International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842
Issue 10, Volume 07 (November 2020) https://www.irjcs.com/archives

PREDICTION OF BREAST CANCER USING MACHINE LEARNING

Shudipti Rani Mondal
Department of Computer Science and Engineering,
Chhattisgarh Swami Vivekanand Technical University, Bhilai, India
rakeshpal1051998@gmail.com
Aafreen
Department of Computer Science and Engineering,
Chhattisgarh Swami Vivekanand Technical University, Bhilai, India
aafreenaaf786@gmail.com
Rakesh Pal
Department of Computer Science and Engineering,
Chhattisgarh Swami Vivekanand Technical University, Bhilai, India
rakeshpal1051998@gmail.com

Publication History
Manuscript Reference No: IRJCS/RS/Vol.07/Issue10/NVCS10083
Received: 02, November 2020
Accepted: 17, November 2020
Published: 25, November 2020
DOI: https://doi.org/10.26562/irjcs.2020.v0710.002
Citation: Aishika saha (2020). Application of Threads. IRJCS:: International Research Journal of Computer Science, Volume VII, 276-280. https://doi.org/10.26562/irjcs.2020.v0710.002
Peer-review: Double-blind Peer-reviewed
Editor: Dr.A.Arul Lawrence Selvakumar, Chief Editor, IRJCS, AM Publications, India

Abstract: In order to support and supervise patients, the key detection and estimation of cancer type should establish a compulsion in the cancer research. Many research teams from the biomedical and bioinformatics fields have been advised to learn and evaluate the use of machine learning (ML) methods because of the relevance of classifying cancer patients into high or low risk clusters. To predict breast cancer, the logistic regression method and many classifiers have been proposed to generate profound predictions about breast cancer data in a new environment. This paper discusses the various approaches to data mining using classification to create deep predictions that can be applied to Breast Cancer data. In addition, by testing datasets on different classifiers, this analysis predicts the best model that delivers high efficiency. In this paper, the UCI machine learning repository has 699 instances with 11 attributes collected from the Breast cancer dataset. First, the data set is pre-processed, visualized and fed to different classifiers such as Logistic Regression, Support Vector Classifier, K-Nearest Neighbour, Decision Tree and Random Forest. 10-fold cross validation is implemented and testing is carried out in order to create and validate new models. Effective analysis shows that Logistic Regression generates the deep predictions of all classifiers and obtains the best model delivering strong and precise outcomes, followed by other methods: Support Vector Classifier, K-Nearest Neighbour, Decision Tree and Random Forest. Most models were less reliable compared to the approach of logistic regression.
Keywords: learning machine prediction cancer breast

I. INTRODUCTION

Cancer refers to the development of abnormal cells that divide uncontrollably and destroy normal body tissue. There are different types of cancer like lung cancer, kidney cancer, breast cancer, bladder cancer, colorectal cancer and many more. Among these, breast cancer is one of the widely spread disease in the world. Breast cancer is the abnormal growth of breast cells in women and rarely to men. The cause of breast cancer is multi-factorial. Several risk factors for breast cancer have been known nowadays. The risk factors are classified into non modifiable risk factors: age, sex, genetic factors (5-7%), family history of breast cancer, history of previous breast cancer and proliferative breast disease; modifiable risk factors: menstrual and reproductive factors, radiation exposure, hormone replacement therapy, alcohol and high fat diet; and environmental factors: organochlorine exposure, electromagnetic field and smoking [14]. This tumor can be classified as benign and malignant. Benign is non-cancerous, it does not invade nearby tissue or spread to other parts of the body but malignant is cancerous that can invade and kill nearby tissue and spread to other parts of the body. Breast cancer occurs when a malignant tumor (mass of tissue) occurs in the breast [15]. In this article, benign is denoted by 0 and malignant is denoted by 1.

It is very hard and time taking task for the doctors to diagnose breast cancer in a patient at commencing stage. But it becomes easy by using application of artificial intelligence, machine learning helps in prediction as well as detection of breast cancer effectively and accurately. Machine learning gives system the capacity to learn automatically and improve for experience. It uses different algorithms for prediction and computation of accuracy. By using effective models like logistic regression, SVC, KNN, decision tree and random forest which give high accuracy assist in prediction of breast cancer. The highly effective model is judged on 10 fold cross validation of testing. The validation is done on the bases of these parameters: accuracy, RMSE Error, sensitivity, specificity, F-Measure, ROC Curve Area and Kappa statistic and time taken to build the model [1].

II. LITERATURE SURVEY

In this part of article, we instigate previous research related to breast cancer detection. TABLE I shows a summary of literature survey.

TABLE I- SUMMARY OF LITERATURE SURVEY

Author	Year of Publication	Classifiers	Area of application/Disease	Accuracy achieved
Jabeen Sultana, Abdul Khader Jilani [1]	2018	Logistic Regression	Breast cancer	97.18%
Anji Reddy Vaka, Badal Soni, Sudheer Reddy K. [3]	2020	Deep Neural Network with Support Value	Breast cancer	97.21%
Vikas Chaurasia, Saurabh Pal, BB Tiwari [10]	2018	Naive Bayes	Breast cancer	97.36%
Dada Emmanuel Gbenga, Ngene Christopher, DaramolaComfortYetunde [5]	2017	Support Vector Classifier	Breast cancer	97.07%
David A. Omondiagbe, Shanmugam Veeramani, Amandeep S. Sidhu [6]	2019	Support Vector Machine	Breast cancer	98.82%
R. Chtihrakkannan, P. Kavitha, T. Mangayakarasi, R. Karthikeyan [7]	2019	Deep Neural Network	Breast cancer	96%
Farahnaz Sadough, Zahra Kazemy, Farahnaz Hamedan, Leila Owji, Meysam Rahmanikatigari, Tahere Talebi Azadboni [9]	2018	Support Vector Machine	Breast cancer	Ultrasound- 95.85%, Mammography- 93.069%, Thermography- 100%
Ricvan Dana Nindra, Teguh Aryandono, Lazuardi, Iwan Dwiprahasto [14]	2018	Support Vector Machine	Breast cancer	99.51%
Hiba Asri, Hajar Mousanni, Hassan Al Moatassime, Thomas Noel [8]	2016	Support Vector Machine	Breast cancer	97.13%
Aindrila Bhattacherjee, Payel Roy, Sourav Roy, Noreen Kausar, Sneha Paul, Nilanjan Dey [16]	2016	Back Propagation Neural Network	Breast cancer	99.27%

III. METHODOLOGY

A. Dataset

A dataset is an intrinsic requirement to produce a robust method for the detection of breast cancer. It is very difficult to collect dataset due to unavailability of sample and privilege of the patients. In this article, we have collected the dataset from UCI machine learning repository [1]. The data is retrieved from Wisconsin Breast Cancer Database, source - University of Wisconsin Hospital Madison, Wisconsin USA. The data contains 699 instances and 11 attributes (as of 15th July, 1992). The dataset contains 458- Benign and 241- Malignant, benign is denoted by 0 and malignant is denoted by 1. The data features are as follows: sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, class.

B. Machine Intelligence Libraries

The libraries used in this article are: matplotlib and seaborn for data visualization, pandas for data manipulation or analysis and numpy for numerical computation. To implement all machine learning algorithms scikit-learn is used.

C. Data Pre-processing

To make a viable analysis, the raw data is cleaned such that more than one machine learning algorithm is executed in one dataset to pick the right one out of them. By deleting comma, large space and unused attributes (sample code number is removed), the raw data is washed. If the dataset includes null values, the mean value of the row or column is replaced, or if several null values exist, the row or column is deleted. The dependent attribute is binarized, which transforms the values of the attribute type to binary values , so that it is possible for breast cancer diagnosis.

D. Data Visualization

Data visualization provides an important suite of tools for gaining a qualitative understanding. This helps in exploring and getting to know the dataset and helps in identifying patterns, corrupt data, outliers and much more. Data visualization is used to express and demonstrate key relationships in plots and charts. In this article, data visualization is used to know about patients having benign and malignant in the breast cancer dataset. Fig. 1 is a bar graph between class and count which shows that benign patients are more than malignant patients, there are 458 benign patients and 241 malignant patients. Fig. 2 is a scatter plot graph between class and features which shows that every patient is either benign or malignant; no other category is added in the dataset.

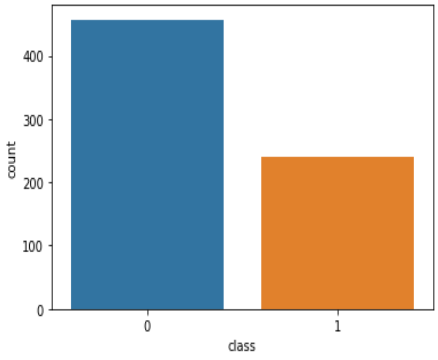

Fig. 1 Graph between class and count

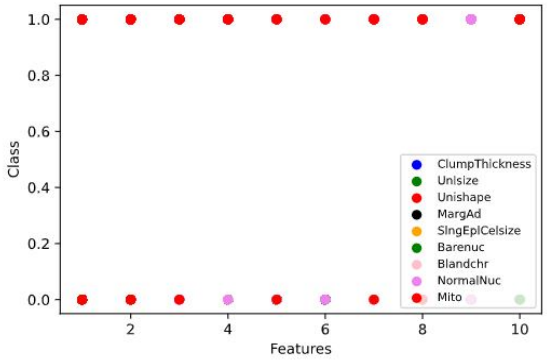

Fig. 2 Graph between class and features

E. Methods used

1) Logistic Regression: It is a supervised learning algorithm used to predict the probability of a target variable. The target variable or dependent variable should have two possible classes like in this article, there are two classes: Benign and Malignant.
2) Support Vector Classifier: The aim of SVC (Support Vector Classifier) is to respond to the information that we have, returning our data to a best match hyper plane that separates or categorizes it. From there we can then fed in any functionality to the classifier after getting the hyper plane to see what the predicted class is.
3) K-Nearest Neighbour: Also known as KNN, K-Nearest Neighbour is a supervised learning algorithm that can be used both for regression and classification issues. But in machine learning, KNN is widely used for classification problems. KNN operates under a theory that implies that each data point that falls next to each other is of the same class.
4) Decision Tree: One of the statistical modelling methods used in analytics, data processing and machine learning is decision tree learning. To go from assumptions about an object (represented by branches) to predictions about the target value of the item (represented by leaves), it uses a decision tree (as a statistical model).
5) Random Forest: Random forests or random decision forests are an ensemble learning tool for classification, regression and other tasks that function by creating a number of decision trees at training time and generating the class that is the class mode (classification) or mean/average predictor (regression) of the individual trees.

F. Parameters used

Cancer is extracted by analysing the results obtained and are evaluated by considering various parameters and are explained in detail here [1].

1) The percentage of test instances that are correctly classified on a given test set is determined as the accuracy of a classifier.

   Accuracy = (Number of correctly classified instances by rules ÷ Total number of instances by rules) * 100
2) Sensitivity and Specificity is calculated from Confusion Matrix obtained in the model.
3) Sensitivity is the proposition of the positive instances that are correctly identified

   TP = (TP/ TP +FN) * 100
4) Specificity is the proposition of the negative instances that are correctly identified

   TN = (TN/TN + FP) * 100

5) RMSE Root Mean Square Error is a measure of the difference between values predicted by a model and the values actually observed.
6) F-measure, ROC curve area and Kappa Statistics are also calculated using Confusion Matrix with the help of WEKA tool.

G. Train and Test:

It is a technique for evaluating the performance of a machine learning algorithm. It is the method of measuring the accuracy of the models. Here firstly, the dataset is split into two parts: features (except class) is stored in X and class is stored in Y. Then the data is divided to train and test: 80% of data is trained and 20% of data is tested. The 80% of data is fed to the machine learning models and accuracy is calculated using the 20% of the data. In this article, 80% of the data that is 558 instance with 9 attributes from X and 558 instances with 1 attribute from Y is fed to the models: logistic regression, support vector classifier, KNN, decision tree and random forest. Then it is tested to calculate the accuracy with 20% of data that is 140 instances and 9 attributes from X and 140 instances with 1 attribute from Y. Logistic Regression gave the highest accuracy among the models, with 99.2% accuracy. TABLE II shows the accuracy obtained by the models.

TABLE II - ACCURACY OF THE MODELS

Models	Accuracy
Logistic Regression	99.2%
Support Vector Classifier	97.8%
K-Nearest Neighbour	98.5%
Decision Tree	95%
Random Forest	98.5%

H. Confusion Matrix

A comparison is drawn between the actual class labels and the predicted class labels based on the class labels by the classifiers. The following describes the case when we deal with two-class classification problem. The generated confusion matrix is 2 * 2 matrixes [1]. Fig. 3 shows the confusion matrix of the logistic regression model.

TP	FN
FP	TN

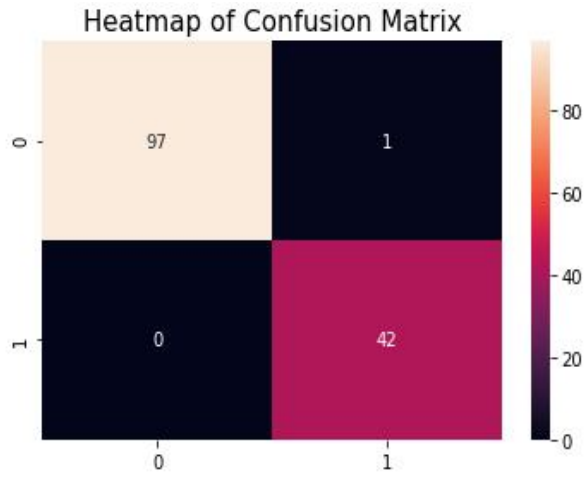

Fig. 3 Confusion Matrix of Logistic Regression

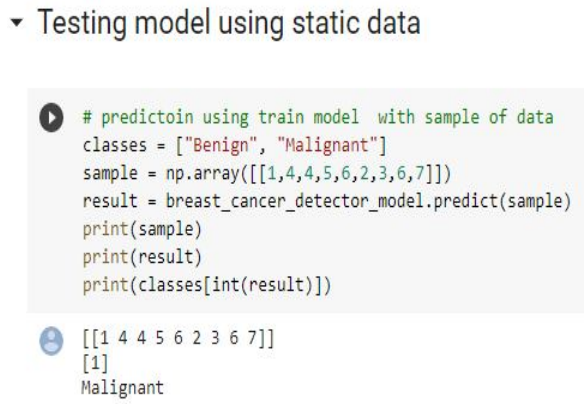

Fig. 4 Prediction of breast cancer using Logistic Regression

I. Testing the model

In this the highest accuracy giving model is tested using static data. To test the model firstly, the highest accuracy giving model that is logistic regression is stored in a variable (named breast_cancer_detector_model) using pickle and then it is used for prediction of cancer. Pickle is the standard way of serializing objects in a file and then it can be deserialized by loading the file. The variable is now used to predict the breast cancer using static data. Fig. 4 shows the prediction of breast cancer using logistic regression.

IV. CONCLUSION

Machine learning is an easy and effective way to predict the kind of tumor the patient is suffering from, as there is increase in number of women suffering from breast cancer. We need to predict the cancer class to which a patient will be classified by extracting the hidden information of different attributes that could be used to maximize performance in general by leveraging the best available tools.

In this article, comparing the accuracy of different models, namely: Logistic Regression, Support Vector Classifier, K-Nearest Neighbour, Decision Tree and Random Forest. The result concludes that Logistic Regression obtains the best model with 99.2% accuracy to predict the breast cancer. Logistic Regression gives best performance in comparison with other models in terms of parameters: accuracy, RMSE Error, sensitivity, specificity, F-Measure, ROC Curve Area, Kappa statistic and time taken to build the model [1]. TABLE III shows the parameters for Logistic Regression.

TABLE III- PARAMETERS FOR LOGISTIC REGRESSION

Method	Accuracy	RMSE	TP	FP	ROC	F-Measure	Kappa statistics	Time taken
Logistic Regression	99.2	0.1	0.97	0	0.97	0.99	0.95	0.64

REFERENCES

1. Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifier. Sultana,Jabeen and Jilani, Abdul Khader. 2018, International Journal of Engineering & Technology, pp. 22-26.
2. Using deep learning to enhance cancer diagnosis and classification. Fakoor, Rasool, et al. Arlington : s.n., 2013, Journal of Machine Learning Research, Vol. 28.
3. Breast cancer detection by leveraging Machine Learning. Vaka, Anji Reddy, Soni, Badal and K., Sudheer Reddy. s.l. : Elsevier B.V., 2020, pp. 1-5. ICT Express (2020).
4. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. M. Agarap, Abien Fred. Viet Nam : Association of Computing Machinery, 2018, ICMLSC.
5. Performance Comparision of Machine Learning Techniques for Breast Cancer Detection. Gbenga, Dada Emmanuel, Christopher, Ngene and Yetunde, Daramola Comfort. s.l. : Nova Explore, 2017, Nova Journal Of Engineering and Applied Science, pp. 1-8.
6. Machine Learning Classification Techniques for Breast Cancer Diagnosis. Omondiagbe, David A., Veeramani, Shanmugam and Sidhu, Amandeep S. s.l. : IOP Publishing, 2019, pp. 1-16.
7. Breast Cancer Detection Using Machine Learning. Chtihrakkannan, R., et al. 11, s.l. : Blue Eyes Intelligence Engineering & Science Publication, September 2019, International Journal of Innovative Technology and Exploring Engineering, Vol. 8, pp. 3123-3126.
8. Using Machine Learning Algorithm for Breast Cancer Risk Prediction and Diagnosis. Asri, Hiba, et al. s.l. : Elsevier B.V., 2016, Procedia Computer Science, pp. 1064-1069.
9. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Farahnaz Sadoughi, Zahra Kazemy, Farahnaz Hamedan, Leila Owji, Meysam Rahmanikatigari, Tahere Talebi Azadboni. s.l. : Dovepress, 2018, Breast Cancer- Targets and Therapy, pp. 219-230.
10. Prediction of benign and malignant breast cancer using datamining techniques. Vikas Chaurasia, Saurabh Pal, BB Tiwari. Jaunpur, UP, India : SAGE, 2018, Journal of Algorithms & Computational Technology, pp. 119-126.
11. A support vector machine-based ensemble algorithm for breast cancer diagnosis. Haifeng Wang, Bichen Zheng, Sang Won Yoon, Hoo Sang Ko. s.l. : Elsevier B.V., 2017, European Journal of Operational Research, pp. 687-699.
12. Breast Cancer Histopathological Image classification: Deep Learning Approach. Mehdi Habibzadeh Motlagh, Mahboobeh Jannesari, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, Iman Hajirasouliha. 2018, pp. 1-8.
13. A five-year (2015 to 2019) analysis of studies focused on breast cancer prediction using machine learning: A systematic review and bibliometric analysis. Zakia Salod, Yashik Singh. s.l. : pagepress, 2020, Journal of Public Health Research 2020, Vol. 9, pp. 65-75.
14. Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. Riecvan Dana Nindrea, Teguh Aryandono, Lazuardi, Iwan Dwiprahasto. 7, 2018, Asian Pacific Journal of Cancer Prevention, Vol. 19, pp. 1747-1752.
15. Application of Machine Learning Techniques to Predict Diagnostic Breast Cancer. Vikas Chaurasia, Saurabh Pal. 2020, SN Computer Science, pp. 1-11.
16. Classification Approach for Breast Cancer Detection Using Back Propagation Neural Network: A Study. Aindrila Bhattacherjee, Payel Roy, Sourav Roy, Noreen Kausar, Sneha Paul, Nilanjan Dey. 2016, IGI Global, pp. 210-221.

