Breast Cancer Detection Using Supervised Machine Learning Algorithm

Shashidhar R[1], Arunakumari B N[2], Naziya Farheen H S[3], Puneeth S B [4],Santhosh Kumar R[5], Roopa M[6]

[1]Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysruru, India
[2]Departement of Computer science & Engineering, BMS Institute of Technology and Management, Bengaluru, India
[3]Department of Electronics and Communication Engineering Presidency University Bangalore, India,
[4]Dept. of Electronics and Communication Engineering, Navkis College of Engineering, Hassan, India
[5,6]Department of Electronics and Communication Engineering Dayananda sagar College of Engineering Bengaluru, India,
[1]shashidhar.r@sjce.ac.in, [2]arunakumaribn@bmsit.in, [3]farheen.naziya14@gmail.com [4]puneeth.sb@presidencyuniversity.in, [5]rsanthosh.kumar665@gmail.com,[5]surajroopa@gmail.com

***Abstract*— The most commonly causing cancer among Indian women is breast cancer and it effecting all over world with its impact. According to the medical reports of breast cancer patients in India were unable to hold the pain and about half of them are dying. In the proposed work used a machine learning algorithm to decrease the pre-processing time and to detection the symptoms and for better accuracy. The system is trained pre-processed image of fed to the system which are in the form of mammograms in common the X-ray of breast. The system which has the data segregated into the training and testing datasets analyses the input images based on the characters or the labels assigned to them done with the application of few of the algorithms which are present in the machine learning we compare the data or the image and probable output based on the character labels is obtained in the form of result. Compared to existing work and the proposed machine learning model as a serious of combination and permutations of algorithms lead to increase in the efficacy of the result and got the accuracy of 97.4% using random forests algorithm.**

Keywords*— *Breast cancer, machine learning, X-ray, random forests algorithm.

1. Introduction

Breast cancer signifies unique of the diseases that as more losses each year. Breast cancer is the utmost collective cancer amongst women universal secretarial for 25% of all cancer cases and pretentious 2.1 million persons in 2015 primary diagnosis suggestively rises the likelihoods of persistence. The existing methods are Machine learning, method of training machines with data to make the decision for same conditions and its application can be observed in various domains such as medical, network, object identification and security etc. There are 2 machine learning types that is single and hybrid approaches as for instance Support vector machine

(SVM), Artificial neural network m (ANN), Gaussian mixture model (GMM), Linear regressive classification (LRC), K- Nearest neighbor (KNN), Weighted hierarchical adaptive voting ensemble (WHAVE), etc. Classification and data mining method is an actual way to categorize statistics [1]. Particularly in the field of medical, these are broadly used in identification and analysis to sort results. Support Vector Machine, Decision Tree, Naive Bayes and k-NN on the Wisconsin Breast Cancer (original) datasets is accompanied [12]. Categorizing data with reverence to competence and proficiency of individually algorithm in relation to precision, accurateness, intuition and specificity is done. Outcomes illustration that SVM gives the utmost accuracy with lowest error rate [2]. The key challenges in cancer recognition are how to categorize tumors into malignant or benign machine learning techniques can theatrically improve the accurateness of diagnosis".

"Breast Cancer is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. In the year 2018 according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women".

The main tribute of using machine learning in early breast cancer detection is to enable the prediction and improving accuracy of decision making. By using this machine learning the tumor can be identified as malignant or benign hence the unnecessary surgeries and painful operations can be decreased. The machine learning can be more accurate by providing more dataset. Machine learning does not require human intervention it gives the ability for the machine to learn on its own. These algorithms increases accuracy and effectiveness as the machine gains experience this helps to get the better decision outcome. Breast cancer detection using machine learning has achieved successfully with accuracy up to 97.4%. By using this machine learning the output is effective and faster and reduces the complexity. Here we have used combination of classifiers & algorithms such as decision tree algorithm, random algorithm and logistic regression helped to achieve high accurate and efficient model [11].some of the image compression techniques explained based on region on interest [20]. In 2020 30% of newly diagnosed cancer in women as per the survey. Proposed work used the supervised machine learning

algorithm used for detection. Here three algorithms used namely Logistic Regression, Decision tree algorithm, Random forests algorithm.

2. Methodology

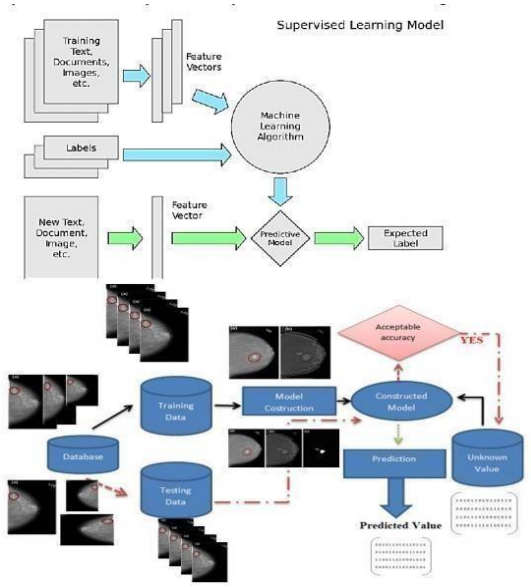

Fig. 1 Block diagram of Supervised Machine learning algorithm

In this section discuss the proposed methods. Here, 569 patients' images were used for analysis, who had a wide range of 33 class labels assigned to them. Each class label is a parameter of the cancerous non- cancerous cells, and we make an attempt to predict the cancer just by the input image of scanned breast image. The approach described states the image is resized to required number of pixels, and weper form both the model optimization and predictions on these down-scaled images [3]. This model uses the clinical dataset which contains images of mammogram of around 569 patients the dataset which is used for the preprocessing has been assimilated in the usage of mammograms and then for the Saturn parameters which are stated below to acquire these parameters the pre-processing has been implemented then obtain data set has been divided into training data set under supervise machine learning models this medical image dataset has been then distributed into practice and testing data sets for the purpose of evaluation and validation of the project the training data group is further cut into the feature vectors weather in to the machine learning algorithm a predictive model has been generated with the help of the labels then with the use of the algorithms which have been defined in the machine learning algorithms the output would be predicted by the process of voting that is which algorithm oats for the highest number of near the expected values those would be given as the expected output labels which are in the form of ones and zeros in the current project which would help us to accommodate more data at a large scale [6]. In the methodology steps are followed as given below.

• Select the input image • Pre-processing • Image Augmentation • Building Convolutional Architecture • Model Testing • Testing for output • Accuracy Testing • Result Comparison with actual dataset Select the input image: The First step is need to select the image containing a breast cancer cell in it. Use various Python libraries to read an image. Here using OpenCv.

Resize the Image: Resize the image: The height, length pixel ideals of the images are improved according to the requirements.

Image = cv2.resize(image, (height, width)).

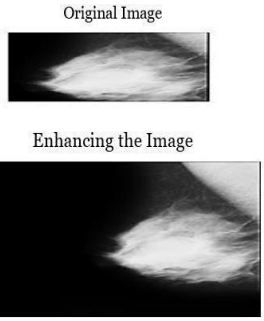

Fig. 2 Resized Images

Converting an Image into Array: Further the image data is converted into an array and stored in an array.

Image = img_to_array (image).

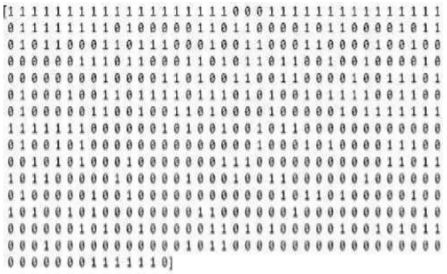

Fig. 3 Image as array form

Image Augmentation: Next step is to rotate the image to different angles for better accuracy and training purposes. Image Data Generator () function is used for augmentation

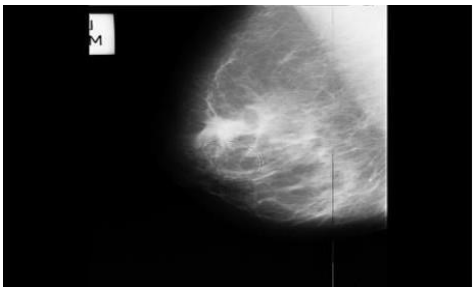

Fig. 4 Augmented Images

Model Training: Model is trained using the data set using Epochs.

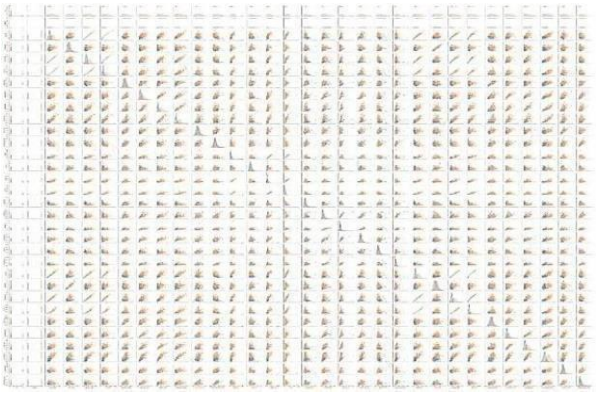

Fig. 5 Model Training

3. Proposed Methodology

In this section the different algorithms used in the proposed approach is explained in detail.

3A. Logistic Regression

A process to evaluate a data-group that has a needy variable and single or added autonomous parameter to forecast the consequence in a binate variable meaning it will have only two outcomes.

Linear regression equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots. + \beta_n X_n$$

where, *y* indicates reliant parameter that is expected. β_0 is the *y*-intercept, that is fundamentally the point on the line which traces the y-axis. β_1 is the slope of the line. *x* here represents the autonomous variable that is used to foresee our subsequent needy value.

- Sigmoid operation $p = \frac{1}{1+e-y}$
- Sigmoid operation applied for the linear regression.
- Logistic regression equation:

$$p = \frac{1}{1 + e - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_n x_n)}$$

$$\frac{p}{1-p} = \exp\,(b_0 + b_1 x)$$

b_0 is the logistic regression constant; it moves the arch in left and right. b_1 is the gradient of the arch.

By artless transformation the logistic regression equation can be written in terms of an odds ratio [11][5].

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

Lastly, enchanting and add logarithmic on both the sides. Transcribe the mathematical model in terms of log odds (logit) that a linear function of the forecasters. The b_1 coefficient is the amount the log it changes with changes with a one-unit changes in x.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots, + b_p x_p)}}$$

As mentioned in the earlier section, logistic regression can hold any numeral and/or definite parameters.

3B. Decision Tree Algorithm

It is tree like structure, wherever an inner knot denotes feature, the division denotes a decision statute, and outcome represented by leaf node. The Gini Catalog reflects a binary splitting for every characteristic. A weighted quantity of the contamination of every divider can be added. If a binary splitting on feature that dividers data D into D1 and D2, the Gini index of D is:

$$Gini_A(D) = \frac{D_1}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2)$$

For example, the discrete valued aspect, the subset values provide the least gini index on the preferred splitting attribute. Another case of continuous valued characteristics, the plan is to choose every brace of neighboring standards as an imaginable split fact and point with slighter gini index preferred as the splitting theme [11][5].

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

3C. Random forests Algorithm

It can be used together for classification and regression and also the utmost flexible and informal to apply the algorithm. Forestry is encompassed of trees.

Algorithm creates decision trees on arbitrarily designated data sections, gets prediction from each tree and selects the best explanation by means of voting.

When using the Algorithm to resolve regression problems, you are using the mean squared error (MSE) to how your data branches from each node [11].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where, *N* = Number of data points, *f*= value returned by the model, y_i is the actual value for data point i

$$MSE = 1/N \sum_{i=1}^{N} (fi - yi)^2$$

4. Result

In this division discuss the outcome of the proposed model, so we have chosen 80*80 sized image models for the testing purpose. The high accuracy is because of combination of the classifier that we have used. If the number of epochs, batch size and image size are increased we may get more accuracy in the model. Below is the graphical image representation of classifiers heat-map during accuracy test.

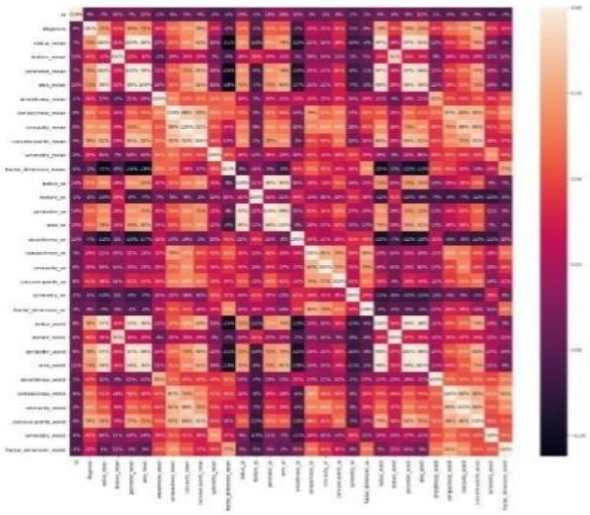

Fig. 6. Heat map plot of the model

```
[[86  4]
 [ 4 49]]
Model[0] Testing Accuracy = "0.94405594405594405!"

[[89  1]
 [ 5 48]]
Model[1] Testing Accuracy = "0.958041958041958!"

[[87  3]
 [ 2 51]]
Model[2] Testing Accuracy = "0.965034965034965!"

[[88  2]
 [ 3 50]]
Model[3] Testing Accuracy = "0.965034965034965!"

[[85  5]
 [ 6 47]]
Model[4] Testing Accuracy = "0.9230769230769231!"

[[84  6]
 [ 1 52]]
Model[5] Testing Accuracy = "0.951048951048951!"

[[87  3]
 [ 2 51]]
Model[6] Testing Accuracy = "0.965034965034965!"
```

Fig. 7. Accuracy Plot of a different classifier Model

Table 1. Different Algorithm Training Accuracy

S.N.	Model	Training Accuracy
1	Logistic Regression	99.06%
2	K Nearest Neighbor	97.6%
3	Support vector machine (Linear Classifier)	98.82%
4	Support vector machine(RBF Classifier)	98.35%
5	Gaussian Naïve Bayes	95.07%
6	Decision tree Classifier	100%
7	Random Forest Classifier	99.53%

```
Model  1
              precision    recall  f1-score   support

           0       0.95      0.99      0.97        90
           1       0.98      0.91      0.94        53

    accuracy                           0.96       143
   macro avg       0.96      0.95      0.95       143
weighted avg       0.96      0.96      0.96       143

0.958041958041958

Model  2
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143

0.965034965034965

Model  3
              precision    recall  f1-score   support

           0       0.97      0.98      0.97        90
           1       0.96      0.94      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.96      0.97      0.96       143

0.965034965034965

Model  4
              precision    recall  f1-score   support

           0       0.93      0.94      0.94        90
           1       0.90      0.89      0.90        53

    accuracy                           0.92       143
   macro avg       0.92      0.92      0.92       143
weighted avg       0.92      0.92      0.92       143

0.9230769230769231

Model  5
              precision    recall  f1-score   support

           0       0.99      0.93      0.96        90
           1       0.90      0.98      0.94        53

    accuracy                           0.95       143
   macro avg       0.94      0.96      0.95       143
weighted avg       0.95      0.95      0.95       143

0.951048951048951

Model  6
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143

0.965034965034965
```

Fig. 8 Model 1 – 6 accuracy using confusion matrix & precision values

4A. Accuracy of various Algorithms

Table .2 Comparison with Existing Algorithm

S. N.	With Comparison	Architecture used	Accuracy (%)
1	[1]	Resnet50	92.1
2	[2]	ANN	96.23
3	[3]	SVM, RVM	96.5
4	[4]	Decision tree & Logistic Regression	95.23
5	Proposed work	Supervised decision tree	97.4%

4B. Heat Image

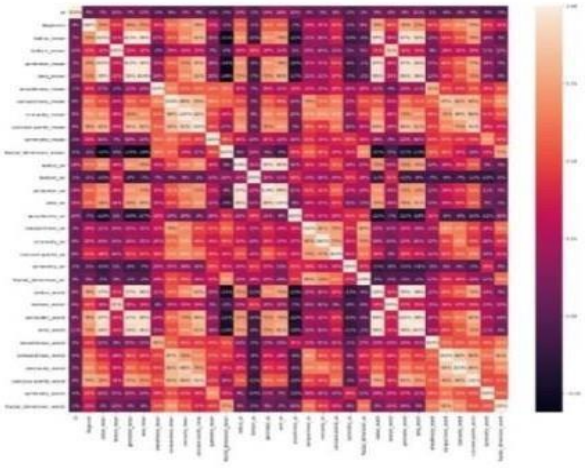

Fig. 9. Heat map of independent variables X & Y

Conclusion

Breast cancer detection using machine learning has achieved successfully with accuracy up to 97.4%. By using this machine learning the output is effective and faster and reduces the complexity. Here we have used combination of classifiers & algorithms such as decision tree algorithm, random algorithm and logistic regression helped to achieve high accurate and efficient model. The results shown in decision tree classifiers prediction and in the actual classification of the patients which presenting ones as malignant (cancerous) and zeros as benign (non- cancerous). This model can predict a greater number of correct values than negatives. By detecting the breast cancer at early stage, the cancer can be curable and the patients can avoid painful surgeries. The overall computational time for the preprocessing would be 3.5sec & the time for the processing stage would be around 5sec for the number of dataset considered, this time could vary depending upon the number of dataset that has been chosen.

References

[1]. M.M.Mehdy, E.E.Shair and P.Y.Ng, "Artificial Neural Networks in Image Processing for Earlier Detection of Breast Cancer", Hindawi, Computational and Mathematical Methods in Medicine, Volume 2017, Article ID 2610628.

[2]. Vishnukumar K.Patel, Prof.Syed Uvaid and Prof.A.C.Suthar, "Mammogram of Breast Cancer Detection Based Using Image Enhancement Algorithm", Internationa Journal of Engineering Technology and Advanced Engineering, Volume 2, Issue 8, August 2012.

[3]. Moh'd Rasoul A Al-Hadidi, Mohammed Y. Al-Gawagzeh, "Solving Mammography Problems of Breast Cancer Detection Using Artificial Neural Networks and Image Processing Techniques", Indian Journal of Science and Technology, Vol 5, No.4 (April 2012), ISSN: 094-6846.

[4]. Bhagyashri k Yadav, Dr. Prof. M. S. Panse, "Virtual Instrumentation Based Breast Cancer Detection and Classification Using Image-Processing", International Journal of Research and Scientific Innovation (IJRSI), Volume V, Issue IV, April 2018.

[5]. Melanie A. Sutton, "Breast Cancer Detection Using Image Processing Techniques", IEEE International Conference on Fuzzy System Febraury 2000.

[6]. A. D. Belsare and M. M. Mushrif, Histopathology Image Analysis Using Image Processing Technique, Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.4, August 2012.

[7]. "Latest Global Cancer Data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018", International Agency for Research on Cancer, World Health Organization, 12 September 2018.

[8]. Oeffinger, K. C. et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA 314, 1599–1614,2015.

[9]. Lehman, C. D. et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiol. 283, 49–58, 2016.

[10]. Elter, M. Horsch, A. CADx of mammographic masses and clustered micro classifications: A review. Med. Phys. 36, 2052–2068, 2009.

[11]. Fenton, J. J. et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. New Engl. J. Medicine 356, 1399–1409 2007.

[12]. Cole, E. B. et al. Impact of Computer-Aided Detection Systems on Radiologist Accuracy With Digital Mammography. Am. J. Roentgenol. 203, 909–916 2014.

[13]. Lehman, C. D. et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer- Aided Detection. JAMA Intern. Medicine 175, 1828– 1837, 2015.

[14]. LeCun, Y., Bengio, Y. Hinton, G. Deep learning. Nature, volume 521, pp 436–444 , 2015.

[15]. Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening. Clin Cancer Res. 2018;24(23):5902- 5909. doi:10.1158/1078-0432.CCR-18-1115.

[16]. Kim, E., Kim, H., Han, K. et al. Applying Data-driven Imaging Biomarker in Mammography for Breast Cancer Screening: Preliminary Study. Sci

Rep 8, 2762 (2018). https://doi.org/10.1038/s41598-018- 21215-1.

[17]. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology, an overview and future trends. Med Image Anal.2018;47:45-67. doi:10.1016/j.media. 2018.03.006.

[18]. Burt JR, Torosdagli N, Khosravan N, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. Br J Radiol.2018;91(1089):20170545. doi:10.1259/bjr.20170545

[19]. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303- 312. doi:10.1016/j.media.2016.07.007.

[20]. Shreekanth T., Shashidhar R. (2018) An Application of Image Processing Technique for Compression of ECG Signals Based on Region of Interest Strategy. In: Hemanth D., Smys S. (eds) Computational Vision and Bio Inspired Computing. Lecture Notes in Computational Vision and Biomechanics, vol 28. Springer, Cham. https://doi.org/10.1007/978-3-319-71767-8_85