Breast Cancer Detection and Classification

Poonam Kathale

Department of Electronics Engineering,

Government College of Engineering Amravati,

Amravati, India

poonamakathale2@gmail.com

Snehal Thorat

Department of Electronics Engineering,

Government College of Engineering Amravati,

Amravati, India

snehsnehu@gmail.com

2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.367

***Abstract* —Breast Cancer is more common hence, identification of BC and detection of region of breast affected is more important. Mammography screening images two views CC and MLO are widely use in diagnosis process. This paper presents the method to detect cancer region and classify normal and cancerous patient. Pre-processing operation perform on the input Mammogram image and undesirable part removed from the image, tumor region segmented from the image using morphological operation and highlighted the region on original mammogram image or if mammogram image is normal case then it shows that patient is normal. Random Forest (RF) classifiers is used for classification of BC patient and normal patient. Classification accuracy of RF is 95% for image of different patient. Processing time of RF classifier is 6.25s.**

Keywords—Breast Cancer, Craniocaudal, GLCM, Mammography Screening, Mediolateral Oblique, Random Forest

I. INTRODUCTION

Cancer is disease including irregular cell germination with potential to spread into different parts of body which means a serious health issue and it is forward in reason of death worldwide. BC is the most widespread invasive cancer in ladies and a secondary leading reason of death in women and it is becoming the main reason for the cause of disability and death in the developing countries. In Breast tumour cell growth is uncontrolled, and the cell becomes shapeless as cancer grows rapidly.

In 2019, predicted 268,600 current causes of invasive BC diagnosed in women and approximately 2,670 states diagnosed among men. Expanding, predicted 48,200 instances of DCIS diagnosed with ladies. Around 41,760 women moreover, 500 men are suspected of death from BC in 2019[1]. The survival rate of BC is incredibly influenced by malignancy's stage in the course of diagnosis. To give proper treatment to patients, early diagnosis needed and thus reduce mortality and morbidity rate. For various kinds of cancer, a High-performance diagnosis will be helpful for a medical expert to support them to diagnose and adopt appropriate treatment.

Normally, BC is treated by surgery, which is pursued by chemotherapy, hormone therapies, and radiations. At any time, disease may recur if cancer patients treated initially. Yet maximum recurrences cases tend to appear in the first Five after the treatment. The disclosure of breast cancer by choosing imaging methods. MRI, Digital imaging, and ultrasound imaging are widely used in imaging methods. Mammography is worthy and most efficient utensil for irregularities detection inside breast[2]. Mammography screening is effective for BC mortality reduction by 30-75%. Two aspects of mammography screening assist experts for BC identification. However, diagnosis precision depends upon individual approaches and perceptions of a medical specialist.

This is possible with the help of CAD system. The CAD scheme possessing two natures; primary is unilateral including the following is bilateral. An individual aspect is practiced in the unilateral. CC and MLO views both are typical mammographic aspects [1]. In bilateral CAD, the combination among CC plus MLO views had been proved to enhance the efficiency of BC identification.

II. LITERATURE SURVEY

M. A. Nasser [1] matching strategy is done to identify the relationship in candidate positions in multiple mammographic views SIFT is adopted to find candidate points. M. A. Berber [2] introduced a feature extraction method for a dangerous mammogram and its class. 7 features for GLCM offered. It also introduced three composite classifications named Wavelet-CT1, Wavele-CT2 and ST-GLCM. SVM is used for classification. Specificity, sensitivity and accuracy for GLCM is 96.88%, 98.43% and 97.91%. F. Ting [3] to BC ranking that algorithm which is a self-regulated perceptron neural-network (MLNN) is designed. ML-NN categorized the medicinal information pictures as a healthy patient, malignant and benign sufferer with sensitivity, and accuracy, the specificity of 91.23%, 90.73% and 90.68 % respectively. N. Tariq [4] suggested approach, texture features of mammogram estimated implementing Gray Level Co-occurrence Matrix, of estimation components most powerful characteristics producing enormous participation to achieve the desired output were obtained and executed to Artificial Neural Network to train and analyse, as ANN is commonly followed in various fields in medical diagnosis, pattern recognition, machine learning. To this task, mini-MIAS dataset used and overall specificity, and sensitivity, and accuracy obtained through adopting the recommended method is 100%, 99.3%, and 99.4% sequentially. Huda Al-Ghaib et al [7] presented the details about mammography, mammogram is a x-ray image of human breast which is used for detection and diagnosis of changes in breast tissue. It is two-dimensional projection of a complex three-dimensional object. Breast cancer types depending upon

978-1-7281-4142-8/$31.00 ©2020 IEEE

their shapes and texture. Details about types of mammography techniques i.e. screening and diagnostic mammography. CAD system help for detection of suspected lesion locations and reduction of false-positive rates.

Early disclosure of breast carcinomas increases the survival rate with more successful treatment options. To detect lesions at their early stages, CAD helpful in reading screening. [9] For the diagnosis of microcalcification, the morphological bandpass filter (MBF) is used. Employing MBF can ROI among True Positive Rate and False Positive Rate are 93.07% and 4.31% respectively.

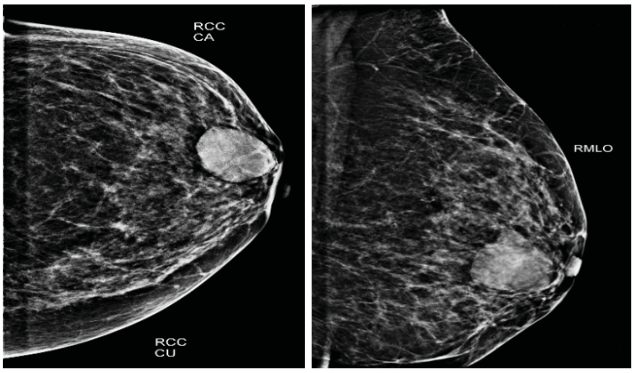

Fig -1: CC and MLO views of Mammogram

III. METHODOLOGY

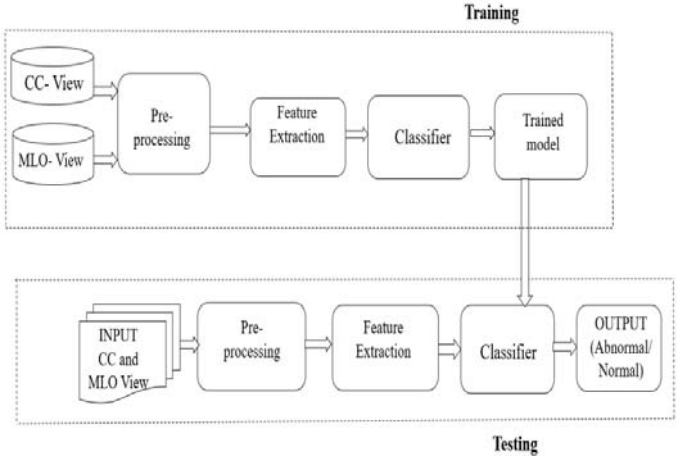

Fig -2: Block Diagram of System Implemented

A. Pre-processing of Mammographic Images

The purpose of the pre-processing step is to produce data which are compatible with the breast cancer region detection and classification system. Input image is grey image and resize operation performed to get specific size image for further operation. For calculating the grey level of an image and the number of pixels Histogram, plot action is performed. Image transformed into a white and black (binary) image to locate ROI with a specified threshold rate. Small spots remove those that possess fewer pixels than a specific value. To prepare refined binary image tiny spaces filled. An annoying portion of the image is expelled.

- Median Filtering

Image Enhancement includes refining, adjustment, and resizing procedure. Median filtering eliminates undesired noise from an image. One of the traditional image enhancement methods is Median filtering.

- Resizing

Resize is important for images because the images having different resolution so it is better to invert all the images in same resolution for further processing.

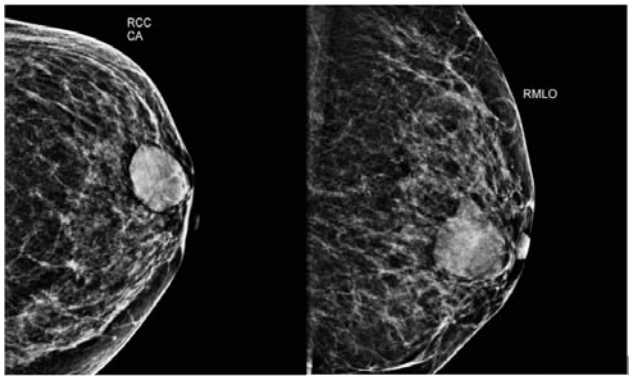

Fig -3: Original Input Image of CC and MLO view

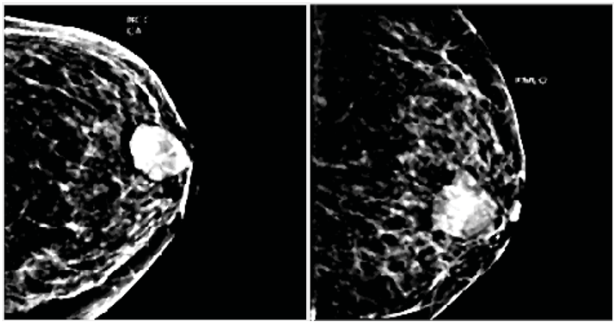

Fig -4: Preprocessed Input Images

B. Feature Extraction

Feature extraction described in the form of features dataset produced after loading training database, feature evaluation, and analysis conducted following loading the experiment training database. Feature extraction gives Miscellaneous Features such as Mean and Texture based features such as GLCM and Entropy.

Entropy: The statistical pattern of randomness is Entropy is a that, employed to determine the texture of input greyscale image.

$$E= -sum(p.*log2(p))$$

Mean: Estimates mean intensity about the gray level of an image. In the reference image, all the pixels intensity values summing up then dividing by the total number of pixels to produce Mean intensity.

Grey Level Co-occurrence Matrix: The GLCM used to characterize images based on texture. GLCM measures how frequently a pixel with grey level value occurs either horizontally or vertically. Co-occurrence matrix can be represented like P(i,j|d,θ) where i and j are the grey level values at distance d with an angle θ.

μ= Mean value of P

μ_x & μ_y = Mean value of P_X & P_Y

σ_x & σ_y = Standard Deviation of

G = Size of co-occurrence matrix

Energy: Gives the total of squared components in GLCM. Likewise, known uniformity either angular second moment.

$$\text{Energy} = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j)^2$$

Correlation: Count the collective probability appearance of this particularized pixel pairs.

$$\text{Correlation} = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \frac{\{i,j\} \times P(i,j) - (\upsilon_x - \mu_x)}{\sigma_x \times \sigma_y}$$

Contrast: Count local varieties in GLCM.

$$\text{Contrast} = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j) \right\}, |i-j| = n$$

Homogeneity: Count the close-ness of the combination of details in GLCM to GLCM diagonal.

$$\text{Homogenity} = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \frac{P(i,j)}{1+|i-j|}$$

Feature Extraction is an important stage before any classification. Features of the normal patient and cancer patient mammogram image has been calculated as shown in Mean, Entropy and GLCM show a significant difference in the features of both the images. From these feature values, images of the cancer patient and the normal patient can be classified.

C. Classification

Classification of the cancer patient and normal patient is achieved using RF classifier. Classification is mainly based on features extracted from training dataset. To classify accurately training dataset should be proper and perfect. Mammographic views of cancer patient and normal patients collected from local hospital. Random Forest is a supervised learning algorithm and it can be applicable for both regression and classification. This algorithm constructs forest by adopting number of decision trees. Random forest is more prosperous with a greater number of trees, but it has certain limitation about number of trees which affects the out of bag (OOB) error. Here in random forest to modify decision trees, information gain or Gini index approach is not adopted.

Decision tree has certain downsides like over fitting, high variance, and low biased tree. Over fitting occurs in algorithm when data hold noise. High variance means due to small variation in database, the system become unstable. The model unable to work properly on fresh new data called low biased, this is due to highly complicated decision trees. All these drawbacks are overcome in the Random forest tree ensemble algorithm.

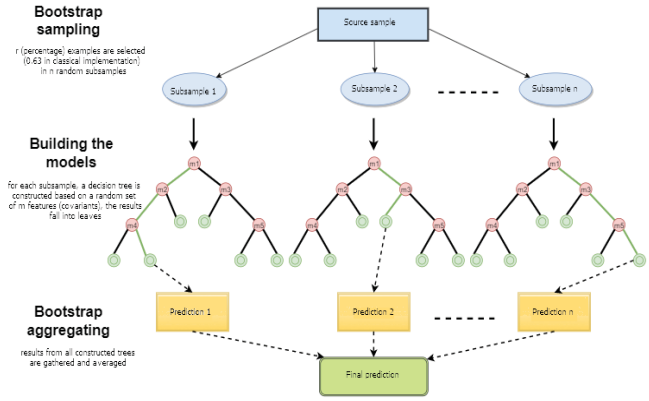

Fig -5: Flow of Random Forest Classifier

The significant features are presented for feature description of input images the feature vectors are obtained from these are classified by the rf method the outcome of the distribution process is the confusion matrix and OOB error rate. The classification process is an indicator for the efficiency of technique. The database was divided 64% training set and 36% test set approximately. The value of number of trees set to 20 for optimizing the performance of the RF classifier.

D. Segmentation of Cancer Region and Detection

The preprocessed image is then converted into a binary image by applying a certain threshold. All the connected components in the image are removed that having fewer than a certain pixels. Components in the image are suppressed which are lighter than their surroundings and connected to an image border. Finally, boundaries of the

segmented image are highlighted in the original image. In this way cancer region is detected in the breast as shown in figure 6.

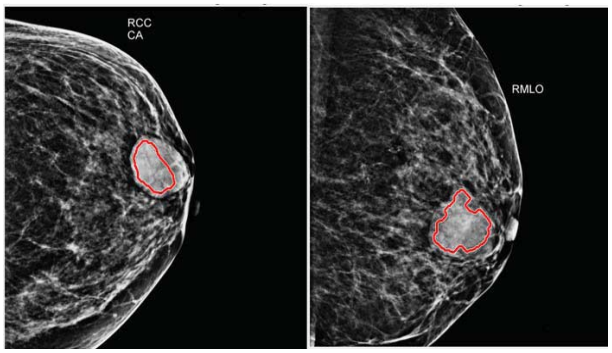

Fig -6: Cancer Region Highlighted in Both CC and MLO Mammogram

IV. Result and Diccussion

Performance of the system is tested on Mammographic Images of total 56 patients. Dataset consist of 72 of right and left breast of 18 patients and 152 views of different normal 38 patients. Initially RF classifier trained using training dataset of 24 mammographic view images of normal and abnormal patients. RF classifier gives the accuracy of 100% for the training dataset. Both the classifier tested on testing dataset and achieves the classification accuracy of 95.3 %. All the performance parameters are analyzing using confusion matrix.

The figure 7 shows that great efficiency is achieved for learning and testing of RF classifiers using a combination of Mean, GLCM, and Entropy. Therefore, those features performed in the method during the classification of normal and cancerous patient. As shown in the confusion matrix, out of 12 views, 2 views are incorrectly detected and out of 116 normal mammographic views 4 are incorrectly detected. Hence an error rate of 5% has occurred. Accuracy of 95.03% is achieved by RF classifier, therefore, the error rate is 4.7%, specificity is 83.33%, sensitivity is 96.55%, recall is 96.55, precision is 98.25%.

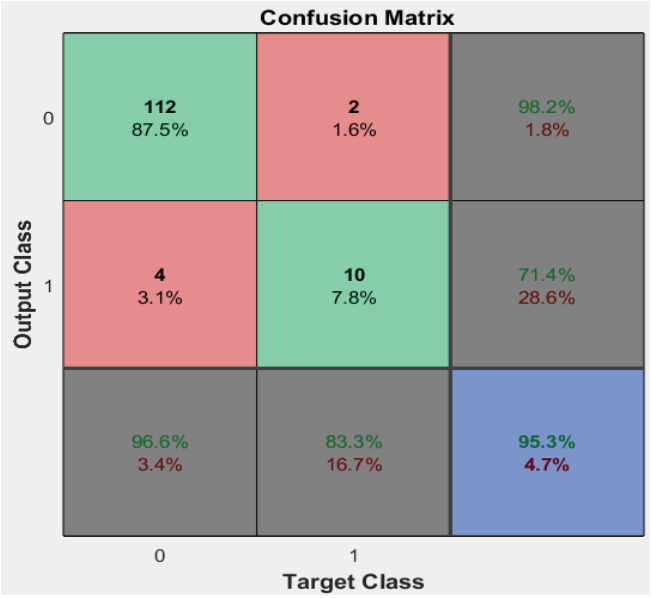

Fig -7: Confusion Matrix of Random Forest

V. Conclusion

The present study demonstrates the effectiveness of different features and the Random Forest Algorithm for breast cancer detection and classification. The cancerous area is segmented upon the base of the gray level intensity of the mammographic image. The random forest classifier achieves the classification accuracy of 95.3%. GLCM, Entropy and Mean are features utilized to examine the texture of the image and perform a vital role in the classification process. The processing time of the RF classifier for training is 6.25 sec. and for testing is 3.16 sec.

Eventually, to attain the higher accuracy training database need to be proper and fitting features are expected during the feature extraction. Consequently, accuracy can be increased in the future by performing different features.

Acknowledgment

I would like to express my deep gratitude to Mrs. S. S. Thorat, Assistant Professor, Government College of Engineering Amravati, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank Dr. P. R. Somwanshi, Dean of Dr. Panjabrao Deshmukh Memorial Medical College (PDMMC), Amravati for granting permission to take CT scan dataset. My grateful thanks are also extended to Dr. S. P. Kothari, Head of Radio-diagnosis Department, PDMMC, Amravati for her help in allowing CT scan dataset. Also, to I would also like to extend my thanks to the Mr. K. V. Tayade Mammography technicians of the Radio-diagnosis department, PDMMC, Amravati for their help in offering me the resources in running the program.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

REFERENCES

[1] Mohamed Abdel-Nasser, Antonio Moreno, Mohamed A. Abdel wahab, Adel Saleh, Saddam Abdulwahab1, Vivek K. Singh and Domenech Puig, "Matching Tumour Candidate Points in Multiple Mammographic Views for Breast Cancer Detection", 2019 International Conference on Innovative Trends in Computer Engineering (ITCE'2019), Aswan, Egypt, 2019.

[2] F. F. Ting, K. S. Sim, "Self- regulated Multilayer Perceptron Neural Network for Breast Cancer Classification", International Conference on Robotics, Automotion and Sciences (ICORAS), 2017.

[3] Mohamed A. Berbar, "Hybrid methods for feature extraction for breast masses classification", Egyptian Informatics Journal, 2017.

[4] Assistant Prof T. Krishna Chaitanya, P. Chandra Sekhar Azad, "Neural Network Based Classification of Digital Mammograms using DCT Coefficients", International Journal of Advance Engineering and Research Development, 2017.

[5] Mohamed Abdel-Nasser, Jaime Melendez, Antonio Moreno, and Domenec Puig, "The Impact of Pixel Resolution, Integration Scale, Pre-processing, and Feature Normalization on Texture Analysis for Mass Classification in Mammograms", Hindawi Publishing Corporation International Journal of Optics Volume 2016.

[6] The top 10 causes of death. World Health Organization. http://www.who.int/mediacentre/factsheets/fs310/en/. Accessed,2015

[7] Sami Dhahbi n, Walid Barhoumi, Ezzeddine Zagrouba, "Breast cancer diagnosis in digitized mammograms using curvelet moments", 2015.

[8] Loris Nanni, Sheryl Brahnam, Stefano Ghidoni, Emanuele Menegatti, Tonya Barrier, Dierent "Approaches for Extracting Information from the Cooccurrence Matrix", 2013.

[9] R. R. Janghel, Anupam Shukla, Ritu Tiwari, Rahul Kala "Breast Cancer Diagnosis using Artificial Neural Network Models", 3rd International Conference on Information Sciences and Interaction Sciences, 2010.

[10] Rangaraj M. Rangayyan, Thanh M. Nguyen, F_abio J. Ayres, and Asoke K. Nandi, "Effect of Pixel Resolution on Texture Features of Breast Masses in Mammograms", Journal of Digital Imaging, 2009.

[11] Leonardo de Oliveira Martins, Aristofanes, Anselmo Cardoso de Paiva and Marcelo Gattass, "Detection of Breast Masses in Mammogram Images using Growing Neural Gas Algorithm and Ripley'K Function", Journal of Signal Processing Systems Volume 55, 2008.

[12] Arnau Oliver, Xavier Llado, Jordi Freixenet, and Joan Mart, "False Positive Reduction in Mammographic Mass Detection Using Local Binary Patterns", MICCAI, 2007.

[13] Arnau Oliver, Xavier Llado, Jordi Freixenet, and Joan Mart, "False Positive Reduction in Mammographic Mass Detection Using Local Binary Patterns", International Conference on Medical Image Computing and Computer Assisted Intervention, 2007.

[14] Saskia van Engeland, Nico Karssemeijera, "Combining two mammographic projections in a computer aided mass detection method", 2007.

[15] JuCheng Yang, DongSun Park, "Detecting Region-of-Interest (ROI) in Digital Mammogram by using Morphological Bandpass Filter", IEEE International Conference on Multimedia and Expo (ICME), 2004.

[16] Hamid Soltanian-Zadeh, Farshid Raee-Rad, Siamak Pourabdollah Nejad D,"Comparison of multiwavelet, wavelet, Haralick, and shape features for micro calcification classification in mammograms", Journal Of Pattern Recognition Society, 2004.

[17] SCIENCEphotoLibrary Breast cancer, adenocarcinoma, mammogram:

[18] https://www.sciencephoto.com/media/910389/view/breast cancer adeno carcino mamammogram