BREAST CANCER DIAGNOSIS AND RECURRENCE PREDICTION USING MACHINE LEARNING TECHNIQUES

Mandeep Rana[1], Pooja Chandorkar[2], Alishiba Dsouza[3], Nikahat Kazi[4]

[1]*Student, FRCRCE, Mumbai University*
[2]*Student, FRCRCE, Mumbai University*
[3]*Student, FRCRCE, Mumbai University*
[4]*Assistant Professor, FRCRCE, Mumbai University*

Abstract

Breast Cancer has become the common cause of death among women. Due to long hours invested in manual diagnosis and lesser diagnostic system available emphasize the development of automated diagnosis for early diagnosis of the disease. Our aim is to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period. To achieve this we have 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. We have compared the accuracies of different techniques and observed the results. We found SVM most suited for predictive analysis and KNN performed best for our overall methodology.

Keywords: *Breast Cancer, SVM, KNN, Naive Bayes, Logistic Regression, Classification.*

----------***----------

1. INTRODUCTION

Breast Cancer is the most common type of cancer World Wide and is the leading cause of death among women. The most effective way to reduce breast cancer deaths is by detecting it earlier. This is possible by performing various tests like MRI, mammogram, ultrasound and biopsy. Breast Cancer refers to uncontrolled growth of cells in the breast tissue. If these cells are not stopped or controlled then it might cause an adverse effect on the whole body. Breast cancer can occur in men too having a higher mortality rate [1].

Diagnosis of breast cancer is done by classifying the tumor. Tumors can be either *benign* or *malignant* but only latter is the cancer. Malignant tumors are more cancerous than the benign. Unfortunately, not all physicians are expert in distinguishing between the benign and malignant tumors. So we need a proper and reliable diagnostic system that can detect the malignant tumor [1, 2]. We also need a system that could predict the recurrence and non-recurrence of breast cancer from the malignant cases after the patient has undergone treatment for certain time period [3].The data available in manual diagnosis is noisy and raw that increases the cost of management of data. Thus there is a need of proper parameter and feature selection so that the error rate and cost is minimised.

In this paper, *SVM* (SMO) model using linear and Gaussian kernels for separable and non-separable data are proposed respectively. We have implemented *K- nearest neighbour* where similarity criteria are Euclidean distance and Manhattan distance. *Naïve Bayes* and *logistic regression* are also implemented. Regularization parameter lambda is implemented in logistic regression for solving the problem of overfitting of data. All these techniques are performed on *UCI depository* (WDBC and WPBC). Based on the observed accuracy values certain conclusions are made.

2. PREVIOUS WORK

The increase in health problems especially breast cancer has instigated many researchers to make further developments in finding the most reliable and efficient diagnostic system. *Xiaowei Song* [4] used various machine learning techniques and obtained good mortality rate. Strong and sophisticated algorithm called least square SVM for predictive analysis has been implemented. Logistic regression based on sigmoid function was also proposed. All the techniques were proposed on four datasets under ROC curve (AUC).

Tarigopulla V.S Sriram [5] proposed SVM, KNN and Naïve Bayes techniques on Parkinson disease dataset obtained from UCI depository and made conclusions based on different values of accuracy.

Work by *S. Kharya* [6] states that artificial neural network have been the most widely used predictive technique in medical prediction, though its structure is difficult to understand. The paper lists out the benefits and limitations among various machine learning techniques such as Decision trees, Naïve Bayes, neural network and SVM.

H. Yusuff [7] proposed logistic regression model for breast cancer analysis, where he worked on the observed as well as the validated mammogram samples that were collected through survey. We are proposing different machine learning algorithms for benign/malignant classification and recurrence/non-recurrence prediction. The algorithms implemented include: SVM (SMO) – linear and RBF,

Logistic regression and Logistic regression with regularization parameter, KNN – Euclidean and Manhattan measures, Naive Bayes. All these algorithms perform in a different manner, thereby giving different values of correct classification and prediction on both the datasets.

3. MATERIALS AND METHODS

Materials that we have used include; MATLAB software for coding and breast cancer data from UCI depository. Our methodology involves use of machine learning techniques such as; SVM, KNN, logistic regression and Naïve Bayes.

3.1 Data

We have used two datasets from UCI depository [8] – one for diagnosis (WDBC) and the other for prediction (WPBC).

WDBC consists of 32 attributes (ID, diagnosis, 30 real-valued input features).
Attribute Information (WDBC):
1) ID number
2) Diagnosis (M = malignant, B = benign)
(3-32) ten real-valued features are computed for each cell nucleus:
a) Radius (mean of distances from center to points on the perimeter)
b) Texture (standard deviation of gray-scale values)
c) Perimeter
d) Area
e) Smoothness (local variation in radius lengths)
f) Compactness (perimeter^2 / area - 1.0)
g) Concavity (severity of concave portions of the contour)
h) Concave points (number of concave portions of the contour)
i) Symmetry
j) Fractal dimension ("coastline approximation" -1)

WPBC consists of 34 attributes (ID, outcome, 32 real-valued input features)
Attribute information (WPBC):
1) ID number
2) Outcome (R = recur, N = nonrecur)
3) Time (recurrence time if field 2 = R, disease-free time if field 2 = N)
(4-33) Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)
34) Tumor size - diameter of the excised tumor in centimeters
35) Lymph node status - number of positive axillary lymph nodes

3.2 Block Diagram

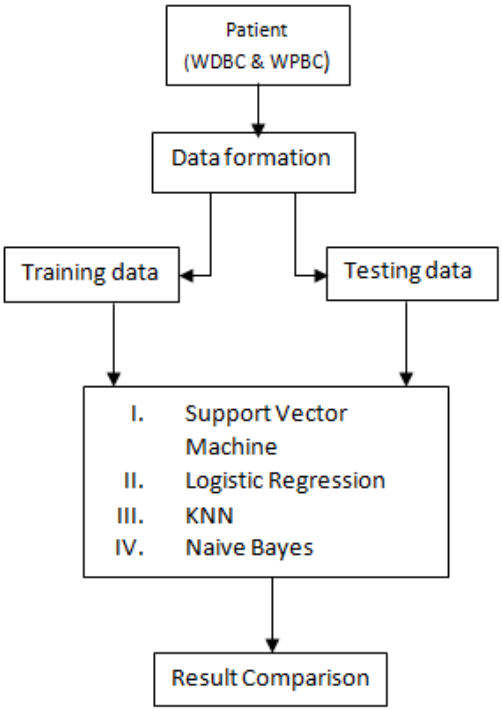


Fig 1 Block diagram of overall methodology

3.3 Support Vector Machine

Support vector machine is a very strong and sophisticated machine learning algorithm especially when it comes to predictive analysis. We have studied and implemented SVM using two kernels: *linear* and *Gaussian*. When it comes to classifying separable dataset we prefer linear kernel (as shown in figure 2) whereas for non-linear dataset classification we opt for kernel selection such as *Gaussian* (as shown in figure 3), *polynomial* [9].

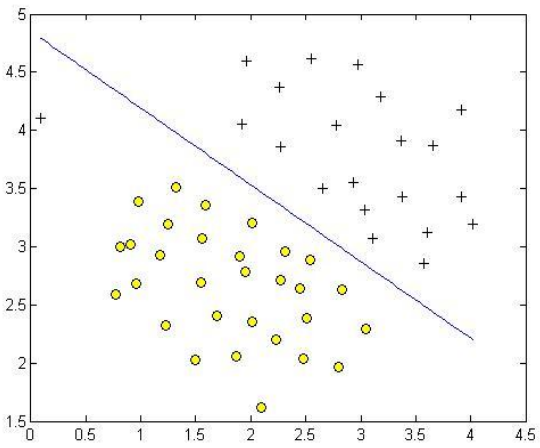


Fig 2 Linear SVM hyperplane construction

SVM focusses in determining the *hyperplane* such that it divides the region into two classes [10].

$$Hyperplane\ eq\ (y) = w * X' + b$$

This equation is dependent on weight vector (w), bias element (b) and the support vectors (X). Vectors that lie closest to the hyperplane are *support vectors.* Support vectors are responsible for determining the hyperplane.

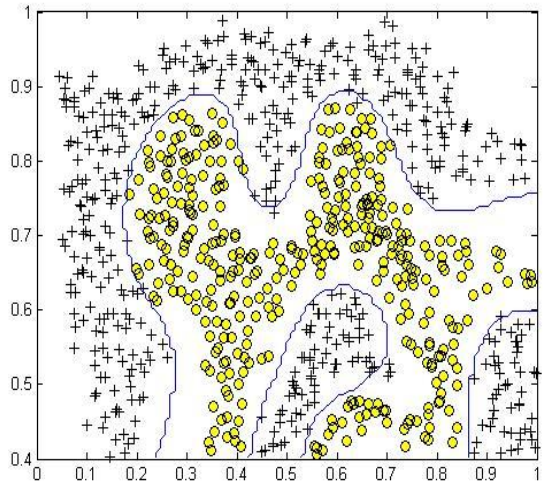


Fig 3 Gaussian SVM hyperplane construction

3.4 Logistic Regression

Logistic regression can be binomial or multinomial. Binomial or binary logistic regression can have only two possible outcomes: for example, "dead" vs. "alive". The outcome is usually coded as "0" or "1", as this leads to the most straightforward interpretation. If possible outcome is success then it is coded as "1" and the contrary outcome referred as a failure is coded as "0". Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase. Logistic regression technique uses *sigmoid* function to carry out the classification [11].

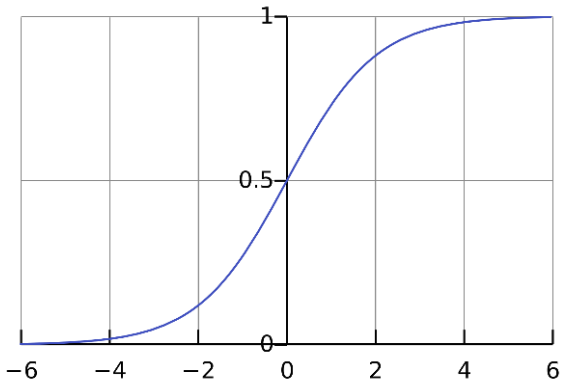


Fig 2 Sigmoid Function graph

We have also implemented the logistic regression technique using the *regularization* parameter 'lambda' that solves the *overfitting* problem thereby giving us better results than the generalised logistic regression technique.

3.5 KNN

The *k-Nearest Neighbour algorithm* is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The KNN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, it can be useful to weight the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbour a weight of 1/*d*, where *d* is the distance to the neighbour [12]. We have considered *Euclidean* and *Manhattan* distance measures to assign weights and determine the neighbours [13].

If we consider two point's x_i and y_i in n dimensional space then:

$$Euclidean(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$Manhattan(x, y) = \sum_{i=1}^{n}|x_i - y_i|$$

3.6 Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple *probabilistic* classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. *Naive Bayes* is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [14].
Using Bayes' theorem, the conditional probability can be decomposed as:

$$P(C_k \mid x) = \frac{p(Ck)\,.p(x|Ck)}{p(x)}$$

Here we have implemented the classifier by considering the *normal* distribution and the *kernel* distribution.

4. RESULTS AND DISCUSSIONS

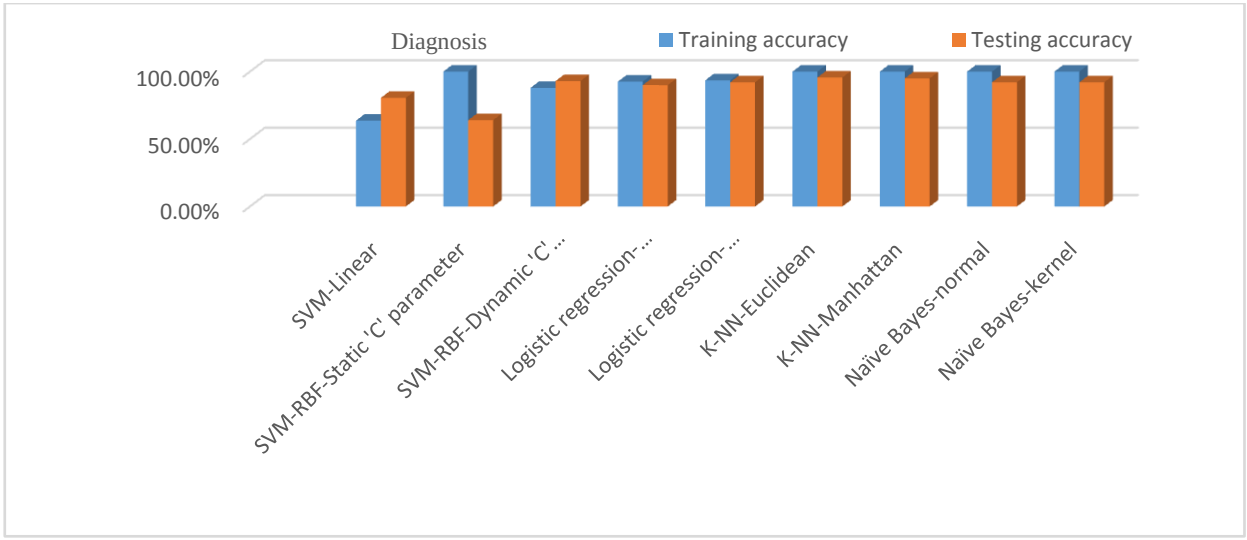
The correctly classified data for diagnosis and prognosis of breast cancer has been observed and its accuracy is calculated as shown below:

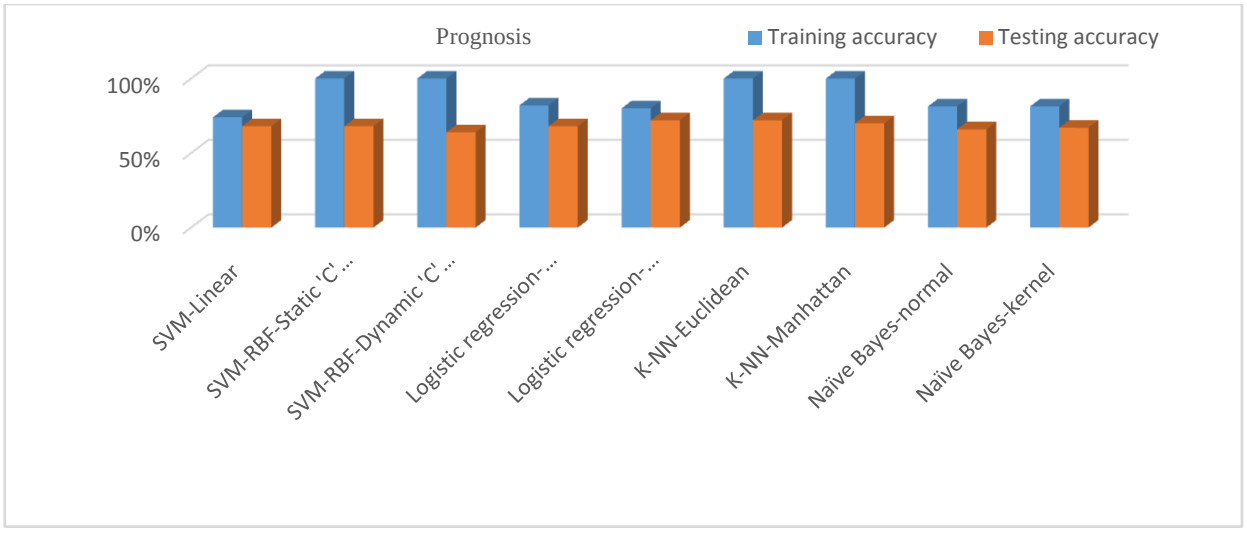
Results for diagnosis of breast cancer:

Machine learning techniques	Training accuracy (%)	Test accuracy (%)
SVM-linear	63.64	80.58
SVM-RBF-static 'C' parameter	100	64.03
SVM-RBF-dynamic 'C' parameter	88	93
Logistic regression-generalized	92.59	90
Logistic regression-regularized	93.54	92.08
k-NN-Euclidean	100	95.68
k-NN-Manhattan	100	94.96
Naïve Bayes-normal	100	92.1
Naïve Bayes-kernel	100	92.1

Results for predicting the recurrence and non-recurrence of breast cancer:

Machine learning techniques	Training accuracy (%)	Test accuracy (%)
SVM-linear	74	68
SVM-RBF-static 'C' parameter	100	68
SVM-RBF-dynamic 'C' parameter	100	64
Logistic regression-generalized	82	68
Logistic regression-regularized	80	72
k-NN-Euclidean	100	72
k-NN-Manhattan	100	70
Naïve Bayes-normal	81.33	66
Naïve Bayes-kernel	81.33	67





For both diagnosis and prognosis we observe that proper parameter selection plays an important role in correct classification. Gaussian SVM gives best results when the value of sigma is small and parameter C is large. On the other hand, the regularized logistic regression gives better performance than the generalized logistic regression. This is because of the regularization parameter "*lambda*". The performance of logistic regression increases as we go on decreasing the lambda value. But at certain value of lambda (0.0001 as in our case) the performance obtained is the best. The performance of SVM (linear and Gaussian both) increases as we increase the number of iterations of classifying the dataset but it causes increase in the training time.

5. CONCLUSION

In this paper, each algorithm performs in a different way depending on the dataset and the parameter selection. For overall methodology, KNN technique has given the best results. Naive Bayes and logistic regression have also performed well in diagnosis of breast cancer. As said earlier, SVM is a strong technique for predictive analysis (in section 3.3) and owing to the above finding, we conclude that SVM using Gaussian kernel is the most suited technique for recurrence/non-recurrence prediction of breast cancer.

FUTURE SCOPE

The SVM (SMO) that is used in the analysis in this paper is only applicable when the number of class variable is binary i.e. we can't have more than 2 classes. To solve this problem scientists have come up with multiclass SVM. Further research in this domain such as the creation of SVM classes like *LIBSVM* [15] has taken place. Further fine tuning of parameters used in algorithms can result in better accuracy. This field has further scope of research.

REFERENCES

[1]. National Breast Cancer Foundation Inc., http://www.nationalbreastcancer.org/about-breast-cancer.

[2]. T. Subashini, V. Ramalingam, and S. Palanivel, ―Breast mass classification based on cytological patterns using RBFNN and SVM, Expert Systems with Applications, 2009. 36(3): p. 5284-5290.

[3]. Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR - Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, J Health Med Inform 2013,4:2,http://dx.doi.org/10.4172/2157-7420.1000124.

[4]. Xiaowei Songa, Arnold Mitnitskib,c, Jafna Coxb, Kenneth Rockwood - Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes, MEDINFO 2004 M. Fieschi et al. (Eds) Amsterdam: IOS Press © 2004 IMIA.

[5]. Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, T Pandu Ranga Vital - Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms, International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013.

[6]. S.Kharya, D. Dubey, and S. Soni - Predictive Machine Learning Techniques for
Breast Cancer Detection, IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023-1028.

[7]. H. Yusuff, N. Mohamad, U.K. Ngah & A.S. Yahaya – Breast Cancer Analysis Using Logistic Regression, www.arpapress.com/Volumes/Vol10Issue1/IJRRAS_10_1_02.pdf.

[8]. UCI depository - http://archive.ics.uci.edu/ml/

[9]. Colin Campbell, Nello Cristianini - Simple Learning Algorithms for Training.

[10]. R. Berwick - An Idiot's guide to Support vector machines (SVMs)

[11]. Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll - An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, September/October 2002 [Vol. 96(No. 1)].

[12]. Abraham Karplus - Machine Learning Algorithms for Cancer Diagnosis, Santa Cruz County Science Fair 2012.

[13]. Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou - Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules, International Journal of Computer Applications (0975 - 8887) Volume 62 - No. 1, January 2013.

[14]. XindongWu · Vipin Kumar · J. Ross Quinlan - Top 10 algorithms in data mining, Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007

[15]. LIBSVM: A library for SVM , Chih Chung chang , National Taiwan University.