA=(Number of images correctly classified/Total number of images)/*100 (1)

From equation (1), CA represents the classification accuracy.It is classified in terms of percentage (%). Higher the classification accuracy, the method is said to be more efficient for cancer detection.

Fig. 2 illustrates the classification accuracy for detecting cancer from the given mammogram images in the range of 10 to 100. The classification accuracy results are plotted in the two-dimensional graphical representation. The results of the proposedGLGBDTC technique and existing methods are illustrated with three different colors of lines. The results show that the proposed GLGBDTCtechnique achieves better performance than the deep CNN [12] and SD-CNN [13]. This improvement of the GLGBDTC technique is to use the ensemble classifier. Then the images are given to the input of the base classifier namely c4.5 classifier.The base classifier constructs the leaf wise decision trees to partition the images into classes based on the Kullback–Leibler divergence. The divergence is used for finding the probability of the disease occurrences with the features. The classified results are combined into single one for attaining accurate breast cancer detection. The ensemble classification technique finds the best classifier with minimum training loss. As a result, breast cancer is accurately detected from the input images.

Fig. 2.classification accuracy versus number of images

Simulation results

Totally ten various accuracy results are attained for three different classification techniques. Then the classification accuracy results of proposed GLGBDTC technique is

Compared with the other two existing classification techniques,the average comparison results show that the GLGBDTC technique increases the classification accuracy by 13% and 8 % when compared to existing deep CNN [12] and SD-CNN [13] respectively.

B. Performance Result of classification Time

The time to Classified the algorithm is defined as the amount of time required to classify the images as normal or abnormal. The mathematical formula for computing the classification time is expressed as follows,

CT=Total Images * Time taken for classification (2)

From equation (2), *CT* represents the classification time which is measured in milliseconds (ms). The calculations for classification time are provided below using three different techniques. The result displays that the proposed GLGBDTC technique effectively minimizes the classification time than the Deep CNN [12] and SD-CNN [13]. The results of the classification time with three classification techniques are plotted in the graph.

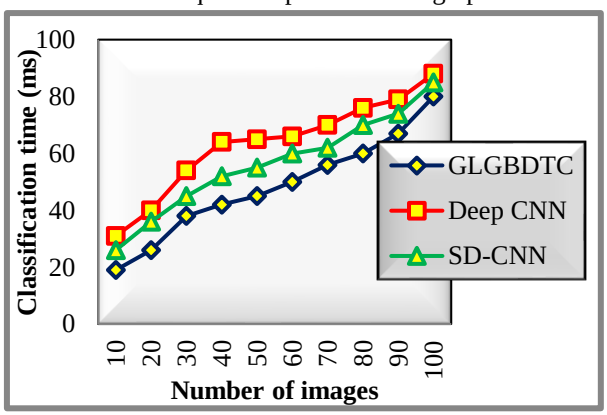


Fig. 3.classification time versus number of images

Fig. 3 depicts the simulation results of the classification time with respect to a number of breast images. The numbers of breast images are taken from the databases for detecting the brain cancer through the classification. The above graphical result shows that the classification time of GLGBDTC technique is comparatively minimized than the Deep CNN [12] and SD-CNN [13]. The existing deep convolutional neural network uses more layers for processing the input images. This process takes more time for disease classification. But the proposed ensemble classifier accurately finds the best classifier results with minimum loss using steepest descent function. This helps to reduce the classification time. In addition, the ensemble classifier categorizes the input images with the extracted features. This process also minimizes the classification time. The various results are observed with different input breast images. Then the performance results of GLGBDTC technique is compared with the two existing technique namely Deep CNN [12] and SD-CNN [13] results. After that, the average results clearly show that the classification time is significantly minimized by 26% and 16% using GLGBDTC technique than the existing Deep CNN [12] and SD-CNN [13] respectively.

The discussion of the parametric results clearly shows that the proposed GLGBDTC technique accurately detects the breast cancer through the high classification accuracy with minimum time.

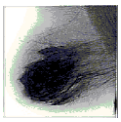


Fig.4. Feature Extracted Image

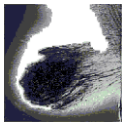
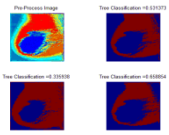


Fig 5. Light Gradient DetectionFig



6. Ensemble Classification

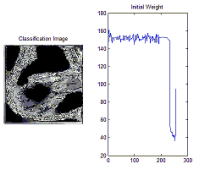


Fig 7.Weight Classification

Fig 8. Classified Image

TABLE. III Result Analysis of GLGBDTC

Total number of Images	Accuracy	Classification Time(ms)
10	80	19
20	85	26
30	90	38
40	88	42
50	92	45
60	88	50
70	91	56
80	93	60
90	89	67
100	93	80

Table IIIshows the results ofbetteraccuracy and classification time extracted from different images, usingGaussian Light Gradient Boost Ensemble Decision Tree Classification (GLGBDTC) Algorithm. The analysed outcomes are shown in the Fig 4 to Fig.8. From the above extracted figures/images, enhanced forms of normal and abnormal tissues of breast cancer are to be identified and analysed.

VI. CONCLUSION

The accurate detection of breast cancer is the ultimate goal of many experts, which was not attaineduntil now even though there were various stages of research were conducted.Thus, in this studysome of the methods like Light Gradient Boost Machine Learning ensemble classifier areapplied through the proposed classification algorithm. This method detects the cancer from theimages. By using the ensemble classifier technic, the numbers of images were split into normal and abnormal on the basis ofKullback–Leibler divergence method. For the above technic of splitting, the c4.5 leaf wise decision tree was used as a base classifier. By the end of this classification, the outcomes of base classifier are combined into a strong one by reducing or minimizing the training loss. Minimum Time Classification and Accurate Classification are the metrics acquired(which were tabulated in Table 1 and 2; and displayed in fig.2 and 3). At the end the comparison was made with already existing above mentioned two methods. In future for the further analysis other forms of similar metrics would preferred to be applied and compared for some better outcome.

REFERENCES

1. Anuj Kumar Singh and Bhupendra Gupta, “A Novel Approach for Breast Cancer Detection and Segmentation in a Mammogram”, Procedia Computer Science, Volume 54, 2015, pp. 676 – 682
2. M. M. Mehdy, P. Y. Ng,1 E. F. Shair, N. I. MdSaleh, and C. Gomes, “Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer”, Computational and Mathematical Methods in Medicine, Hindawi, Volume 2017, April 2017, pp. 1-15
3. J. Dheeba, N. Albert Singh, S. Tamil Selvi, “Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach,” Journal of Biomedical Informatics Volume 49, 2014, pp. 45–52
4. X. Castells, M. Roman, A. Romero, J. Blanch a, R. Zubizarreta c, N. Ascunce, D. Salas, A. Buron, M. Sala, the Cumulative False Positive Risk Group, “Breast cancer detection risk in screening mammography after a false-positive result”, Cancer Epidemiology, Elsevier, Volume 37, 2013, pp. 85–90
5. SungHwan Kim, “Weighted K-means support vector machine for cancer prediction” Springer Plus, Volume 5, Issue 1162, 2016, pp. 1-11
6. M. MohsinJadoon,Qianni Zhang, IhsanUlHaq, Sharjeel Butt, and AdeelJadoon, “Three-Class Mammogram Classification Based on Descriptive CNN Features”, Hindawi, BioMed Research International, Volume 2017, January 2017, pp. 1-11
7. Qinwei Li, Xia Xiao, Liang Wang, Hang Song, HayatoKono, Peifang Liu, Hong Lu, and TakamaroKikkawa, “Direct Extraction of Tumor Response Based on Ensemble Empirical Mode Decomposition for Image Reconstruction of Early Breast Cancer Detection by UWB”, IEEE Transactions on Biomedical Circuits and Systems, Volume 9 , Issue 5 , 2015, pp. 710 – 724
8. Jinyu Cong, Benzheng Wei, Yunlong He, Yilong Yin, and YuanjieZheng, “A Selective Ensemble Classification Method Combining Mammography Images with Ultrasound Images for Breast Cancer Diagnosis’’, Computational and Mathematical Methods in Medicine, Volume 2017, June 2017, pp. 1-7
9. Xiaoming Liu and ZhigangZeng, “A new automatic mass detection method for breast cancer with false positive reduction”, Neurocomputing, Elsevier, Volume 152, 2015, pp. 388-402
10. Teresa Araújo ,GuilhermeAresta, Eduardo Castro, José Rouco, Paulo Aguiar, CatarinaEloy, AntónioPolónia, AurélioCampilho, “Classification of breast cancer histology images using Convolutional Neural Networks”, PLoS ONE, Volume 12, Issue 6, 2016, pp. 1-14
11. Muhammad Talha, “Classification of mammograms for breast cancer detection using fusion of discrete cosine transform and discrete wavelet transform features”, Biomedical Research, Volume 27, Issue 2, 2016, pp. 322-327
12. Xiaofei Zhang, Yi Zhang, Erik Y. Han, Nathan Jacobs, Qiong Han, Xiaoqin Wang, Jinze Liu, “Classification of Whole Mammogram and Tomosynthesis Images Using Deep Convolutional Neural Networks”, IEEE Transactions on NanoBioscience, Volume 17 , Issue 3, July 2018, pp. 237 – 242
13. FeiGao, Teresa Wu, Jing Li, Bin Zheng, Lingxiang, Ruan, Desheng Shang, Bhavika Patel, “SD-CNN: a Shallow-Deep CNN for Improved Breast Cancer Diagnosis”, Computerized Medical Imaging and Graphics, Elsevier, Volume 70, December 2018, pp. 53-62

AUTHORS PROFILE

S.Vahini Ezhilramanis, pursuing her Doctoral Degree in Department of Computer Science, VelsInstitute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. She did her MCA Degree in Krishnasamy College of Engineering and Technology, Cuddalore and M.Phil Degree in the Department of Computer Science, in Alagappa University, Karaikudi.

Dr. Sujatha Srinivasan is Associate Professor in Computer Science in SRM Institute for Training and Development, Chennai. She got her Doctorate degree in Computer Science in January 2014 from Bharathidasan University. She has PG degrees in Mathematics and Computer Applications. She has published more than 25 papers in International Journals and Conference Proceedings. Her research interest includes Artificial Intelligence, Human Computer Interaction, Data Analytics, Evolutionary Computing, and Simulation Modelling. Currently she is working in the area of Big Data analytics, Network Security and Social Modelling.

Dr.G.Suseendran received his M.Sc., Information Technology and M.Phil., degree from Annamalai University, Tamil Nadu, India and Ph.D., degree in Information Technology-Mathematics from Presidency College, University of Madras, Tamil Nadu, India. In additional qualification, he has obtained DOEACC ‘O’ Level AICTE Ministry of Information Technology and Honor Diploma in Computer Programming. He is currently working as Assistant Professor, Department of Information Technology, School of Computing Sciences, VelsInstitute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India which is well known University. He has 10 years of teaching experience in both UG and PG Level. His research interests include Wireless Senor Network, Ad-hoc networks, IOT, Data Mining, Cloud Computing, Image Processing, Knowledge-based systems and Web Information Exploration.