Breast Cancer Detection and Prediction using Machine Learning

[1]Gemechu Keneni, [2]Raghavendra R

[1]Student, [2]Assistant Professor
Department of Master of Science in Information Technology
Jain (Deemed-to-be) University Bangalore, India

***Abstract*: One the top type of cancer in women takes twenty-five percent of all cancer death around the globe is breast cancer. Proper and early treatment is the best solution for best diagnosis. Manual diagnostic needs experienced pathologists and much amount of time. Automated technique of detecting breast cancer improves accuracy and saves the specified diagnosis time. Therefore, the aim of this thesis to build up a methodology which enable detecting to maximize the number of breast cancer, identified at infant stage increase effectiveness of the treatment so that to reduce the number of death from breast cancer. Detecting breast is one of the solutions to effective treatment of breast cancer. I use different machine learning algorithm like Logistic regression, decision tree and random forest classifier to forecast if the tumor is not cancer.**

The proposed techniques were evaluated employing a confusion matrix, and classification performance report back to assess which features a higher classification potential. The logistic regression algorithm has achieved an average accuracy of 95%, average precision of 95.0%, average recall 95.0% and an average F1 value of 95.0% over a test data-set of previously unseen 143.The decision tree algorithm has achieved an average accuracy of 93%, average precision of 93.0%, average recall 93.0% and an average F1 value of 93.0% over a test data-set of previously unseen 143.The random forest classifier algorithm has achieved an average accuracy of 96%, average precision of 96.0%, average recall 96.0% and an average F1 value of 96.0% over a test data-set of previously unseen 143.From the analysis of the experimental results, the random forest algorithm gives better results than the other supervising machine learning classifiers. The accuracy of the model is 96 % so we can see a few wrong predictions but mostly this model is successful in predicting a tumor Malignant (M) (harmful) or Benign (B) (not harmful) based upon the features provided in the data and the training given.

Index Terms: Breast Cancer, random forest, logistic regression, decision tree, benign, malignant.

I. INTRODUCTION

The breast is an organ of the human frame, which is classified as an exocrine gland, designating its functionality to provide breast milk. The breast is specifically constructed up of fat, connective tissue, and masses of small structures, known as lobules. [1] The cell is the tiniest organic unit with diverse systems and able to unbiased continuation.[2] cancer can be defined as a nation of sickness that terminates the cells to answer to regular stimuli. An irregular cell that spreads and breaks out of control will supply rise to a tumor. A tumor that has expanded beyond the layer of tissue where in it advanced, and is spreading into encircling healthful tissue is said to be persistent. [3] Breast cancer is described as a cancerous tumor that arises from breast tissue, regularly from the epithelial cells within the wall of milk ducts that supply the ducts with milk of the breast.[4,5] The primary peril determinants for breast cancer are being female, antique age, and genetic elements. other relevant factors include past due first childbirth, null parity and menopause [6, 7]. some researchers also show evidence of alcohol ingesting being a hazard factor [8], while breastfeeding and physical interest have proven to lower the uncertainty of breast most cancers [9, 10]. Breast cancer prognosis is the method of locating the existence of most cancers mobile in the patient's breast and multidisciplinary attempt. at some stage in the breast most cancers diagnostic the doctors selects and use distinctive forms of checking out techniques to come across most cancers and to check if the cancer has increase to different parts of the body out of doors the breast and the lymph nodes under the arm [11]. The preponderance of occurrences of irregularities within the breast is identified by way of screening using imaging technologies inclusive of mammography, magnetic resonance imaging , or ultrasound and biopsy.[11,12,13] The breast most cancers-detecting result is reported in any of the five general FNA cytology categories is inadequate pattern (not sufficient epithelial cells for analysis), Benign (now not cancer), unusual, atypical/uncertain but probable benign, Suspicious, and probably malignant (cancer) and Malignant (most cancers) for the unique patient. most cancers can be handled in many different methods, relying on the most cancers kind, its region, grade, and its metastasis level (whether or not it has spread or not). The maximum common forms of treatments for breast most cancers are as follows: surgery, Hormone remedy, Radiotherapy, Chemotherapy biological treatment options. The guide breast cancer detection method is accomplished by using a health practitioner/pathologist without the want of laptop or pc machine (which is in particular designed for slide microscopic digital photo analysis) the use of a mild microscope. Breast cancer detection and prediction approach is a the use of machine gaining knowledge of Algorithms specifically designed for detecting the breast cancer from the given at the Wisconsin Diagnostic Breast cancer data-set that's derived from a digitized photo of an magnetic resonance imaging. In a computerized breast cancer discovery device, no want for an experienced health practitioner or a crew of medical doctors for breast cancer-detecting result verification.

II Machine Learning Technique

Machine learning deals with generating algorithms and strategies that allow computers to routinely learn and make accurate forecasts primarily based on previous investigations. Machine learning is to obtain statistics from data-set automatically, with the aid of the use of computational and statistical ways. [14, 15] Machine learning gaining knowledge of algorithms may be prepared

into 3 agencies. Supervised, unsupervised, and semi-supervised machine learning algorithm. Supervised machine learning algorithms include decision tree, Support Vector Machines, naïve Bayes, random forest algorithm, and multi-layer preceptor and K-Nearest Neighbors algorithms and logistic regression etc the following Block diagram in determine explains the running principle of system learning set of rules., and multi-layer preceptor and k-Nearest associates algorithms and logistic regression and many others. the subsequent Block diagram in determine explains the running precept of machine mastering algorithm.

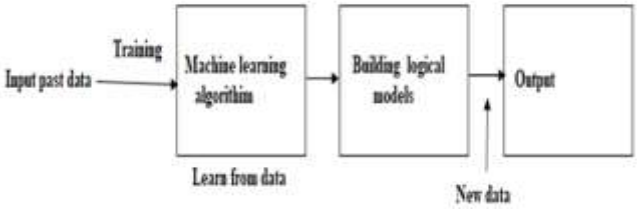

Figure 1: Working with Machine Learning Algorithm

Logistic regression

Logistic regression may be classification algorithms which help to allocate observation to a distance group of classes. In statistics this model is used to form the probability of certain class of interst.Few instances of classification problems are Tumor Malignant or Benign alive/dead or healthy/sick. But these regressions convert its output using the logistic sigmoid function to return a probability value. The output of this Regression is a categorical variable but for other regression models their output is value. Binary output can also be predicted from the independent variables.

The confusion matrix is employed to live the performance of two class problems for a given dataset. True positive (TP) and True negative (TN) means correctly classify instances also as false positive (FP) and False negative (FN) means incorrectly classify instances. Table 1 shows both correctly and incorrectly classified instances.

Here the matrices are of form
[TP FP]
[FN TN] where

Correctly Classified		Incorrectly Classified	
TP	TN	FP	FN
86	50	3	4

Table1: Logistic Regression Confusion Matrix

The confusion matrix of Logistic regression is somewhat better. There are only three observations that are Incorrectly Labeled as not cancer and four observations are misclassified as cancer and the experiment results show LR algorithm achieves 95 % of accuracy.

Decision Tree classifier

This Decision Tree classifier some another machine learning device used in categorizing problem to predict the class of example taken for the process moreover region and health issues. Decision tree is like structure and of the decision tree test the characteristic of the instance and each associate tree shows the out the attribute split. Leaf nodes stands for the class of instance formed in the model of the decision tree. The uppermost node in a decision tree is referred the root node. [17]

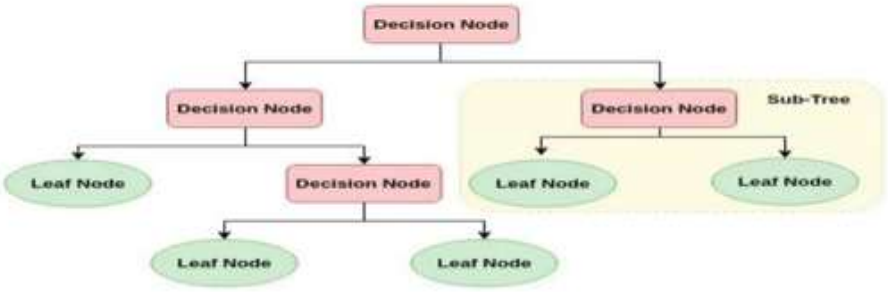

Figure 2: The DT Classifier Algorithm Diagram

The confusion matrix of decision tree Classifier is only two observations that are Incorrectly Classified as not cancer and seven observations are misclassified as cancer and the experiment results show DT algorithm achieves 93% accuracy.

Correctly Classified		Incorrectly Classified	
TP	TN	FP	FN
83	51	2	7

Table 2: DT Classifier Confusion Matrix

Random Forest

It is another machine learning algorithm which incorporate many individual decision tree .each decision tree votes the classification of a given data .Therefore, this algorithm acknowledges the classification that has obtained a maximum number of vote from a single tree .this a type of altogether learning algorithm that generates a number decision tree from the selected data set to forecast the output of sample data..[18]

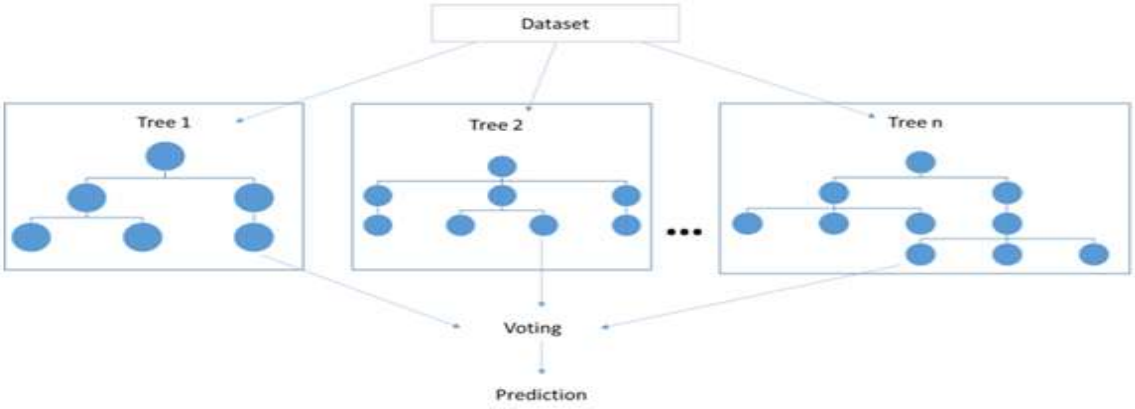

Figure 3: The Random Forest Algorithm Diagram

The confusion matrix of random forest algorithm is higher than other two algorithms. There are only two observations that are Incorrectly Labeled as not cancer and three observations are misclassified as cancer and the experiment results show RF algorithm achieves 96 % of accuracy.

.

Correctly Classified		Incorrectly Classified	
TP	TN	FP	FN
87	51	2	3

Table 3: Confusion Matrix For RF Classifiers

III PROPOSED METHODOLOGY

Dataset explanation

A dataset rows are 569 patient information and columns 33 features were downloaded from Kaggle. The data set row hold 569 rows and 33 columns .For detection and prediction purpose we use 'Diagnosis 'column and M or 0 refers for Malignant, 1 or B refers Benign.

Performance Metrics

Performance measures, such as recall (sensitivity), precision, and F-measure are also used for calculating other accumulated performance measures. To measure the performance of each method we will use different metrics such as:-

Accuracy is the dimension of examples that are labeled properly amongst the whole samples of the dataset. To calculate the accuracy we will use the following formula:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN}$$

Precision:

Precision is a measure for the optimistic predictive value and is given calculated by the formula as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall:

Recall is a measure for the true positive rate and the formula how to calculate is as follows:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Excessive precision refer a low false-positive rate and a excessive recall refer a low false-negative rate. Excessive precision and excessive recall suggest that you have correct consequences however when you have a high recall and low precision, then it means that most of the predicted values are false. If you have a low recall and high precision at the equal time, then it means that most of the predicted values are accurate. The satisfactory case for a model is when it has a high precision and a excessive recall. One way to summarize both metrics precision and recall is the f-score.

F-score

F-score is a consonant mean for both recall and precision is calculating as in the following:

$$\textbf{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

IV IMPLEMENTATION AND RESULT ANALYSIS

The implementation of this research work is done under a machine that has the following Specification details. Experiments and related analysis processes are done: Computer with Intel (R) Core (TM) i5- 4210U CPU ,Speed 2.4GHz ,6.00 GB RAM ,750 GB hard disk space and Windows 10 (Pro) installed. As a programming language Python 3.7 is used which an open-source, with variety is of free libraries, rich documentation, including contributor support. The supportive libraries and Software tools are listed next. Numpy, Matplotlib, Pandas, Seaborn and Scikit-learn. The experimentation is performed using three machine learning algorithms which are Random forest, Logistic Regression and Decision Tree algorithms. In machine learning, two types of data sets are used for training and, test purpose. Training data is used for building classification model and for measuring the performance of the model we will utilize test data.

Accuracy for Training Model

Training data model for accuracy is listed below:
Training Accuracy for Logistic Regression (LG): 0.99061032863849
Training Accuracy for Decision Tree Classifier (DT): 1.0
Training Accuracy for Random Forest Classifier (RF): 0.9953051643

So we can see that the Decision Tree Classifier has the best accuracy among the entire models ie.100%. In the below table we will test and check accuracy performance metrics of all Algorithms.

Performance Metrices	Algorithms with Percentage split		
	Logistic Regression	Decision Tree	Random Forest
Precision	95%	94%	97%
Recall	95%	94%	97%
f1-score	95%	94%	97%
Accuracy	95%	93%	96%

Table 4: Performance Metrics Of All Algorithm

V PREDICTION OF MODEL

The first data shows the actual result of which patient had cancer and the second data is the one predicted by the model. The accuracy of the model is 96.5% so we can see a few wrong predictions but mostly this model is successful in predicting a tumor Malignant (M) (harmful) or Benign (B) (not harmful) based upon the features provided in the data and the training. Thus supervised machine learning systems will be very significant in the early investigation and prediction of a cancer type in the scheme.

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]
```

Figure 4: Printed the Predictions

VI. CONCLUSION AND FUTURE WORKS

Machine learning systems have been extensively utilized in the pharmaceutical field and have worked as a helpful diagnostic instrument that assists physicians in examining the usable data as well as designing medical expert systems.

In general, the experimentation gives encouraging solution for the detection and prediction of breast cancer. It provides useful contribution to the field of breast cancer detection in health sectors. The evaluation results show that Random forest, algorithm performs 96% of accuracy and the other algorithms Logistic Regression and Decision Tree scores 95 % and 93% accuracy, respectively. The model with RF classification algorithm also attains better results in all performance evaluation metrics of precision, recall and F1measure to build and evaluate the models.

It has been observed that a good quality dataset provides better accuracy. The Selection of appropriate algorithms with good home dataset will lead to the increase accuracy of prediction systems and decrease error. These systems can help in proper treatment methods for a patient diagnosed with breast cancer. There are many remedies for a affected person based totally on breast most

cancers level device studying can be a very good help in finding out the line of treatment to be accompanied by using extracting information from such appropriate databases.

Future Works

This platform for the breast cancer detection and prediction provides many useful and interesting directions in this area. There may still be a gap to be improved on the breast cancer detection system using machine learning approaches, since the problem is a fatal Public health issue as stated on the problem statement. The main purpose of public health concern problem solving is to serve for performance enhancement to be done continuously till the highest accuracy level is attained. Accordingly, the following are some important for future work recommendations perceived while implementing this study work.

If more labeled data sets are found it will lead to better result

- We want to decrease the error rates with maximum accuracy rate
- the Grading level of cancer or identifying how quickly the cells are growing or increase
- Identifying the sub-types of breast cancer

REFERENCES

[1] S. D. Tzikopoulos, M. E. Mavroforakis, H. V.Georgiou, N. Dimitropoulos, S. eodoridis, A fully automated scheme for mammographic segmentation, and classification based on reast density and asymmetry, computer methods and programs in biomedicine, 102 (2011) 47-63

[2] M. Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N. Gurcan, "Computerized Classification of Intraductal Breast Lesions UsingHistopathological Images", IEEE Transactions on Biomedical Engineering, Vol. 58, No. 7, July 2011, pp. 1977-84.

[3] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., Molecular Biology of the Cell, Garland science, Taylor & Francis Group, LLC, 2008.

[4]. World Health organization "Cancer" retrieved from http://www.emro.who.int /health-topics/cancer/ index.html, last accessed on 28 May 2017.

[5]. National Cancer Institute, "Breast Cancer", retrieved from http://www.cancer.gov/ cancertopics /types/breast, last accessed on 29 June 2014.

[6]. Verywell, "Mammary Epithelial Cells: Function and Abnormalities" retrieved from https://www.verywell.com/ last accessed on 29 May 2017.

[7] Boyle, P., Levin, B. (Eds.), World Cancer Report 2008. Ann Oncol; 2008; 19(9): 1519-1521, "Need for global action for cancer control".

[8] Collaborative Group on Hormonal Factors in Breast Cancer, Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease, British Journal of cancer 87 (11) (2002) 1234–1245.

[9] Thune, I., Brenn, T., Lund, E., Gaard, M., Physical activity and the risk of breast cancer, New England Journal of Medicine 336 (18) (1997) 1269–1275.

[10] Collaborative Group on Hormonal Factors in Breast Cancer, Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease, Lancet 360 (9328) (2002) 187–195)

[11]. Cancer.Net, "Breast cancer diagnosis", retrieved from www.cancer.net, last accessed on 19 May 2017.

[12]. Cochrane Nordic, "Screening for breast cancer with mammography",retrivedfrom http://nordic.cochrane.org/screening-breast-cancer-mammographyy last accessed on 15 July 2017.

[13]. National Breast Cancer Centre Incorporating the Ovarian Cancer Program "Breast fine needle aspiration cytology and core biopsy: a guide for practice", First Edition Prepared by the National Breast Cancer Centre Funded by the Department of Health and Ageing. National Breast Cancer Centre 2004, ISBN Print: 174127036 7 Online: 174127 042 1, CIP: 618.190758.

[14]. NHS Cancer Screening Programs, "Guidelines for Non-Operative Diagnostic Procedures and Reporting in Breast Cancer Screening, Stream line Offset, Hoddesdon, Herts, Publication No 50, ISBN 1 871997 44 5, June 2001.

[15] Muhammad Jamil Moughal, "Which Machine Learning algorithm to use? Muhammad Jamil Moughal – Medium," 2018.

[16]"javatpoint.com/machine-learning-decision-tree-classification-algorithm"Feb 13th,2019.

[17] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," Ensemble Mach. Learn., no. February 2014, 2012, doi: 10.1007/978-1-4419-9326-7.

[18] Abhay Padda, "Introduction to Random Forest" March 2018.