The Machine Learning based Optimized Prediction Method for Breast Cancer Detection

Nirdosh Kumar
M.Tech
MNIT, Jaipur
Rajasthan, India- 302017
Email: 2017peb5169@mnit.ac.in

Gaurav Sharma
Research Scholar
MNIT, Jaipur
Rajasthan, India- 302017
Email: 2015rec9014@mnit.ac.in

Lava Bhargava
Professor
MNIT, Jaipur
Rajasthan, India- 302017
Email: lavab@mnit.ac.in

Abstract—**Breast Cancer is the most prevalent form of cancer and significant reason for high mortality rates among women. Manual diagnosis of this disease requires long hours & specialists. Therefore an Automated breast cancer diagnosis has been developed to reduce the time taken for diagnosis and decreases the spread of cancer. This paper presents a comparative study of four machine learning algorithms namely Logistic Regression, SVM, KNN and Naive Bayes by calculating their classification accuracy, sensitivity, specificity and other parameters. The different hyper-parameters used for different ML algorithms were manually assigned. Among all algorithms, SVM performed better with the accuracy of about 98.24%.**

Index Terms—**Logistic regression, support vector machine, k-nearest neighbours, naive Bayes, Wisconsin diagnostic dataset**

I. Introduction

Breast cancer is a common cancer in India and accounts for about 27% of all cancers in women. India accounts for the 3rd highest number of cancer among women [1] after the US and China. Statistically, 1 out of 28 women is likely to suffer from breast cancer in their life span [2]. The factors [3] which cause Breast cancer in women are obesity, hormone replacement therapy during menopause, family history of breast cancer, lack of physical exercise, prolonged exposure to electromagnetic radiation, having children at a later age or not at all and early age at first menstruation etc. 2000 new women are diagnosed with cancer every day, and 1200 are detected at the later stages out of these 2000 women. Late detection [4] reduces the survival rate by 3 to 17 times and costs 1.5 to 2 times higher as compared to early-stage detection. The mortality rate because of breast cancer is 1.6 to 1.7 times higher. In 2017, India had the highest mortality rates globally for breast cancer. Early-stage diagnosis of breast cancer can help significantly in increasing the survival rate of patients . Machine learning algorithms can play a vital role in the diagnosis of breast cancer [5].

A. *Diagnosis using imaging techniques*

Mammography, ultrasound and MRI are imaging techniques [6] which are used for diagnosis of breast cancer. All these techniques characterise the breast cell present in the image using different features like cell size, the texture of cell etc. Based on these findings, we can predict whether someone has breast cancer or not.

B. *Machine Learning Algorithms*

We can diagnose breast cancer by using all these features found using Imaging Tests with the assistance of different machine learning algorithms. In this paper, we will be using four different machine learning algorithms, namely: Logistic regression [7], SVM, KNN and Naive Bayes for the diagnosis of breast cancer.

II. Tools and Dataset

A. *Tools & Libraries*

Anaconda navigator was used for implementing the machine learning algorithms in the Python programming language in this study along with the libraries such as Numpy, Pandas, Scikit-Learn and Matplotlib.

B. *Dataset*

We are using Wisconsin Breast Cancer Data Set [8] downloaded from the UCI repository. It includes 569 no of samples. Out of these 569 samples, 357 are benign & rest 212 are malignant as shown in Figure 8.

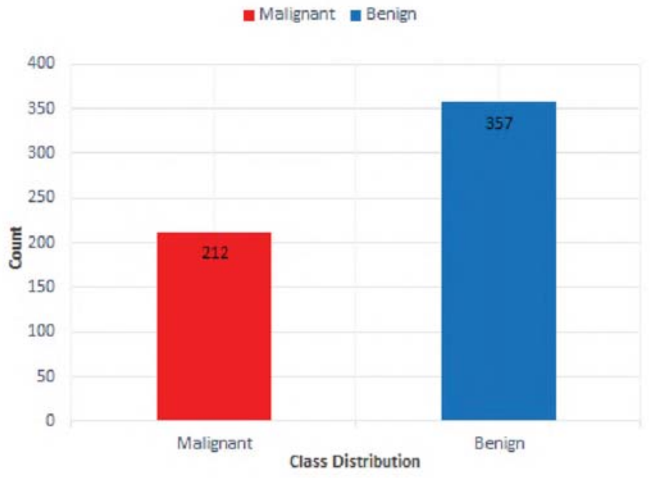

Fig. 1. Wisconsin breast cancer dataset.

For the nucleus of each cell, 10 real-valued features [9] have been computed:

i) Radius - average distance between the centre and perimeter
ii) Perimeter
iii) Texture - standard deviation in grey-scale values
iv) Area
v) Compactness - calculated using the following formula given in Equation 1.

$$compactness = (\frac{perimeter^2}{area} - 1)g \quad (1)$$

vi) Concave points - number of concave portions of the contour
vii) Smoothness - local variation in the length of the radius
viii) Symmetry
ix) Concavity - The intensity of concave portions of the contour
x) Fractal dimension - ratio providing a statistical index of complexity of the pattern.
For all these 10 features of breast cell, the mean, standard error and worst values are calculated, resulting in overall 30 features. All these features are computed to four significant digits after the decimal point.

III. Methodology

Figure 2 shows the flowchart of the breast cancer diagnosis model using supervised machine learning. In supervised machine learning algorithms, a machine is trained using labelled data, such as an input where the desired output is known and based on this new data is classified. On processing the training data, the algorithm generates a mapping function which predicts the output for the new data after adequate training. In our work, the data is labelled as either malignant or benign. There are different methods such as regression, classification, gradient boosting and others which are used for predicting the output in supervised machine learning.

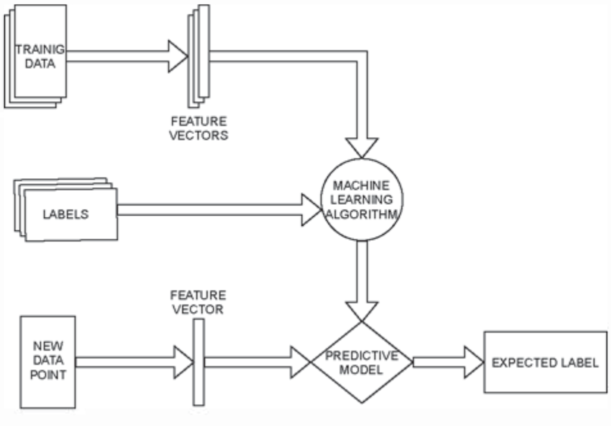

Fig. 2. Flowchart of diagnosis of breast cancer using machine learning.

We have employed four machine learning algorithms, namely: Logistic Regression, KNN, SVM [10], & naive Bayes. A brief description, along with their mathematical representation, is given below:

A. Logistic Regression

Logistic regression is a type of regression model. It predicts the dependent variable (result) which is categorical in nature, i.e. 0/1, pass/fail, yes/no etc., after finding a relation between the dependent variable and the given independent variables. Logistic regression uses the sigmoid function [11] given in Equation 2

$$y = \frac{1}{1 + e^{-x}} \quad (2)$$

for predicting the value of the dependent variable, which is dichotomous (binary) in nature. The Sigmoid activation function is drawn in Figure 3.

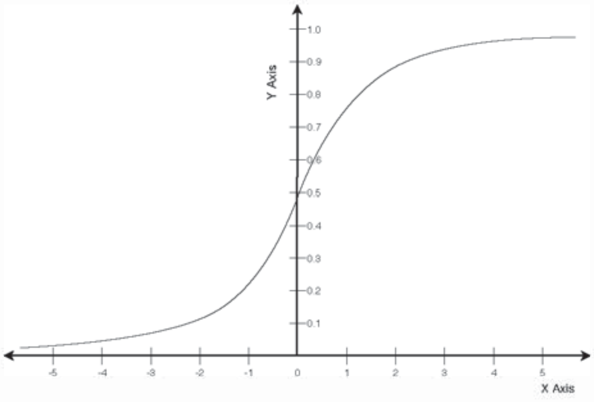

Fig. 3. Sigmoid Activation Function.

The generalised equation for logistical regression is given in Equation 3.

$$y = \frac{1}{1 + e^{-(b + c_1x_1 + c_2x_2 + c_3x_3 + ... + c_nx_n)}} \quad (3)$$

where x_1,x_2,x_3, ... are independent variables, y is a dependent variable and b, c_1,c_2,c_3, ... are constants.
For our work, We have taken n=30 as there are 30 different cell features in the dataset.

B. Support Vector Machine

SVM [12], a supervised ML algorithm, is employed in both regression and classification [13]. Mostly it's used in classification problems. In this ML algorithm, we plot each and individual data as a point in the n-dimensional space where each dimension represents a particular feature of the dataset (also called the independent variable) as shown in Figure 4. A hyperplane or sets of hyperplanes are constructed for classifying different classes. The equation of hyperplane is $W^TX=0$, where W is the normal vector to the hyperplane. Optimal hyperplane [14], one which classifies the dataset very well maximises the margin for the training dataset, is chosen from a number of hyperplanes.

We have used radial basis function (RBF) kernel [15] in our work. The mathematical representation of the RBF kernel is given in Equation 4.

978-1-7281-6387-1/20/$31.00 ©2020 IEEE

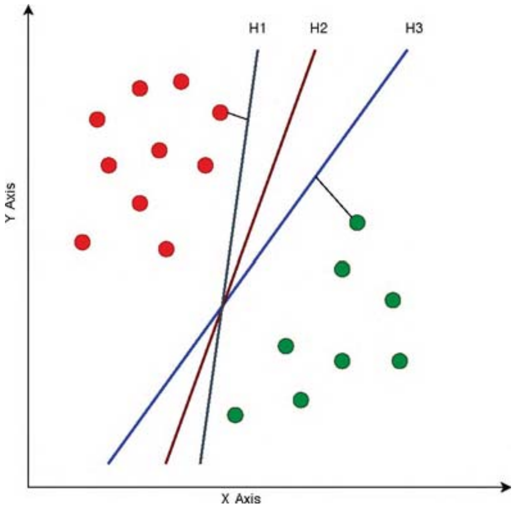

Fig. 4. Support Vector Machine.

$$k(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \quad (4)$$

C. k-Nearest Neighbours

k-nearest-neighbour [16] algorithm (KNN), is a type of supervised non-parametric machine learning algorithm [17], [18]. It classifies the new data point into one of the available categories based on the similarity principle [19]. The class of the new data point is found out based on the majority vote of its neighbouring data points [20]. Number of the neighbours to be used for classification [21] are decided manually. We have used the Euclidean distance our study to estimate the similarity. Euclidean distance [22] is calculated using the formula given in Equation 5.

$$EuclideanDisatance(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (5)$$

Depending on the Euclidean distance, calculated with the help of above-given equation, k neighbours are selected and based on the majority vote of these k neighbours new data point is classified among one of the given classes. Figure 5 shows the k-nearest algorithms for two classes of datasets: class A and class B. Based on the value of k, we classify the new data point between one of them using majority voting.

D. Naive Bayes

A Naive Bayes [23] classifier is a probabilistic classifier model. It is used for binary (dichotomous) or multi-class classification problems [24]. It is based on Bayes theorem. Bayes theorem is given in equation 6.

$$P(\frac{y}{X}) = \frac{P(\frac{X}{y})P(y)}{P(X)} \quad (6)$$

where y is dependent variable & X is a dependent feature vector of size n, shown in Equation 7.

$$X = (x_1, x_2, x_3, ..., x_n,) \quad (7)$$

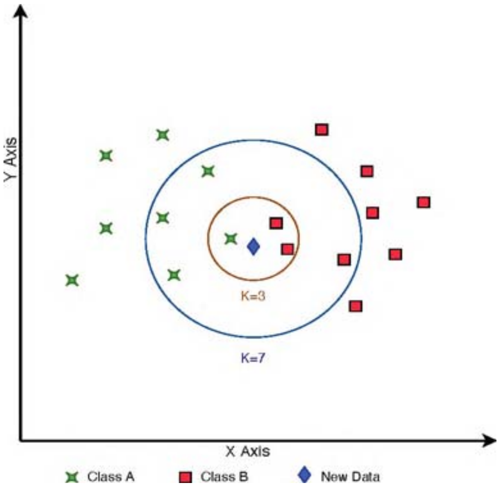

Fig. 5. k-nearest neighbours.

Figure 6 shows a typical naive Bayes classifier which classifies the new data using conditional probability. Maximum a posteriori (MAP) [25] decision rule has been used for constructing the classifier.

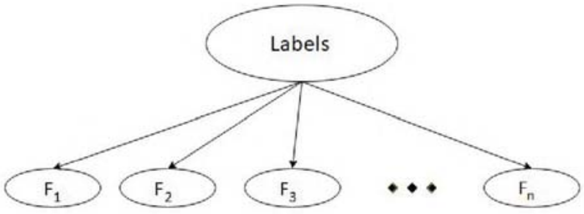

Fig. 6. Naive bayes classifier.

Naive Bayes classifiers are of three types, namely, Gaussian naive Bayes Classifier [26], multinomial naive Bayes, Bernoulli naive Bayes. We will be using only Gaussian naive Bayes in this paper. Equation 8 shows the mathematical model of the Gaussian function.

$$P(\frac{x_i}{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma^2{}_{\text{y}}}) \quad (8)$$

E. Data Analysis

We implemented all the four machine learning algorithms in two phases: (1) Training phase, and (2) Testing phase. Wisconsin data set was partitioned into 2 subsets for training subset and testing subset, assigning different values for data selection in the algorithm manually. We estimated the parameters for all the four machine learning algorithms, namely: (1) Test Accuracy, (2) Sensitivity, (3) Specificity, (4) Positive predictive value (PPV) and (5) Negative predictive value (NPV).

IV. Results and Discussion

We calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each of the algorithms separately. Fig 7 shows the bar chart comparing accuracies for all the four machine learning algorithms. All other parameters along with the accuracy are given in Table I. Among all the four algorithms, Support vector machine gave the best results with accuracy up to 98.24%, specificity .9714, sensitivity 1.0, positive predictive value .9565 and negative predictive value 1.0 for regularisation constant C=1000 and radial basis function (RBF) kernel. k-nearest neighbours performed second best with accuracy being 97.20%.

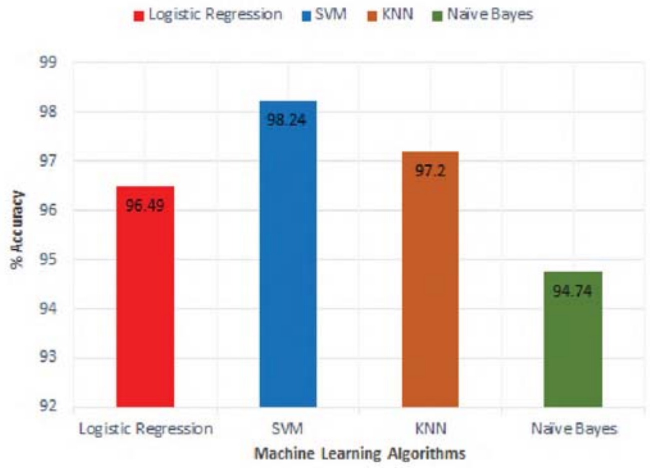

Fig. 7. Accuracy for different Algorithms.

TABLE I
Comparison of accuracy, specificity, sensitivity, NPV and PPV for different Algorithms

Algorithms	***Accuracy***	***Specificity***	***Sensitivity***	***PPV***	***NPV***
Logistic Regression	96.49 %	.9706	.9565	.9565	.9706
SVM	98.24 %	.9714	1.0	.9565	1.0
KNN	97.20 %	.9560	.9808	.9273	.9886
Naive Bayes	94.74 %	.9429	.9545	.9130	.9706

The comparison chart in Fig. 8 depicts the comparison between different algorithms. The graph clearly indicated the better accuracy of SVM over all other existing techniques. The technique dominates over other ML algorithms because of optimal paths selections to reduce processing time. The k-neareast path trimming provides an advantage to improve other factors including accuracy.

V. Conclusion

In this paper, we proposed the implementation of breast cancer diagnosis model using four different machine learning algorithms, namely: Logistic Regression, SVM, KNN and Naive Bayes in Anaconda Navigator using Python Language. These above-given algorithms gave satisfactory results. Subsequently, all other parameters like sensitivity, specificity, PPV and NPV were also found competent.

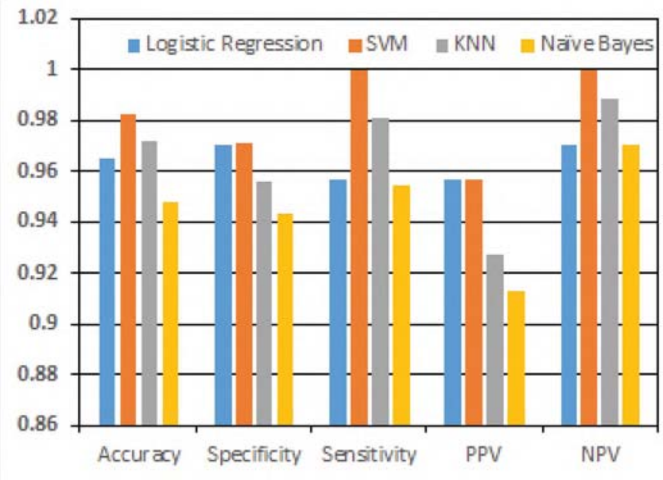

Fig. 8. Comparison chart.

The application of machine learning algorithms will not only provide the diagnosis of cancer with better accuracy as compared to manual diagnosis, but also help in minimising cost and time of diagnosis.

Acknowledgement

We are heartily thankful to our supervisor Prof. Lava Bhargava Department of Electronics and Communication Engineering, MNIT Jaipur for his effortless support & providing invaluable suggestions that greatly assisted the research. We are also thankful for the Department of Electronics for providing us with the facilities to carry out our research work.

References

[1] https://timesofindia.indiatimes.com
[2] https://www.medanta.org
[3] https://en.wikipedia.org/wiki/Breast-cancer.
[4] Z. Yan, H. Yanzhen and Y. Peng, "Computer Based Breast Cancer Diagnosis," 2009 Third International Symposium on Intelligent Information Technology Application Workshops, Nanchang, 2009, pp. 59-62.
[5] B. Bektaş and S. Babur, "Machine learning based performance development for diagnosis of breast cancer," 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, 2016, pp. 1-4.
[6] https://nbcf.org.au/about-national-breast-cancer-foundation
[7] A. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," 2015 SAI Intelligent Systems Conference (IntelliSys), London, 2015, pp. 150-154.
[8] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.
[9] https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin
[10] Abien Fred M. Agarap. 2018. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18). ACM, New York, NY, USA, 5-9.
[11] Val erie Bourd'es, St ephane Bonnevay, Paolo Lisboa "Comparison of Artificial Neural Network with Logistic regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient outcomes"
[12] Shang Gao and Hongmei Li, "Breast cancer diagnosis based on support vector machine," 2012 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering, Jalarta, 2012, pp. 240-243.
[13] W. Yi and W. Fuyong, "Breast Cancer Diagnosis via Support Vector Machines," 2006 Chinese Control Conference, Harbin, 2006, pp. 1853-1856.

[14] C. Cortes V. Vapnik "Support-vector networks" Machine learning vol. 20 no. 3 pp. 273-297 1995.

[15] X. Yang, H. Peng and M. Shi, "SVM with multiple kernels based on manifold learning for Breast Cancer diagnosis," 2013 IEEE International Conference on Information and Automation (ICIA), Yinchuan, 2013, pp. 396-399.

[16] H. R. H. Al-Absi, B. Belhaouari Samir and S. Sulaiman, "A computer aided system for breast cancer detection and diagnosis," 2014 International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2014, pp. 1-4.

[17] Pandian, A. Pasumpon. "Identification and classification of cancer cells using capsule network with pathological images." Journal of Artificial Intelligence 1, no. 01 (2019): 37-44.

[18] Vijayakumar, T. "Neural network analysis for tumor investigation and cancer prediction." Journal of Electronics 1, no. 02 (2019): 89-98.

[19] D. O. Loftsgaarden C. P. Quesenbery "A nonparametric estimate of a multivariate density function" Ann. Math. Stat. vol. 36 pp. 1049-1051 June 1965.

[20] R. Delshi Howsalya Devi and P. Deepika, "Performance comparison of various clustering techniques for diagnosis of breast cancer," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, 2015, pp. 1-5.

[21] R. Radha and P. Rajendiran, "Using K-Means Clustering Technique to Study of Breast Cancer," 2014 World Congress on Computing and Communication Technologies, Trichirappalli, 2014, pp. 211-214.

[22] H. Jegou, M. Douze and C. Schmid, "Product Quantization for Nearest Neighbor Search," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, Jan. 2011.

[23] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4.

[24] Cooper, 1999, An overview of the representation and discovery of causal relationships using Bayesian networks

[25] J. Gauvain and Chin Hui Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," in IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, April 1994.

[26] R. Pundlik, "Comparison of Sensitivity for Consumer Loan Data Using Gaussian Naïve Bayes (GNB) and Logistic Regression (LR)," 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), Bangkok, 2016, pp. 120-124.