Breast cancer Prediction using Machine Learning Techniques

Ashok Deulkar, Dr.J.A.Laxminarayana
Dept.of Computer Science & Engg.,Goa College Of Engineering, Farmagudi-Goa, India,
[2]Head of Department, Computer Science & Engg.,Goa College Of Engineering, Farmagudi-Goa, India.

***Abstract*:** There are many types of cancers that need our attention and a lot of human time spent in researching for their cure by analyzing a lot of symptoms. Many patients with similar health problems receive different kinds of treatment and eventually different extents of cure. Breast Cancer is one of the most exquisite and internecine disease among all of the diseases in medical science. It's mainly effective for women. It is one of the crucial reasons of death among the females all over the world. Various supervised machine learning techniques such as Logistic Regression,Decision tree Classifier,Random Forest ,K-NN,Support Vector Machine has been used for classification of data .The very famous data set such as Wisconsin breast cancer diagnosis (WBCD) data set has been used for classification of data. The experimental result shows that the Random Forest classifier gives the highest accuracy of 96.50% among the other classifier. The aim is for early detection of cancer because the early detection of cancer can be helpful to remove the cancer completely.

Keywords*: *Machine Learning, Classification, Decision tree Classifier, K-Nearest Neighbor, Logistic Regression, Random Forest,Support Vector Machine,accuracy etc.

I. Introduction

Machine learning (ML) is a significant method for data analysis that iteratively learns from the available data with the aid of learning algorithms. Machine learning can be defined as a technique for programming computers to optimize a performance criterion using example data or past experience i.e, defining a model up to some parameters, and learning by executing a computer program to optimize the parameters of a model using the training data or past experience. The model has been made predictive to make predictions in the future or descriptive to gather some knowledge/information from data. Machine learning is an artificial intelligence (AI) application that enables systems to learn and improve automatically without being explicitly programmed. Machine learning focuses on developing programs that are able to use the data and to learn from it.

Breast Cancer is the prime reason for death of women. It is the second most dangerous cancer after lung cancer**.** Breast cancer is the most common cancer among women worldwide accounting for 25 percent of all cancer cases. 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 [1]. Most of the experienced researchers come to conclusion that the most experienced physicians/oncologist can diagnose cancer with 79 percent accuracy while 91 percent correct diagnosis has been achieved using machine learning techniques. The procedure for early diagnosis of breast cancer must be accurate and reliable to distinguish between benign breast tumours from malignant ones.

II) Related Work

1) The problem of detection of breast cancer from the set of symptoms attracted many researchers worldwide. Ebrahim et. al [2], have proposed the experiments using Wisconsin Diagnosis Breast Cancer database to classify

the breast cancer as either benign or malignant. Supervised learning algorithm such as Support Vector Machine (SVM) with kernels like Linear and Neural Network (NN) are used for comparison to achieve this task. The implementation of Neural Network (NN) and Support Vector Machine (SVM) approach for classifying breast cancer as either benign or malignant was carried out. The performances of the models are analyzed where Neural Network approach provides more 'accuracy' and 'precision' as compared to Support Vector Machine in the classification of breast cancer and seems to be fast and efficient method**.** Neural Network technique is more efficient compared to SVM technique in breast cancer detection.

2) The work titled "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", proposed by Ch. Shravya et al.[3], is a relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) on the dataset. For the classification of benign and malignant tumor, they have used machine learning techniques in which they have learned from the past data and can predict the category of new input. The efficiency of each algorithm is measured and compared with respect to the accuracy, precision, sensitivity, specificity and False Positive Rate of the result. As the dataset contains 32 attributes, dimensionality reduction decreases the multi-dimensional data to a few dimensions. Out of the three applied algorithms Support Vector Machine, k Nearest Neighbor and Logistic Regression, they showed that SVM gives the highest accuracy of 92.7% compared to other two algorithms. They also proposed that the experiments have shown that SVM is the best for predictive analysis with an accuracy of 92.7% and on the whole KNN presented well next to SVM. Also further they conclude that SVM is the best suited algorithm for the prediction of Breast Cancer occurrence with complex datasets.

3) Usage of Three Machine Learning Techniques for Predicting Breast Cancer Recurrence was investigated by Ahmad l. LG, et. al. [4] who found that the number and size of medical databases are increasing rapidly but most of these data are not analyzed for finding the valuable and hidden knowledge. They proposed that data mining techniques can be used to discover hidden patterns and relationships and thus Models are developed from these techniques is useful for medical practitioners to make right decisions. Their present research experiments the application of data mining techniques to develop predictive models for breast cancer recurrence in patients that were followed-up for two years. They had used dataset that contained 1189 records, 22 predictor variables, and one outcome variable. Various machine learning techniques were used such as Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) for implementation to develop these predictive models. The goal of this paper is to compare the performance of these three well-known algorithms on data through sensitivity, specificity, and accuracy. The analysis shows that accuracy of DT, ANN and SVM are 0.936, 0.947 and 0.957 respectively. From this result they come to conclusion that the SVM classification model predicts breast cancer recurrence with least error rate and highest accuracy. The predicted accuracy of the DT model is the lowest of all. The results are achieved using 10-fold cross-validation for measuring the unbiased prediction accuracy of each model

4) In another related work Haifeng et.al. [5] Proposed a Breast Cancer Prediction model Using Data Mining Method to compare and identify an accurate technique to predict the incidence of breast cancer based on various patients' clinical records. To achieve this, they used Four data mining algorithms such as support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier and AdaBoost tree. Principal component analysis (PCA) has been used as a dimension reduction method to manifests advantages in terms of prediction accuracy and efficiency. Feature space is discussed in this paper due to its high influence on the efficiency and effectiveness on the learning process. To evaluate the performance of models, two

popular data sets are used i.e., Wisconsin Breast Cancer Database and Wisconsin Diagnostic Breast Cancer. The 10-fold cross-validation method is implemented to estimate the test error of each model and compared the performance of these models. Accuracy is calculated by ,

$Accuracy = TP + TN /(TP + TN + FP + FN)$

The results show that the SVM has the highest accuracy for WBC data and ANN has the best accuracy performance for WDBC data. As a future work, they proposed nonlinear feature reduction methods using k-means techniques. Also for raw data set such as SEER, created and validated by the authors, different data mining techniques can be used.

5) The work by Ayush Sharma et at[6] on Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis aims to predict breast cancer as benign or malignant using data set from Wisconsin Breast Cancer Data using sophisticated classifiers such as Logistic Regression, Nearest Neighbor, Support Vector Machines. The probability of recurrence in affected patients is calculated using the dataset. A concrete relationship between precision, recall and the number of features in the data set is determined . In this they had done the clinical examination to detect the tumor/lump in the breast by using imaging techniques such as Mammography. Then FNA on the lump detected. The FNA procedure provides with attributes such as tumor size, radius, area etc. as features for modelling the data into a classifier. Models such as SVM, k-nearest neighbor and Logistic Regression are trained using the data to make predictions.

6) The work on Comparison of Machine Learning Methods for Breast Cancer Diagnosis by Ebru Aydındag et al [7] presented a study on two most popular machine learning techniques that have been used for classification using Wisconsin Breast Cancer dataset. The performance of these techniques has been compared with each other using the values of accuracy, precision, recall and ROC Area. The algorithm such as Artificial Neural Network and Support Vector Machine are used as ML techniques for the classification of WBC dataset using WEKA tool.The result shows that Support Vector Machine technique has achieved the best performance with the highest accuracy. The effectiveness of applied ML techniques is compared in terms of key performance metrics such as accuracy, precision, recall and ROC area. Based on these performance metrics of the applied ML techniques, SVM has shown the best performance with accuracy of 96.9957 % for the diagnosis and prediction.

7) In another related work on “Breast Cancer Risk Prediction Using Data Mining Classification Techniques” by Peter Adebayo Idowu et. Al.[8] presented a study using two data mining techniques such as naïve bayes and the J48 decision trees to predict breast cancer risks in Nigerian patients. The performance of both classification techniques was evaluated in order to determine the most efficient and effective model. The result shows the J48 decision trees has higher values of accuracy, recall, precision and error rates compared to that of the naïve bayes .The evaluation criteria proved that the J48 decision trees to be a more effective and efficient classification techniques for the prediction of breast cancer risks among patients of the Nigeria.

8) The author Mandeep Rana, et al[9] presented a paper on “Breast cancer diagnosis and recurrence prediction using machine learning techniques” to classify whether the breast cancer is benign or malignant and to predict the recurrence and non-recurrence of malignant cases after a certain period. They had used machine learning techniques such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes.These techniques are coded in MATLAB using UCI machine learning depository. The accuracies of different techniques are compared and observed the results.SVM using Gaussian kernel is the most suited technique for recurrence/non-recurrence prediction of breast cancer and KNN performed best from overall methodology. In future work they had

proposed to solve the problem of multiclass variable the multiclass SVM and also SVM classes like LIBSVM will be used to achieve fine tuning of parameters used in algorithms for better accuracy.

III. METHODOLOGY

A. Dataset and Attributes

The experiment was carried out using a very famous breast cancer data set available from the University of Wisconsin Hospitals Madison Breast Cancer database [10].The dataset consist of 11 attributes for each sample. The instances were represented from attributes 2 to 10 respectively. There are total 699 cases, where some instances are omitted due to missing attributes. There is a class attribute in addition to 9 other attributes. Each instance has one of the 2 possibilities ie.Benign or malignant.The data set includes two classes, as mentioned earlier. They are benign (B) and malignant (M). After further analyzing the data we arrived at 30 attributes with 569 attributes[11].

TABLE 1. ATTRIBUTES OF THE WISCONSIN DIAGNOSTIC BREAST CANCER (WDBC) DATASET

Attribute	**Representation**	**Information Attribute**
ID Number	ID	Numerical
Diagnosis	Diagnosis	Nominal
Radius	radius_mean	Numerical
Texture	texture_mean	Numerical
Perimeter	perimeter_mean	Numerical
Area	area_mean	Numerical
Smoothness	smoothness_mean	Numerical
Compactnes s	compactness_mea n	Numerical
Concavity	concavity_mean	Numerical
Concave points	concave points_mean	Numerical
Symmetry	symmetry_mean	Numerical
Fractal dimension	fractal_dimension _m	Numerical
Radius	radius_se	Numerical
Texture	texture_se	Numerical
Perimeter	perimeter_se	Numerical
Area	area_se	Numerical
Smoothness	smoothness_se	Numerical
Compactnes s	compactness_se	Numerical
Concavity	concavity_se	Numerical
Concave points	concave points_se	Numerical

IV. EXPERIMENTS AND RESULT

In this section, we discuss the Breast Cancer dataset, Experiments and the evaluation scheme. In this we implement many algorithms for data mining clustering, classification, regression, and analysis of results.

The proposed architecture is shown in figure 1.

A. Experimental Setup

This Section describes the parameters and discusses the Result of the assessment of the implemented machine learning methods.

Accuracy: The accuracy of detection is measured as the percentage of correctly identified instances ie. the number of correct predictions divided by the total number of instances in the dataset. It should be noted that the accuracy is highly dependent on the threshold which was chosen by the classifier and may, therefore, vary between different sets of tests. Therefore,this is not the optimal method to compare different classifiers,but

it can give an overview of the class. Therefore, the accuracy can be calculated using the following equation:

$Accuracy = TP + TN /(TP + TN + FP + FN)$ Where: TP = True positive; FN= False negative; FP= False positive; TN = True negative. Similarly, P and N represent the Positive and Negative Population of Malignant and Benign cases, respectively.

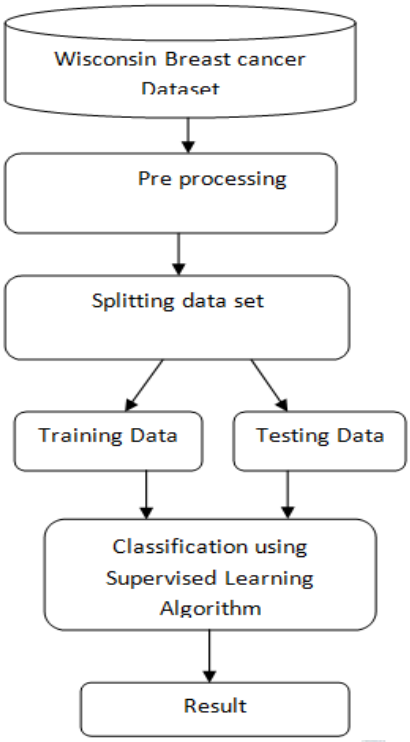

Fig.1. The Proposed architecture

Recall: Recall, also commonly known as sensitivity, is the rate of the positive observations that are correctly predicted as positive. This measure is desirable, especially in the medical field because how many of the observations are correctly diagnosed the sensitivity or the true positive rate

(TPR) is defined by: TP / (TP+FN)

while the specificity or the true negative rate (TNR) is defined by : TN / (TN + FP)

Precision: Percentage of correctly classified elements for a given class:

Precision = TP / (TP + TN).

B. Results

To evaluate the classifiers, we split the original data set to evaluate predictive models that in a training sample ie. 75% to form the model, and into a set of testing ie.25% to evaluate it. After applying the pre-treatment and preparation methods, we try to visually analyze the data and determine the distribution of values in terms of effectiveness and efficiency. We evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy.

TABLE 2: CLASSIFIER PERFORMANCE

Evaluation criteria	Classifier				
	Logistic Regression	Decision Tree	Random Forest	KNN	SVM
Correctly classified Instances	542	534	550	542	546
Incorrectly classified Instances	28	36	20	28	24
Accuracy (%)	95.10 %	93.70 %	96.50%	95.10 %	95.80 %
Precision	0.95	0.94	0.97	0.95	0.96
Recall	0.95	0.94	0.97	0.95	0.96
F1 Score	0.95	0.94	0.97	0.95	0.96

CONCLUSION

The experiments were performed on Wisconsin Diagnostics breast cancer dataset, taken from UCI Machine Learning repository. The datasets have20 features. Five Machine Learning algorithms are applied on the datasets. The prediction of cancerous elements is likely to be found in the dataset. The dataset is divided into two parts. One part is called training data which is 75% of total dataset, and rest remaining 25% is the testing data. The result shows that the random forest technique produced the accuracy of 96.50 % than the Support Vector Machine (SVM) technique which gave the accuracy of 95.80%.

Random Forest technique gives best accuracy of 96.50 % than other ML technique.

REFERENCES

[1] Gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf

[2] Ebrahim Edriss Ebrahim Ali, Wu Zhi , **"Breast Cancer Classification using Support Vector Machine and Neural Network"** March 2016, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[3] Ch. Shravya, K. Pravalika, Shaik Subhani "**Prediction of Breast Cancer Using Supervised Machine Learning Techniques**",April 2019, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6.

[4] Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR "**Using three machine learning techniques for Predicting breast cancer recurrence**", 2013, Journal of health & Medical informatics 4: 124 doi:10.4172/2157-7420.1000124.

[5] Haifeng Wang and Sang Won Yoon "**Breast Cancer Prediction Using Data MiningMethod**"2018, Department of Systems Science and Industrial Engineering,State University of New York at Binghamton.

[6] Ayush Sharma, Sudhanshu Kulshrestha, Sibi Daniel "**Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis**" 2017 IEEE, Department of computer Science,Jaypee inst. information technology U.P. India

[7] Ebru Aydındag Bayrak, Pınar Kırcı and Tolga Ensari "**Comparison of Machine Learning Methods for Breast Cancer Diagnosis**", 2019 IEEE, 978-1-7281-1013-4/19.

[8] Peter Adebayo Idowu,Kehinde Oladipo Williams, Jeremiah Ademola **Prediction Using Data Mining Classification Techniques**",April 2015,Transactions on Networks and Communications, Volume 3 No 2.

[9] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, Nikahat Kazi "**Breast cancer diagnosis and recurrence prediction using machine learning techniques**", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | PISSN: 2321-7308.

[10] Archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[11] M. Lichman, "**UCI Machine Learning Repositry [Online]**," Available: https://archive.ics.uci.edu/, 2013

[12] Youness Khourdifi , Mohamed Bahaj "**Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification**", 978-1-5386-4225-2/18/$31.00 ©2018 IEEE