Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA

Subrato Bharati[1], Mohammad Atikur Rahman[2] and Prajoy Podder[3]
Dept. of EEE
Ranada Prasad Shaha University Narayanganj, Bangladesh
subratobharati1@gmail.com[1], sajibextreme@gmail.com[2] and prajoypodder@gmail.com[3]

Abstract— **At present world, Breast cancer is a second main cause of cancer death in women after lung cancer. Breast cancer occurs when some breast cells begin to raise abnormally. It can arise in any portion of the Breast and it can be prevented if the treatment is started at the early stage of the Breast cancer. Breast cancer is a malignant tumour i.e. a collection of cancer cells arising from the cells of the breast Treatment of breast cancer relies on the cancer type and its stage (zero to fourth) and may include surgery, radiation, or chemotherapy. Mainly this paper focused on diagnosing the Breast cancer disease using various classification algorithm with the help of data mining tools. Data mining of the intelligent accumulated from previously disease detected patients opened up a new aspect of medical progression. In this paper, the capability of the classification of Naïve Bayes, Random Forest, Logistic Regression, Multilayer Perceptron, K-nearest neighbors in evaluating the Breast Cancer Disease dataset culled from UCI machine learning repository, was observed to predict the existence of Breast cancer. Data set has been explored in terms of Kappa Statistics, TP rate, FP Rate and precision.**

Keywords: **Naïve Bayes; Random Forest; KNN classifier; Kappa statistics; Precision; Recall; ROC area.**

I. INTRODUCTION

Now a day, breast cancer is one of the burning issue all over the world. It is one of the major health problem for women. Globally the incidence of breast cancer is only second to that of Lung cancer. The disease represents the main cause of cancer death among women. Breast cancer is developed from breast tissue. Signs of breast cancer may include a breast lump, skin dimpling, fluid coming from the nipple, breast shape change, a newly inverted nipple, or a scaly patch of skin. Breast cancer typically attack postmenopausal women. Both genetic and ancestral factor play a role. About 5-10% of breast cancer are hereditary and occur in the patient with mutation BRCA1, BRCA2 genes [1]. Prolong estrogen exposure associated with early menarche, late menopause uses of hormone replacement therapy (HRT) has been associated with increased risk. Other risk factor includes Obesity, Alcohol intake, nulliparity and late first pregnancy [2]. Breast cancer usually present as a Palpable mass with nipple discharge. Breast cancer may metastasis to bone (70%), lung (60%), liver (55%) and other organ [4].

II. LITERATURE REVIEW

JA Baker et. al. categorized malignant and benign type cancer with the help of ANN based on BI-RADS standardized lexicon. Vikas Chaurasia and his team [5] has implemented Sequential Minimal Optimization (SMO), Best First Tree and IBK data mining algorithm to obtain the classification accuracy. They declared in their research work that SMO will provide better results with good positive accuracy with minimum error rate. Zarei.S, and et.al [6] discussed about the multivariate linear regression, logistic regression, the KNN method in their paper and determined tumour category in a patient using Wisconsin Breast Cancer (WBC) database. B.Padmapriya and T.Velmurugan [5] showed the performance of J48, CART and Alternative decision Tree classification algorithms are given for the purpose of analysis the breast cancer dataset of 250 patients at the Cancer Institute, Adyar, Chennai, India. They calculated some parameters like specificity, sensitivity and kappa statistics in their paper using data mining tool. E Venkatesan et. al.[9] analysed the breast cancer dataset by some classifiers namely J48, Classification and Regression Trees (CART), Alternating Decision Tree and Best First Tree for accurate cancer prediction. Wenbin Yue discussed some machine learning techniques such as Artificial Neural network, SVM, KNN, Decision tree in order to detect the breast cancer and classify the patients according to the malignant and benign groups [11]. Ayush Sharma et. al. performed breast cancer prediction as benign or malignant type using Wisconsin Breast Cancer Dataset applying Logistic Regression, Nearest Neighbors and Support Vector Machines classifier. They showed a relationship among the value of recall and precision and the number of features that exist in dataset [12].

978-1-5386-8279-1/18/$31.00 ©2018 IEEE

III. Theoretical description of the Classifiers

In this paper, Five Classification algorithm have been discussed. They are: Naïve Bayes, Random Forest, Logistic Regression, Multilayer Perceptron, K- Nearest Neighbor classifier.

A. Naïve Bayes

Naïve Bayes classifiers assume that the effect of a variable value on a given class does not depend on the values of the other variables. This assumption can be called as class conditional independence [13]. When training data set is small, it is used to recognize the parameters essential for classification. The Naïve Bayes classifier combines the Bayes probability model with a decision rule.

B. Logistic Regression

For binary classification Logistic regression is widely used. The coefficients of logistic regression provide a linear and additive summary of the influence of a variable on the logged odds of having a characteristic on an event [14]. Here, one or more independent variable which determine the outcome. It has two outcome 1(for true) or 0(for false). It is used to illuminate data and to narrate the relationship between one dependent binary variable and one or more independent variable.

C. Multilayer Perceptron

Multilayer perceptron is the good starting point of deep learning. It is the artificial neural network's field often called neural network. It consists of at least more than two layers of nodes [16].

D. K-nearest neighbors classifier

The k-nearest neighbour's algorithm is a technique of classification. It is very simple process in machine learning. Generally for classification and regression k-nearest neighbour's algorithm is usually used. It classify the input data by the vote of its neighbours. The nearest neighbour is determined by the minimum distance point. By this method, sample data can be classified [15].

E. Random Forest

Random decision forests or Random forests are an collaborative learning system for regression, classification and other tasks that is operated by formative a multitude of decision trees at training time. Random decision forests appropriate for decision trees characteristic of over suitable to their instruction set. One important property of Random Forest the reported training error is a unbiased estimator cross validated error rate [18].

IV. Performance Parameters

A. Kappa Statistics

The statistics which assessments inter rate treaty between two categorized data sets. The range of the value from 0 to 1 and the stronger inter rate treaty is indicated by higher values. It is strong more than normal percent treaty prediction. It represents simple value of contract for chance of contract. It can be expressed by K,

$$\boldsymbol{K} = \frac{\boldsymbol{P(A)-P(E)}}{\boldsymbol{1-P(E)}} \qquad (1)$$

Here,

K = Kappa Statistics
$P\ (A)$ = Percentage of agreement
$P\ (E)$ = Agreement chance

B. TP Rate

The proportion of patients(disease present) who tested positive on the diagnostic test is the true positive rate.

C. FP Rate

False positive rate is when the disease is absent and the diagnostic test is positive.

D. Precision

Precision means the proportion of correct positive classifications (true positives) from instances that are considered as positive.

E. Recall

Recall means the ratio of correct positive algorithms (true positives) from instances that are truly positive.

F. F-Measure lining

F-measure considers both the Recall and the Precision of the process to calculate the score. It can be expressed by F,

$$\mathrm{F} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 \qquad (2)$$

G. MCC

Matthew's Correlation Coefficient or MCC is a good option that helps us to choose and represent a single value confusion matrix. Normally it is used when two confusion matrix are not so comparable. Though there is no suitable way of telling the confusion matrix of true and false negatives and positives with a on its own number, the MCC is usually observed as being the finest treatments. It is basically a correlation coefficient between the predicted and actual series. It profits values between -1 and +1. The coefficient of -1 denotes total difference between observation and prediction, +1 represents a accurate prediction and 0 no better than random prediction. It can be considered by the resulting equation,

$$MCC = \frac{TP*TN-FP*FN}{\frac{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}{1}} \qquad (3)$$

H. Receiver Operating Characteristics Area or ROC Area

The area under the ROC curve is known as ROC area or Receiver Operating Characteristics area. It is a graphical plot that illustrate the presentation of a binary classifier. Here some false positive rate and true positive rate at several threshold is plotted for creating essential plot. Each point on the ROC area signifies a specificity/sensitivity parallel corresponding to a particular decision threshold where True positive is acknowledged as sensitivity and the false positive is acknowledged as fall out. It estimates the performance of a classifier based on: excellent (0.90-1), good (0.80-0.90), fair (0.70-0.80), poor (0.60-0.70), and fail (0.50–0.60).

V. Experimental result

The breast cancer data set used in this paper is collected from UCI machine learning repository, USA [17]. This data contains records of 286 instances with 10 attributes to predict and analyse breast cancer. These attributes are age, menopause, inv-nodes, tumour size, deg-malig, node-caps, breast-quad, breast, irradiate and class.

Number of Instances used: 106

Number of Attributes considered: 10

Type of classification: {Recurrence events, Non recurrence events}

Class Distribution: [201 for Recurrence events] [85 for Non recurrence events]

Visualization of all attributes of the dataset has been displayed in fig. 1.

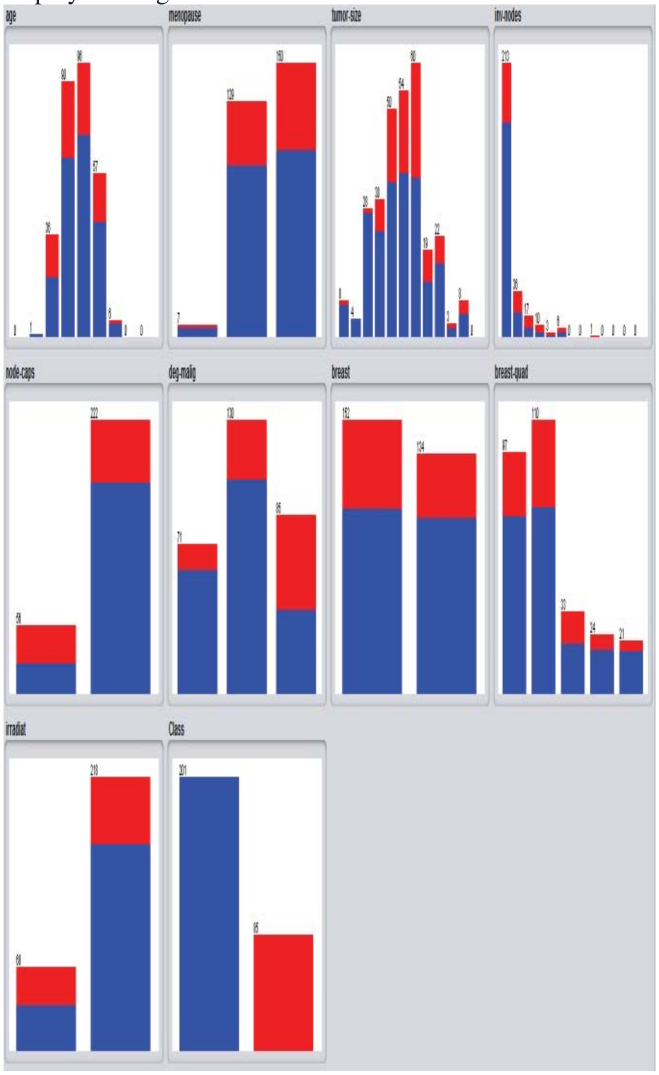

Fig. 1 Visualization of the attributes of breast cancer dataset.

From table 1, it can be stated that the percentage of correctly classified instances of KNN classifier (72.3776% i.e. 207 instances) is comparatively larger than other mentioned classifier and it is approximately closer to the Naïve Bayes classifier (71.6783% i.e. 205 instances). As a result, the percentage of incorrectly classified instances of KNN (27.6224% i.e. 79 instances) is comparatively less than the others classifier where Multilayer perceptron gives highest incorrect classification instances (35.3147% i.e. 101 instances). Here, the total number of instances is 286.

TABLE I
Correctly and Incorrectly Classification

Classifier	Correctly Classified Instances	In Correctly Classified Instances	Kappa statistic
Naïve Bayes	71.6783 %	28.3217 %	0.2857
Logistic Regression	68.8811 %	31.1189 %	0.1979
Multilayer Perceptron	64.6853 %	35.3147 %	0.1575
K-nearest neighbors classifier	72.3776 %	27.6224 %	0.2438
Random Forest	69.5804 %	30.4196 %	0.1736

From table 2, it can be seen that Random Forest has largest value of MAE (0.3727). KNN and Naïve Bayes have comparatively less MAE and approximately equal value (0.3272 and 0.3257). Random Forest has large percentage of relative absolute error (89.0857%).

TABLE II
MSE, RMSE, RAE Calculation

Classifier	Mean absolute error	Root mean squared error	Relative absolute error
Naïve Bayes	0.3272	0.4534	78.2086%
Logistic Regression	0.37	0.4631	88.4196%
Multilayer Perceptron	0.3552	0.5423	84.8811%
K-nearest neighbors classifier	0.3257	0.5101	77.8513%
Random Forest	0.3727	0.4613	89.0857 %

TABLE III
TP, FP, Precision and Recall Calculation

Classifier	TP Rate	FP Rate	Precision	Recall
Naïve Bayes	0.717	0.446	0.704	0.717
Logistic	0.689	0.505	0.668	0.689
Multilayer Perceptron	0.647	0.489	0.648	0.647
K-nearest neighbors classifier	0.724	0.511	0.699	0.724

Random Forest	0.696	0.543	0.664	0.696

From table 3, it can be seen that KNN has highest True Positive rate (0.724) and Multilayer Perceptron has lowest True Positive rate (0.647). Also, KNN and Naïve Bayes have good Precision value (0.699 and 0.704 respectively). Random Forest has largest False positive rate (0.543).

Table 4 illustrates that Naïve bayes has highest MCC, on the contrary Multilayer Perceptron has low MCC value. Naïve Bayes covers highest percentage of ROC area (70.1%).

TABLE IV
F-MEASUREMENT, MCC AND ROC AREA CALCULATION

Classifier	**F-measure**	**MCC**	**ROC Area**
Naïve Bayes	0.717	0.446	0.701
Logistic Regression	0.675	0.202	0.646
Multilayer Perceptron	0.647	0.158	0.623
K-nearest neighbors classifier	0.697	0.261	0.628
Random Forest	0.669	0.184	0.634

VI. CONCLUSIONS

Breast cancer has been predicted and considered for some classifiers such as Naïve Bayes, Random Forest, Logistic Regression, Multilayer Perceptron, K-nearest neighbors classifier. WEKA data mining tool has been used and compared the presentation of these classifier algorithms. It is observed the performance results of K-nearest neighbors classifier algorithm. It provides the highest correctly classified instances of 97.9021%. The second most accurate classifier is Multilayer Perceptron with correctly classified instances of 96.5035 %. This paper mainly visualized 286 instances with 10 attributes to predict and analyse breast cancer dataset has been showed. It discusses the performance of different classification algorithm on the basis of distribution plot. This paper also observed Kappa statistic, Mean absolute error, F-measure, MCC, ROC Area, Relative absolute error, FP rate, TP rate, Root mean squared error, Precision Recall and. It has also been comprised that K-nearest neighbors classifier has highest percentage (99.9%) of ROC Area and Multilayer Perceptron has second highest percentage (98.0%) of ROC Area.

REFERENCES

[1] Paraskevi Apostolou and Florentia Fostira, “Hereditary Breast Cancer: The Era of New Susceptibility Genes”, BioMed Research International Vols. 2013.

[2] Jaimini Majali, Rishikesh Niranjan,Vinamra Phatak, Omkar Tadakhe,“Data Mining Techniques for Diagnosis And Prognosis of Cancer”, Int. Journal of Advanced Research in Computer and Communication Engg., Vol. 4, Issue 3, pp. 613-614, 2015.

[3] K. R. Lakshmi, M. Veera Krishna, S.Prem Kumar, “Performance Comparison of Data Mining Techniques for Prediction and Diagnosis of Breast Cancer Disease Survivability”, Asian Journal of Computer Science and Information Technology, Vol. 3, pp. 81 – 87, 2013.

[4] Joshi, Miss Jahanvi, and Mr. Rinal Doshi, Dr.Jigar Patel. "Diagnosis And Prognosis Breast Cancer Using Classification Rules", Int. Journal of Engineering Research and General Science, Vol. 2, Issue 6, pp. 315-323, 2014.

[5] B. Padmapriya, T. Velmurugan, “Classification Algorithm Based Analysis of Breast Cancer Data”, International Journal of Data Mining Techniques and Applications, Volume 5, Issue 1, Page no.43-49, June 2016.

[6] Vikas Chaurasia, Saurabh Pal, “A Novel Approach for Breast Cancer Detection using Data Mining Techniques” International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, pp.2464, 2014.

[7] Zarei, S., Aminghafari, M., HakimehZali, “Application and comparison of different linear classification methods for breast cancer diagnosis”, International Journal of Analytical, Pharmaceutical and Biomedical Sciences, Vol. 4, Issue2, pp. 123-128, 2015.

[8] Y. Ireaneus Anna Rejani, Dr. S. Thamarai Selvi, “Early Detection Of Breast Cancer Using SVM Classifier Technique”, Int. Journal on Computer Science and Engineering, Vol. 1, Issue 3, pp. 127-130, 2009.

[9] E Venkatesan , T. Velmurugan, “Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification”, Indian Journal of Science and Technology, Vol. 8(29), November 2015.

[10] Jain Lakhmi C, Jain Ashlesha, Jain Ajita, “Artificial Intelligence Techniques In Breast Cancer Diagnosis And Prognosis”, World Scientific, 2000.

[11] Wenbin Yue , Zidong Wang , Hongwei Chen , Annette Payne, Xiaohui Liu, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis”, Designs, Vol. 2, Issue: 2, 13, 2018.

[12] Ayush Sharma, Sudhanshu Kulshrestha, Sibi Daniel, “Machine learning approaches for breast cancer diagnosis and prognosis”, International Conference on Soft Computing and its Engineering Applications (icSoftComp), Changa, India, 1-2 Dec. 2017.

[13] M. Narasimha Murty, V. Susheela Devi, “Pattern Recognition: An Algorithmic Approach”, Springer Science & Business Media, May 25, 2011.

[14] Fred C. Pampel, “Logistic Regression: A Primer”, SAGE Publishers, pp. 35-39, May 26, 2000.

[15] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer Science & Business Media, Jun 24, 2013.

[16] Petra Perner, “Machine Learning and Data Mining in Pattern Recognition”, 6th International Conference, MLDM 2009, Leipzig, Germany, July 23-25, 2009, Proceedings

[17] Matjaz Zwitter & Milan Soklic, “UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Breast%2BCancer] ”, University Medical Centre, Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia, 1988.

[18] Tony Fischetti, Eric Mayor, Rui Miguel Forte, “R: Predictive Analysis”, Packt Publishing Ltd, Mar 31, 2017.