ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter378102020.pdf
https://doi.org/10.30534/ijeter/2020/378102020

Comparative analysis of various classification techniques to predict and detect Breast Cancer Using R

Arnav Manchanda[1], Dr. Anu Manchanda[2], Swati Mathur[3], Dr. Helen Josephine V.L[4]
[1]Student, Greenwood High School, Bengaluru, arnavm@greenwoodhigh.edu.in
[2]Associate Professor, Department of Master of Computer Applications, CMR Institute of Technology, Bengaluru, anumanchanda@gmail.com
[3]Assistant Professor, Department of Master of Computer Applications, CMR Institute of Technology, Bengaluru, swati.1116@gmail.com
[4]Associate Professor, Department of Master of Computer Applications, CMR Institute of Technology, Bengaluru, helenjose.cbe@gmail.com

ABSTRACT
Breast cancer is widely spreading among women in developing as well as developed countries and is a major area of concern among medical researchers. Our objective is to find out the best classification method that can help the practitioner in the medical field to detect and predict the early occurrence of Breast cancer. We have worked on Naive Bayes, KNN, SVM, Decision Tree, and Random Forest in the R tool
.

Key words: Naive Bayes, SVM, KNN, Random Forest and Decision Tree, R, Breast cancer.

1. INTRODUCTION

Data mining involves the extraction of useful information from a huge amount of data using various techniques. Few of the data mining tasks are frequently carried out by data mining systems include association rule mining, classification, clustering, prediction, etc. These techniques are being implemented in various fields like business, mathematical, scientific, and medical. One of the important and vital tasks in data mining is classification. It can be implemented successfully in the medical field due to its effectiveness like in predicting the occurrence or non-occurrence of the disease. It helps in the early detection of the disease and reduces the cost of treatment.

In the medical field, the most threatening and common disease among females is breast. The year 2018, among all the deaths among women 15% of women died of breast cancer alone.

Approximately 627,000 women were accounted to die because of breast cancer in the year 2018[1]. It is one of the main sources of death among women throughout the realm [2]. Breast cancer cases are found to be more in developed countries and rates are growing up globally.

Breast cancer cases can be brought down if it could be discovered at early phases. Detection at the early stages would help to bring down the mortality rate. The early detection procedures are required to categorize the patient into one of the two groups either a “benign” or “malignant”. Data mining techniques can be utilized effectively to classify the patient into any of the two above stated categories of breast cancer. Medical research data are collected in huge amounts, made available to the researchers. These datasets are utilized by the researchers working on various technologies. Data mining techniques are becoming a popular tool for researchers working in the medical field.

These tools assist them in identifying the relationships among different variables and to find out the patterns etc. These methods can be used to forecast and classify breast cancer patterns. In our research work, we have tried to equate the performance of diverse classification algorithms that are suitable to predict and detect breast cancer into two classes “benign” and “malignant”. We have worked in the R tool as not many studies have implemented their work in R and most of them have carried their work using the Weka tool.

2. LITERATURE REVIEW

To diagnose cancer is vital in the field of medicine. Many studies have been carried out so far as to identify breast cancer. These studies have been carried out using different techniques of data mining, artificial intelligence, and neural networks.

Mümine Kaya Keles [3] carried a study to identify the breast cancer studied different data mining IBk, Bagging, Random Forest, Random Committee, and Simple Classification and Regression Tree algorithms using the Weka tool. Accuracy was found to be higher than 90% for Random Committee, Random Forest, Bagging, IBk, and Simple CART algorithms. Therefore, it’s possible to identify the tumor at the early stages. Jesús Silva et al. [4] studied the classification and prediction techniques to find out the repetitive occurrence of breast cancer. The Naive Bayes classifier was selected as the preliminary classifier and, in the 2nd instance, later SVM, J48, and Generalized Regression Neural Network. Weka tool was made use of to find out the effectiveness of the classification algorithm. Naïve Bayes and SVM gave an

accuracy of 89% in comparison to 91% accuracy of J48 and GRNN.In another study done by Keerti et al. [5], used the SEER dataset and classified dataset into "benign" and "malignant". I was found that the C4.5 algorithm gave better accuracy than the Naïve Bayes algorithm.

In a study by B.Padmapriya et al.[6] J48, AD Tree, and CART algorithms of data mining were used for data analysis to analyze Breast cancer data using mammogram images. Weka tool was used to run the algorithms. The highest accuracy of 98.1 % was of J48 followed by 97.7% of ADTree and 98.5 % of CART. It was found that CART algorithm's performance was best among the three to classify mammogram images.

To study early detection of breast cancer. Sudhamathy et al. [7] took the dataset was from the 'mlbench' package of CRAN. R package was used to analyze the decision tree classification methods. The basis of the study of the various algorithm was accuracy, recall, precision, specificity, and sensitivity. Results revealed that the 'randomforest' decision tree classification method was best suited for their objective under study.

C Nalini et al. [8] used two classification techniques namely Decision tree and K nearest neighbor to forecast Breast cancer. Weka tool was utilized to carry out data analysis. The outcomes indicated that the performance of the Decision tree was better than the K Nearest algorithm.

Another study was carried out on three classification techniques, J48, MLP, and rough set [9]. Data analysis was done using ROSETTA for rough set classification and WEKA for neural network and decision tree. Feature selection was done and results were compared with and without feature selection and it was found that the feature selection technique was significant in improving the accuracy of different techniques to increase the Receiver Operating Characteristics (ROC) to diagnose breast cancer disease.

3. METHODOLOGY

a. R language

R is used for statistical analysis in the computing field. It offers a wide range of statistical and graphical techniques to analyze the data. R was developed by GNU [10]. It is a freely available software. It runs on a wide variety of platforms like UNIX, Windows, and macOS. It offers both a command line and a graphical interface. There are many default packages installed in it and each package performs a certain analysis. We have used R Studio with version R- 3.6.2.

b. Dataset

In our study, we worked on [11] Wisconsin Diagnosis Breast Cancer Database (WDBC).

In WDBC, an FNA of the breast mass is used in digitized form to extract features. Cell nuclei characteristics are represented by the image. A total of 569 instances are there in which 357 are benign and the remaining are malignant. The instances are described by 32 attributes. Details of the attribute are: ID number, Diagnosis, and for each cell-nucleus various parameters are calculated: Radius Texture, Area, Perimeter, Smoothness, Concavity, Compactness, Concave points, Fractal Dimension, and Symmetry.

Table 1: Related Work

Research Paper title	**Method used**	**Outcome**
Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: Comparative Study [3]	IBk, Bagging, Random Forest,	Random Committee, Random Forest, Bagging, IBk, and Simple CART algorithms with accuracy greater than 90%
Comparative Study on different Classification techniques for breast cancer dataset [4]	Neural network, Decision tree, and Rough set	Accuracy of J48= 79.97%, Accuracy of MLP =75.35% Accuracy of
Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R [5]	C4.5 and Naive Bayes algorithm	C4.5 accuracy= 98.09% Naïve Bayes =95.85%.
Classification Algorithm Based Analysis of Breast Cancer Data [6]	J48, CART, and AD Tree	J48 accuracy = 98.1 % ADTree accuracy = 97.7% CART accuracy=98.5 %.
Data mining classification technique applied for breast cancer [7]	Decision tree, k-nearest neighbor algorithms	Decision tree Precision=0.75 Recall=0.84 k- nearest neighbor Precision=0.76 Recall=0.82
Comparative Analysis of R Package Classifiers Using Breast Cancer Dataset [8]	Decision tree classification methods namely, 'rpath', 'ctree', and randomforest'	The method of 'randomforest' precision=1
Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Type of Breast Cancer Recurrence [9]	C4.5 (J48), Naive Bayes algorithm, SVM and GRNN	J48 and GRNN had an accuracy rate of 91%. Naïve Bayes and SVM had an accuracy of 89%.

c. Classification Techniques

The classification model is one of the data analyzing techniques using which we can classify our data and assign a label or predict the target value to the given data. There are various techniques for classification. Using the various algorithms, we will check whether a patient is having breast cancer or not. Classification is a two-step process:

i. Training Phase: It is a learning phase in which a general model is designed using the training dataset.

ii. Testing Phase: At this stage, a trained model is used to predict the class label for testing data.

In this paper, we are presenting various classification problems to predict the breast cancer of a patient: Decision Tree, KNN, SVM, Naive Bayes, and random forest.

i) Decision Tree: It's [12] is a predictive based model which generates a tree-like structure, where internal nodes represent attributes and leaf nodes represents class label (output). In R-programming," party" package is used to create a decision tree. ctree() is defined in the party package using which decision tree is created. Ctree means conditional inferences. But before giving data to ctree(), it has to be converted into categorical values. To install all dependent packages of party, install the whole package by running the command

install. packages("party")

ii) K Nearest Neighbour: It is a non-parametric algorithm. It classifies data based on the likeness of all the available data objects [13]. K represents the number of nearest neighbors considered for assigning a label to a given data point. "class" package of the R language, helps to apply the KNN algorithm. To install this package, we have to write:

install. packages("class")

iii) Naive Bayes: It is a probabilistic classification technique centered on Bayes theorem [14] which assumes independence among features. It is suitable for large datasets.

In R-programming, naivebayes() is defined in "e1071" package. To install the dependency of e1071 package type:

install. packages("e1071")

iv) Support vector machine: It is one of the supervised learning algorithm [16]. It divides the datasets into different classes based on the hyperplane. The hyperplane is an n-1 dimensional subspace for n-dimensional space. SVM handles linearly separable and non-linearly separable cases.

There are many packages for SVM supported by R. Among them we have used Kernlab(Kernel-Based Machine Learning) and caret(Classification And REgression Training).

v) Random Forest: It is a modified version of the decision tree. It generates multiple decision trees and then combines all of them and generates a common result [15]. Random forest is a package defined by R, which has implemented Breiman's random forest algorithm.

4. RESULTS AND DISCUSSION

In this paper, medical data related to breast cancer is taken into consideration to forecast and detect cancer. Cancer stages are divided into two stages: Benign and malignant. The experiment is conducted using the R tool. In this, various classification algorithms are used. Among all algorithms, the ratio of training and the testing dataset is 80:20. In the KNN algorithm, how to choose the K parameter? In our dataset, the number of instances is 569. So K = SquareRoot(569)=23 is taken.

In SVM, a linear kernel function is used. Different measures are used to evaluate the performance discussed below:

a. Precision: It is calculated as per the below-given formula.

Precision= TP/ (TP + FP)

Where TP and FP stand for true positive and false positive respectively.

b. Recall or True Positive Rate: It is calculated as per the below-given formula.

Recall= TP/ (TP + FN)

Here FN stands for false negative

c. Accuracy is defined as per the given below:

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

where TN denotes true negatives

d. Depending on the above performance matrices, the confusion matrix is identified for each of the five algorithms which are given below.

Table 2: NAIVE BAYES CONFUSION MATRIX

	B	M
B	74	3
M	5	37

Table 3: KNN CONFUSION MATRIX

	B	M
B	88	0
M	2	24

Table 4: DECISION TREE CONFUSION MATRIX

	B	M
B	64	9
M	1	42

Table 5: SVM CONFUSION MATRIX

	B	M
B	71	4
M	0	38

TABLE 6: RANDOM FOREST

	B	M
B	65	5
M	0	46

After calculating the confusion matrix, performance metrics are calculated which are represented in table 7.

Table 7: Results Comparison of five algorithms based on Accuracy, Precision, and Recall

Algorithm	Accuracy	Precision	Recall
Naive Bayes	0.9328	0.9367	0.6667
K-Nearest Neighbour	0.9646	0.97778	0.6514
Decision Tree	0.9138	0.9846	0.6038
SVM with	0.9825	1	0.7857
Random Forest	0.9652	1	0.5856

From figure 1, we find that the accuracy of SVM is the highest followed by Random Forest.

From Figure 2, it can be seen that the precision of both SVM and Random forest is highest.

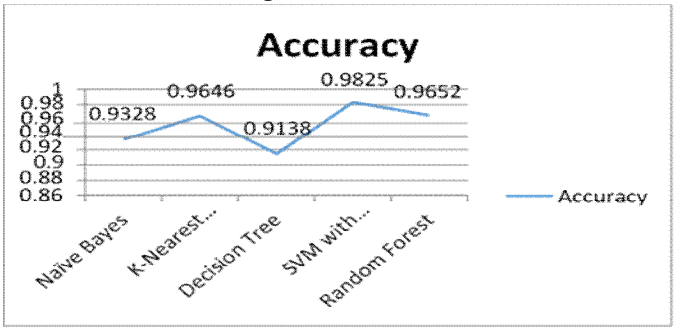

Figure 1: Comparing the accuracy of classification methods

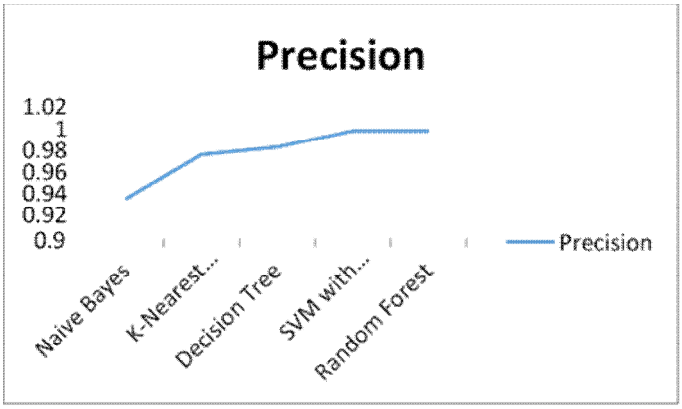

Figure 2: Comparing precision of classification methods

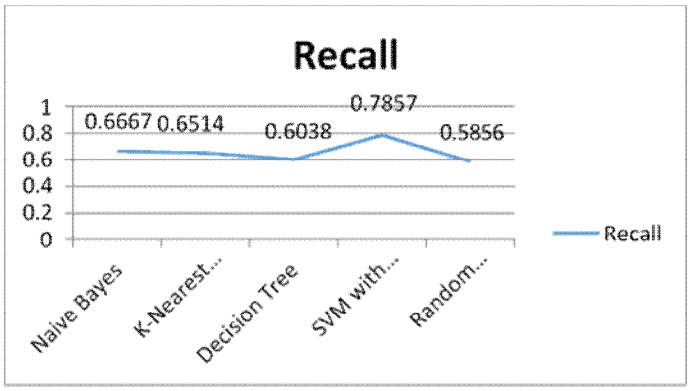

Figure 2: Comparing recall of classification methods

From 3, we can say that the recall parameter of SVM is the highest.

5. CONCLUSION

It's a challenging task to find out the most suitable and precise classifier in predicting breast cancer. We have used five different classification algorithms. With reference to table 7, we can say that SVM is the best algorithm based on accuracy and recall. In the future, different algorithms can be made to work together and comparisons can be done using different statistical tools.

REFERENCES

1.https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-4th-edition.pdf. Retrieved on 13- Mar-2020.

2.https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. Retrieved on 13- Mar-2020.

3. M. Kaya Keleş, **"Breast cancer prediction and detection using data mining classification algorithms: A comparative study,"** *Tehnički vjesnik,* Vol. 26, no. 1, (2019), pp. 149–155, doi: 10.17559/TV- 20180417102943.

4. J. Silva, O. Bonerge Pineda Lezama, N. Varela, L. Adriana Borreroa, **"Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Type of Breast Cancer Recurrence**," *GPC*, 2019.

5. K. Yeulkar, R. Sheikh, "**Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R**," *International Journal for Research in Applied Science & Engineering Technology*, Vol. 5, no. 3,(2017), pp. 406-410.

6. B. Padmapriya and T. Velmurugan, "**Classification Algorithm based analysis of Breast cancer data**," *International Journal of Data Mining Techniques and Applications*, Vol. 5, no. 1, (2016), pp. 43– 49, doi: 10.20894/ijdmta.102.005.001.010.

7. G. Sudhamathy, M. Thilagu, and G. Padmavathi, "**Comparative analysis of R package classifiers using breast cancer dataset**," *International Journal of*

Engineering and Technology, Vol. 8, (2016), no. 5, pp. 2127–2136, doi: 10.21817/ijet/2016/v8i5/160805432.

8. C. Nalini, T. Poovozhi, **"Data Mining Classification Technique Applied for Breast Cancer**," *International Journal of Pure and Applied Mathematics*, Vol. 119, No. 12, (2018), pp. 10935-10945.

9. A. Lebbe, S. Saabith, E. Sundarajan, and A. A.Bakar "**Comparative study on different classification techniques for breast cancer dataset**", *International Journal of Computer Science and Mobile Computing*, Vol. 3, no. 10, (2014), pp. 185-191.

10. https://www.r-project.org/about.html Retrieved on 10-Mar-2020.

11.https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Retrieved on 15- Dec-2019.

12. Z. Zhang, **"Decision tree modeling using R**," *Ann. Transl. Med.*, Vol. 4, no. 15, (2009), pp. 1–8, doi: 10.21037/atm.2016.05.14.

13.https://towardsdatascience.com/k-nearest-neighbors-knn-algorithm-bd375d14eec7 Retrieved on 1- Mar-2020.

14.Z. Zhang, "**Naïve bayes classification in R**," *Ann. Transl. Med*., Vol. 4, no. 12, (2016), pp. 1–5, doi: 10.21037/atm.2016.03.38.

15.https://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-challenges/?utm_source=blog&utm_medium=understandingsupportvectormachinearticle Retrieved on 1- Mar-2020.

16. A. Karatzoglou, D. Meyer, and K. Hornik, "**Support Vector Algorithm in R**," *Journal of Statistical Software*, Vol. 15, no. 9, (2006), pp. 1–28.