American Journal of
Electronics & Communication

Sayani Ghosh et. al., American Journal of Electronics & Communication, Vol. II (1), 4-9

Automated Breast Cancer Diagnosis Based on Machine Learning Algorithm

Sayani Ghosh
Dept. of BCA, University of Engineering & Management, Kolkata, INDIA
Email Id: sayanighosh37@gmail.com

Sayantan Dey
Dept. of BCA, University of Engineering & Management, Kolkata, INDIA
Email Id: sayantandey42@gmail.com

Souvik Chatterjee
IEMA Research & Development Pvt. Ltd., Kolkata, INDIA
Email Id: souvik.chatterjee@iemcal.com

Abstract – Breast Cancer classification is becoming more important with the increasing demand of automated applications especially interactive applications. It can be used to improve the performance of classifiers like Logistic Regression, Decision Tree, Random Forest, SVC etc. This study is based on learning genetic patterns of patients with breast tumors and machine learning algorithms that aim to demonstrate a system to accurately differentiate between benign and malignant breast tumors. The aim of this study was to optimize different algorithm. In this context, we applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on accuracy, precision and the roc curves. The present report prepared by us proves that genetic programming can automatically find the best model by combining feature preprocessing methods and classifier algorithms by reducing False Positive rate. In this paper, there were two challenges to automate the breast cancer diagnosis: (i) determining which model best classifies the data and (ii) how to automatically design and adjust the parameters of the machine learning model. We have summarized the experimental studies and the obtained results, and lastly presented the main conclusion.

Keywords *- Breast cancer classifiers, Training and Testing, Machine learning, Accuracy*

I. INTRODUCTION

Breast Cancer, a very common type of cancer widespread among woman worldwide, and is the prevalent cause of death majorly. It develops in the breast cells. It is the leading cause of death among middle aged and older women. So to tackle this problem. Machine Learning plays a very important role. ML algorithms helps to determine whether the cells are Malignant or Benign .ML algorithms can determine cancer cells more efficiently.

Nowadays, the demand for machine learning is growing until it becomes a service in every aspects of life. Classification and data mining methods are an effective way to classify data in ML algorithm. In medical field, these methods are widely used in diagnosis and analysis to make decisions for better curing the diseases. In this paper, we performed a comparison between different machine learning algorithms is conducted to observe the accuracy rate. The foremost objective is to extend the correctness in data classification with reference to efficiency and effectiveness of every algorithm in terms of accuracy, precision and the roc curves. Moreover we used a large data set containing 32 features from 569 female patients to train and evaluate the system which provides greater reliability and accuracy.

Section 2 of this work describes related works in this domain. Methodology is discussed in section 3. Section 4 gives detailed explanation of the proposed method. Simulation results are shown and discussed in section 5. Paper concludes in section 6.

II. RELATED WORK

[1] Random forest classifier was implemented in their project to seek out sensitivity, time consumed and mean accuracy of two data set WBCPD and WBCDD. [2] In their paper, they have tested algorithms like C4.5, ANN, SVM to seek out

classification accuracy in carcinoma dataset. Their research shows SVM had produced higher accuracy in classification. [3] Adaboost algorithm was used to predict the cause and effect of breast cancer and the reason for death. Modest Adaboost algorithm was used. [4] Performance criterion of classifiers is compared by Vikas Chaurasia and Saurabh Pal for SVM with the RBF kernel, naïve bayes, rbf kernel in neural networks, simple cart and algorithm in decision trees in breast cancer dataset to seek out the simplest classifier. Their experimental results say, SVM-RBF kernel produces an accuracy of 96.84% which is above than other classifiers. The performance and efficiency of the algorithms such as SVM, Random Forest, Logistic Regression and Naïve bayes were compared to the similar works mentioned above. [5] According to Breiman, a single training and test partitions are not effective estimators of a classification error scheme on a limited dataset. Thus, it was decided that a random subsampling scheme should be used in this experiment to minimize any estimation bias. With the aim of preventing the overfitting, the cross-validation is a powerful concept against this problem. [6] Ultrasound characterisation of breast masses by S. Gokhale written by proposed a system where they found that doctors have known and experienced that breast cancer occurs when some breast cells begin to grow abnormally. In this study, they have used four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, ANN and Random Forest. [7] Another project by Pragya Chauhan and Amit Swami, which is based on the ensemble method usually used to increase the prediction accuracy of breast cancer. A Genetic algorithm based weighted average method that has crossover and mutation is taken for the prediction of multiple models. [8] Further more, a project by Abien Fred M. Agarap uses different methods like GRU-SVM, NN, multilayer perceptron (MLP), softmax regression to classify the dataset into benign or malignant. [10] A project by Priyanka Gupta shows the comparison of the lesser invasive techniques like Classification and Regression Trees (CART), random forest, nearest neighbour and boosted trees. These four classification models are chosen to extract the foremost accurate model for predicting cancer survivability rate. [11] Another project by Muhammet Fatih Aslan, Yunus Celik , and Kadir Sabanci, Akif Durdu that uses the blood analysis dataset from UCI. Their result shows that those are from methods like Extreme Learning Machine (ELM), ANN etc. it also has MATLAB GUI environment for classification with ANN. [12] Also a project by Yixuan Li and Zixuan Chen gives a performance evaluation using three indicators, i.e. prediction accuracy values, F-measure metric and AUC values are used to compare the performance of those five classification models. [13] A project by Mumine Kaya Keles, which is a comparative study of data mining classification algorithms. Another project by Sang Won Yoon and Haifeng Wang that uses four data mining models are applied in this paper, i.e., support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree. Furthermore, feature space is highly deliberated in this paper due to its high imapct on the efficiency and effectiveness of the learning process. [14] A project by Wenbin Yue and Zidong Wang that shows the algorithms that helped them with the diagnosis and prognosis of their dataset. [15] Cancer Prediction by the Priyanka Gandhi and Prof. Shalini L of VIT university, Vellore. In this paper, ML techniques are explored so as to spice up the accuracy of diagnosis. Methods such as CART, Random Forest, K-Nearest Neighbours are compared. The datasheet used is derived from UC Irvine Machine Learning Repository. It is found that KNN algorithm has much better performance than the other techniques used for comparison. [16] Detecting and Classifying Breast Cancer by different Machine Learning Algorithms using blood analysis data by Muhammet Fatih Aslan, Yunus Celik, Kadir Sabanci and Akif Urdu for carcinoma early diagnosis. During this paper, four different machine learning algorithms are used for the early detection of carcinoma. The objective of this project is to process the results of routine blood analysis with different ML methods. Methods used are ANN, ELM, SVM and K NN. [17] Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction by Yixuan Li and Zixuan Chen used two datasets in the study. The results of this study provide a reference for experts to distinguish the character of carcinoma. In this study, there are still some limitations that ought to be solved in further work.

III. METHODOLOGY

The main objective of this work was to classify and detect breast cancer in women from the raw breast cancer data by scaling the features and using several model classifiers to test the accuracy rate in each of them by determining their prediction tables. In order to fulfill our purpose, our proposed framework firstly extracted several features from the datasheet which shows the details of patients who may or may not have breast cancer.

A.*Feature Extraction*

The first objective of this work is to extract features from the dataset containing details of the tumor. A dataset containing details of 569 patients were used in the work. A total no. of 32 features were extracted from them. The features included radius, texture, perimeter, smoothness, concavity, compactness, area and symmetry.

B. *Visualization*

The count of total malignant and benign tumors was visualized using count plot and a pair plot. The **hue** parameter determines which column in the data frame should be used for color encoding to distinguish between benign and malignant tumors.

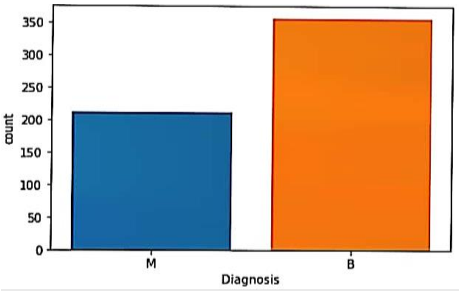


Fig. 1. Total count of Malignant and Benign tumors using count plot.

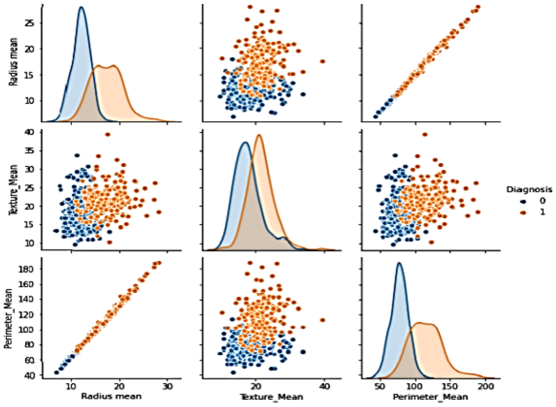


Fig. 2. Occurences of the features using a pairplot

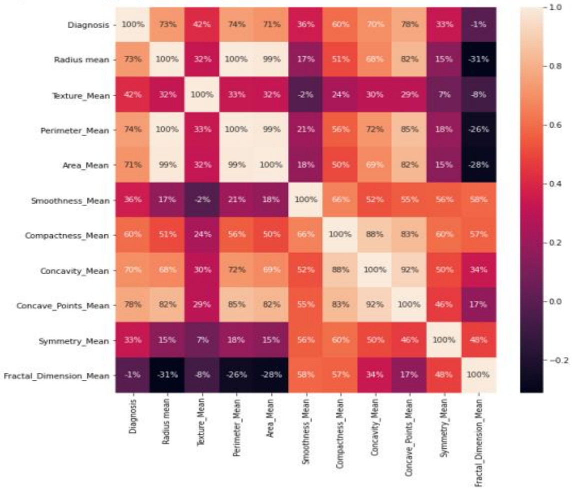


Fig. 3. Correlation among the 1st 11 features using heatmap

The correlation among the first 11 features was visualized from the dataset using heatmap which is a data visualization technique that shows magnitude of a phenomenon with different color intensities.

C. *Feature Scaling*

After splitting the data into 75% training and 25% testing, the unscaled features in the dataset were scaled using StandardScaler. Feature scaling through standardization (or Z-score normalization) is an important pre-processing step for many machine learning algorithms. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

D. Training and Testing

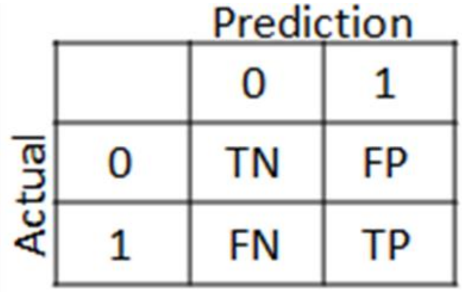
Algorithms were created using sklearn for training and testing of 7 model classifiers: *Logistic Regression Classifier*
K Nearest Neighbor Classifier
Support Vector Machine (Linear Classifier)
Support Vector Machine (RBF Classifier)
Gaussian Naive Bayes Classifier
Decision Tree Classifier
Random Forest Classifier

The features of the patients' diagnosis were obtained to train all the classification models to perform with atmost accuracy.

E. Accuracy Test using Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.
Condition positive (P): The number of real positive cases in the data.
Condition negative (N): The number of real negative cases in the data.

True positive (TP): Sensitivity (also called the **true positive rate**, the epidemiological/clinical sensitivity, the recall, or probability of detection in some fields) measures the proportion of actual positives that are correctly identified. TPR is calculated as TP/(TP+FN).
True negative (TN): Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified. TNR is calculated as TN/(TN+FP).
False positive (FP): The false positive rate is the proportion of the individuals with a known negative condition for which the test result is positive. FPR is calculated as FP/(TN+FP).
False negative (FN)
The false negative rate is the proportion of the individuals with a known positive condition for which the test result is negative. This rate is sometimes called the miss rate. FNR is calculated as 100 x FN / (TP+FN).



		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

IV. Proposed Method

Pipeline is the process of tying together some ordered final modules into one to build an automated machine learning workflow. It provides high-level abstraction of the machine learning process and significantly simplifies the complete workflow. Mostly, it is known as Extract, Transform, and Load (ETL) operations. In this work, many applied techniques were tested for the subsequent stages of processing and analysis of the breast cancer dataset.

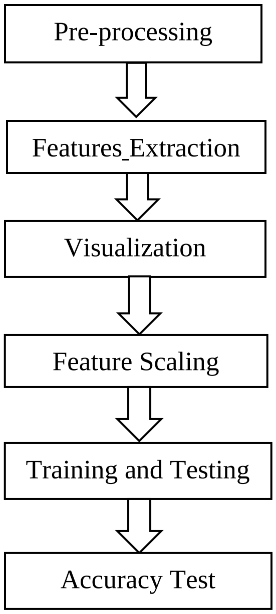


Fig. 4. Proposed Breast Cancer Classification system

After completing the training procedure the system was tested using a large and challenging dataset using different success metric such as Specificity and Accuracy were checked in order to represent evaluation of the success of the proposed system.

V. Results and Discussion

The proposed method has been implemented using Python on Google Colaboratory. After creating predictive model, efficiency can be checked. For this, the models can be compared based on their training and test accuracy rate.

Training Accuracy = $\frac{FP+FN}{TP+TN+FP+FN}\times 100.$

The other metrics derived from a confusion matrix are defined as follows:

$$\text{recall} = \frac{TP}{TP + FN},$$

$$\text{precision} = \frac{TP}{TP + FP},$$

$$F1 = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}.$$

TABLE I: Performance Analysis on the basis of training data

Sr.No.	Name of the classifiers	Accuracy rate
1	Logistic Regression Classifier	99.06%
2	Support Vector Machine (Linear Classifier)	98.82%
3	Support Vector Machine (RBF Classifier)	98.35%
4	Gaussian Naïve Bayes Classifier	95.07%
5	K Nearest Neighbour Classifier	97.65%
6	Decision Tree Classifier	100%
7	Random Forest Classifier	99.5%

TABLE II: Performance Analysis on the basis of the confusion matrix

Logistic Regression Classifier					**Decision Tree Classifier**				
TP	**FP**	**FN**	**TN**	**Accuracy**	**TP**	**FP**	**FN**	**TN**	**Accuracy**
49	4	86	4	94.44%	52	6	1	84	95.10%
Random Forest Classifier					**K N Neighbour Classifier**				
TP	**FP**	**FN**	**TN**	**Accuracy**	**TP**	**FP**	**FN**	**TN**	**Accuracy**
51	3	87	2	96.50%	48	1	5	89	95.80%
SVM (Linear Classifier)					**SVM (RBF Classifier)**				
TP	**FP**	**FN**	**TN**	**Accuracy**	**TP**	**FP**	**FN**	**TN**	**Accuracy**
51	2	87	3	96.50%	50	3	2	88	96.50%

Gaussian Naïve Bayes Classifier				
TP	**FP**	**FN**	**TN**	**Accuracy**
47	5	85	6	92.30%

TABLE III: Performance Analysis on the basis of Classification accuracy and other metrics

K Nearest Neighbour Classifier

	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

Accuracy Rate = 95.80%

Decision Tree Classifier

	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

Accuracy Rate = 95.10%

Random Forest Classifier

	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

Accuracy Rate = 96.50%

Support Vector Machine (Linear Classifier)

	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

Accuracy Rate = 96.50%

Support Vector Machine (RBF Classifier)

	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

Accuracy Rate = 96.50%

Gaussian Naïve Bayes Classifier

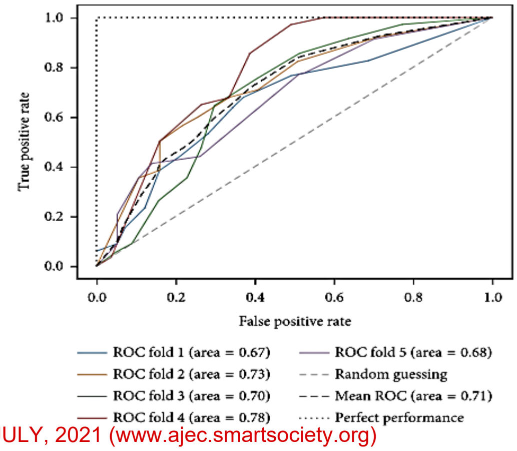
	Precision	Recall	F1-score	Support
0	0.93	0.94	0.94	90
1	0.90	0.89	0.90	53
Accuracy			0.92	143
Macro avg.	0.92	0.92	0.92	143
Weighted avg.	0.92	0.92	0.92	143

Accuracy Rate = 92.30%

From the accuracy and metrics above, the model that performed the best on the test data was the Random Forest Classifier with an accuracy score of about **96.5%.** Hence, this model can be chosen that model to detect cancer cells in patients.

A perfect classifier would fall into the top-left corner of the graph with a true positive rate of 1 and a false positive rate of 0. Based on the ROC curve, we can then compute the AUC to characterize the performance of a classification model. Thus, it is shown that applied models can predict more accurately.

Fig. 4. Random Forest Classifier ROC graph



VI. Conclusion

The importance of breast cancer classification is increasing with the advancement in technology and the extensive demand of new applications. The prime objective of this paper was to achieve the lowest error rate and best accuracy in analysing data of patients with breast tumors and detect whether it is malignant or benign. We trained the system with a large dataset to increase the accuracy of the system and also evaluated the system with a challenging data set in order to prove the robustness and reliability of the system. As a future recommendation we can say that the system, specially Random Forest Classifier could be implemented to detect cancerous cells. As a further extension, the idea behind this work could be used to automatically detect breast cancer faster and efficiently.

And again, this work can be further extended including more features using other classifier to see the changes in the results.

References

1. Poorkiani M, Hazrati M, Abbaszadeh A, Jafari P, Sadeghi M, Dejbakhsh T, Mohammadian Panah M.'s rehabilitation program to improve quality of life in breast cancer patients. Payesh. 2010;9(1):61–68.

2. Aghabarari M, Ahamadi F, Mohammadi E, Hajizadeh E, Farahania V. wrote about physical, emotional and social aspects of quality of life among breast cancer women under chemotherapy. Iranian Journal of Nursing Research. 2005;3:55–65. .

3. Hasanpoor Dehkordi A, Azari S. wrote about quality of life and related factor in cancer patients. Behbood. 2006;10(2):110–119. .

4. Saki A, Hajizadeh E, Tehranian N. wrote about evaluating the risk factors of breast cancer using the Analysis of Tree Models. Ofogh-e-Danesh. Journal of Gonabad University of Medical Sciences. 2011;17(2):60–69. .

5. Safaee A, Zeighami B, Tabatabaee HR, Moghimi Dehkordi B. Quality of life and Related Factors in Breast Cancer Patients under Chemotherapy. Iranian Journal of Epidemiology. 2008;3(4):61–66. .

6. McPherson K, Steel CM, Dixon JM. ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. BJM. 2000;321(7261):624–628. [PMC free article] . .

7. Harirchi I, Karbakhsh M, Kashefi A, Momtahen A. ewrote about breast cancer in Iran women: results of a multicenter study. Asian Pacific J Cancer Prev. 2004;5(1):24–27. . .

8. Hosseini M, Hassannejad R, Khademolghorani SH, Tabatabaeian M, Mokarian F. wrote about pattern identification of breast cancer metastasis among women of Iran, between 1999 and 2009 by Association Rules and Ordinal Logistic Regression. Scientific Research Journal of Health System Research (HSR) 2012;7(6):746– 762.

9. Lynch HT, Watson P, Conway TA. Clinical/ genetic features in hereditary breast cancer. Breast Cancer Res Treat. 1990;15:63–71.

10. Shishegar A. New breast cancer screening. Journal of Army University of Medical Sciences of TheI. R. Iran. 2011;9(1):58–66. .

11. Tabari F, Zakeri Moghadam M, Bahrani N, Monjamed Z. studied about the evaluation of the quality of life in newly recognized cancer patients. HAYAT. 2007;13(2):5–12. .

12. Shakeri J, Abdoli N, Paianda M, Chareh-Ga G. The frequency distribution of depression among patients with breast cancer in Kermaneshah u.m.s chemotherapy centers in 2007. Journal of MCIR of Irans. 2009;27(3):324–328. .

13. Safae A, Moghim-Dhkordi B, Zeighami B, Tabatabae HR, Porhosingholi MA wrote about predictors of quality of life in breast cancer patients under chemotherapy. Indian Journal of Cancer. 2008;45(3):107–111.

14. Zilich AJ, Blumenschin K, Johaneson M, Freeman P wrote about the the relationship between disease severity, quality of life, and willingness to pay in asthma. Pharmacoeconomics. 2002;20(4):257–265.

15. Mardani Hamule M, Shahraky Vahed A. wrote about the assessment of relationship between mental health and quality of life in cancer patients. Scientific Journal of Hamadan University of Medical Sciences. 2009;16(2):33–38.

16. Heravi Karimovi M, Pourdehqan M, Jadid Milani M, Foroutan SK, Aieen F. studied the effects of group counseling on quality of sexual life of patients with breast cancer under chemotherapy at Imam Khomeini Hospital. J Mazandaran Univ Med Sci. 2006;16(54):43– 51.

17. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin. 2009;59:225–249.

18. Sariego J. Breast cancer in the young patient. Am Surg. 2010;76(12):1397– 1400. . .

19. Steiner E, Klubert D. Assessing Breast Cancer Risk in Women. Am Fam Physician. 2008;78(12):1361–1366. . .

20. Yager JD, Davidson NE. Estrogen carcinogenesis in breast cancer. N Engl J Med. 2006;354(3):270–282 .

21. Venturi S. studied the role of iodine in breast diseases. 2001;10(5):379– 382. . .

22. Aceves C, Anguiano B, Delgado G. study of iodine and mammary glands. J Mammary Gland Biol Neoplasia. 2005;10(2):189–196.

23. Stoddard FR-II, Brooks AD, Eskin BA, Johannes GJ studied about iodine and alteration of gene expression in the MCF7 breast cancer cell line: evidence for an anti-estrogen effect of iodine. Int J Med Sci. 2008;5(4):189–196.

24. Labrecque LG, Barnes DM, Fentiman IS, Griffin BE. studied about different virus in epithelial cell tumors: a breast cancer study. Cancer Res. 1995;55(1):39–45.

25. Glaser SL, Hsu JL, Gulley ML. studied about virus and breast cancer: state of the evidence for viral carcinogenesis. Cancer Epidemiol Biomarkers Prev. 2004;13(5):688– 697.