Comparison Of Datamining Techniques For Prediction Of Breast Cancer

Deneshkumar V, Manoprabha M, Senthamarai Kannan K

Abstract— Breast cancer is one of the most challenging deadly diseases. Correct and in-time prediction of such disease is very important. Wisconsin breast cancer dataset with 569 patients and 32 features were included in this study. The Information Gain and Gini Index were used to determine the effectiveness of features on breast cancer. The performance comparisons of the most commonly used statistical methods were also studied to find the best predictive model. The main objective of this manuscript is to make use of the advanced technologies to develop a best predictive model for breast cancer. All performance assessments were carried out using Rapid Miner Studio software.

Index Terms— Breast cancer, Data mining, Prediction, Feature Selection, Gini Index, Information Gain and ROC Curve.

1 INTRODUCTION

BREAST cancer is the most commonly diagnosed cancer among women. It is caused due to the abnormal cell growth in the breast. A tumour may be malignant or benign. Malignant is the cancerous tumour which is more harmful and spread to other parts of the body. Benign is the tumour which grows but will not spread. Detection of cancer in their early stage is very important. For the prediction of such cancerous disease, we used the Wisconsin breast cancer dataset with 569 patients and 32 features. Different predictive models have been computed and their performance has been evaluated. The features describe the characteristics of the cell nuclei which are computed from a digitized image of a breast mass. Each patients diagnosis are put in to two possible categories either Benign (not harmful) or Malignant (harmful). The compactness, area, texture, fractal dimension, smoothness, concave points, perimeter, symmetry, radius and concavity are the ten real-valued features computed from the image of the cell nuclei. The mean, standard error and worst of these features were also computed resulting in 30 features. The features which shows high impact on the cause of the disease is selected using some feature selection methods like correlation, Information Gain and Gini index. For every 4 minutes a woman is diagnosed with breast cancer in India. Every 8 minutes a woman dies for this disease. In 2012, it is estimated that 70,218 woman died due to breast cancer in India which is the highest in the world for that year. Breast cancer occurs mostly in the age group of 30-50. The patients diagnosed with breast cancer increases every year in India. Early prediction of this disease is more important, which is the main objective of this research work.

- *Manoprabha M, Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012. Email Id- manoprabhamurugan@gmail.com*
- *Deneshkumar V, Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012. Email Id- vdenesh77@gmail.com*
- *Senthamarai kannan K, Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012. Email Id- senkannan2002@gmail.com*

2 BACKGROUND OF THE STUDY

Chaurasia et al. [8] compared three data mining techniques to predict breast cancer. The prediction models used are Naive Bayes, RBF Network and J48. 10-fold cross validation method is also used to find the unbiased estimator. Among the three model Naive Bayes is indicated as the best predictor. Alizadehsani et al. [2] used data mining approach for diagnosis of coronary artery disease. They used some of the feature selection and feature creation methods to find the most effective feature on CAD. The highly relevant features are selected using Information Gain and confidence. Performance of the used models are also evaluated, highest accuracy of about 94.08% is achieved. Jena and Kamila [13] derived a procedure for prediction of chronic kidney disease using classification algorithms. This study focus on studying the performance of six different classifiers, which gives high accuracy in detecting the kidney disease. Here multilayer perceptron is evaluated as the best classifier. Dehkordi and Sajedi [10] aimed at predicting a disease based on the prescription of the patient's using data mining methods. They had derived maximum information from the dataset. Using the prescription they predicted the physician of each patient and the type of the disease the patient is suffering from. K-Nearest Neighbor is the appropriate method for predicting such dataset. Arbain and Balakrishnan [3] investigated some of the data mining algorithm to predict liver disease on imbalanced dataset. Comparisons are done based on accuracy and ROC index. KNN performs well than other model with accuracy of 99.7%. Srinivas et al. [19] applied some data mining techniques in healthcare and prediction of heart attack. Rich information can be discovered while using these models. Classification based techniques like Decision tree, Naive Bayes and Artificial Neural Network are used. One Dependency Augmented Naive Bayes classifier and naive creedal classifier 2 are applied to find the most significant factors that are related to heart attack. Jacob et al. [12] carried out a survey on prediction of breast cancer using data mining algorithms. Knowledge discovery plays a vital role in identification of diseases in early stage. Here varied clustering and

classification algorithms are used. From the comparison of the outcome it is cleared that classification algorithms are superior to clustering algorithms. Kunwar et al. [14] have done some analysis on predicting the chronic kidney disease from a huge unmined data using data mining classification technique. All experiments are implemented in Rapid miner. From the result produced Naive Bayes is more accurate than other models. Ayatollahi et al. [4] compared two data mining algorithms for predicting coronary artery disease. This study aimed at comparing the positive predictive value of SVM and ANN algorithms. From the result it is concluded that SVM model provides better classification than ANN. Sanjay et al. [17] proposed a data mining tool to predict breast cancer at an early stage using hybrid feature selection method. This new feature selection technique increases the prediction accuracy and used to find the most relevant feature which causes breast cancer. Shouman et al. [18] used data mining techniques in the diagnosis and treatment of heart disease. They took a new attempt in identifying the suitable treatment for each heart patient's. Also studied whether the data mining technique shows reliable performance on diagnosis of cardiac diseases. Chaurasia and pal [6] detected heart disease using data mining models. This prediction was done eliminating the irrelevant attributes to reach the maximum accuracy and some classifier models where applied. Unbiased estimate of the classifiers were found using 10-fold cross validation method. Abbass [1] experimented breast cancer diagnosis using evolutionary artificial neural networks approach based on the pareto differential evolution algorithm. This approach was termed as memetic pareto artificial neural network. He then compared his result with the standard back-propagated neural network and showed that his approach has better generalization and lower computational cost. Ya-Qin et al. [20] proposed a predictive model based on decision tree on imbalanced data. Sampling is taken to carry out the disadvantages caused by imbalanced data. The performance is evaluated using some basic criteria and the performance is considered best when the distribution of the data is equal. Integration decision tree model is build using bagging algorithm. Bellaachia and Guven [5] presented an analysis on predicting breast cancer survivability on SEER dataset. The C4.5 decision tree algorithm, the back-propagated neural network and Naive Bayes are applied for this data. The prediction performances are evaluated after several experiments and finally the C4.5 algorithm showed better performance than other two algorithms. Chaurasia and pal [7] analysed the performance of data mining algorithms by using heart and breast cancer dataset. Here the most popular five algorithms are used. The performance of each model is evaluated using the prediction accuracy.

3 METHODOLOGY

Data Mining is predicted to be "one of the revolutionary developments of the next decade" [9]. It is the process of discovering interesting patterns and knowledge from large amounts of data [11]. Naive Bayes, Logistic regression, Decision tree, Random forest and Support Vector Machine are the classification algorithms used to analyse the dataset. Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical class labels. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis.

3.1 Algorithm Used

3.1.1 Naive Bayes algorithm

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naive Bayes classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called classconditional independence.

The probability of dataset X having the class label C_i is:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

The class label C_i with largest conditional probability value determines the category of the dataset.

3.1.2 Logistic regression algorithm

Logistic regression is an alternate regression technique. It refers to methods for describing the relationship between a categorical response variable and a set of predictor variables. This algorithm is suitable for binary classification. The value produced by logistic regression is a probability value between 0 and 1. These values are computed by

$$Yi = e^u/(1 + e^u) \quad (2)$$

Where Yi is the estimated probability that the ith case is in a category and u is the regular linear regression equation [16].

$$u = A + B_1X_1 + B_2X_2 + \cdots + B_KX_K \quad (3)$$

3.1.3 Decision Tree algorithm

J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser) during the late 1970s and early 1980s. Decision trees can easily be converted to classification rules. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. It can handle multidimensional data.

3.1.4 Random Forest algorithm

The collection of classifiers is a forest. It is a general technique of random decision forests that are an ensemble learning technique for classification. It is constructed by the multitude of decision trees at training time and outputting the class that is the mode of the classification of the individual trees [15]. It generally controls over fitting and improves the accuracy of decision trees decisions.

3.1.5 Support Vector Machine algorithm

SVM is a method for the classification of both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. It searches for a "decision boundary" separating the tuples of one class from another. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.

3. 2. Feature Selection

3.2.1 Information Gain

Let node *N* represents or hold the tuples of partition *D*. The splitting attribute is chosen based on the Information Gain, which is evaluated as.

$$Info(D) = -\sum_{i=1}^{m} P_i log_2(P_i) \quad (4)$$

Where *pi* is the nonzero probability that an arbitrary tuple in *D*. A log function to the base 2 is used, because the information is encoded in bits. *Info* (*D*) is just the average amount of information needed to identify the class label of a tuple in *D*. *Info* (*D*) is also known as the entropy of *D*.

3.2.2 Gini Index

The Gini index measures the impurity of *D*, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (5)$$

Where *pi* is the probability that a tuple in *D*. The Gini index considers a binary split for each attribute. The attribute that maximizes the reduction in impurity is selected as the splitting attribute.

4 Experimental Results and Discussion

For prediction of breast cancer we have used the Wisconsin breast cancer dataset. This dataset used in our experiment contains the record of 569 patients and 32 features. The dataset is available through the UWCS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/. Table 1 describes about the features which shows the characteristics of the cell nuclei of each patients along with their range respectively. Among 569 patients 212 has abnormal cell nuclei and 357 are benign.

Table 1. Features of Wisconsin dataset

Feature Name	Range
id	-
Diagnosis	B- Benign, M-Malignant
radius_mean	6.981 – 28.110
texture_mean	9.710 – 39.280
perimeter_mean	43.790 – 188.500
area_mean	143.500 – 2501
smoothness_mean	0.053 – 0.163
compactness_mean	0.019 – 0.345
concavity_mean	0 – 0.427
concave points_mean	0 – 0.201
symmetry_mean	0.106 – 0.304
fractal_dimension_mean	0.050 – 0.097
radius_se	0.112 – 2.873
texture_se	0.360 – 4.885
perimeter_se	0.757 – 21.980
area_se	6.802 – 542.200
smoothness_se	0.002 – 0.031
compactness_se	0.002 – 0.135
concavity_se	0 – 0.396
concave points_se	0 – 0.053
symmetry_se	0.008 – 0.079
fractal_dimension_se	0.001 – 0.030
radius_worst	7.930 – 36.040
texture_worst	12.020 – 49.540
perimeter_worst	50.410 – 251.200
area_worst	185.200 – 4254
smoothness_worst	0.071 – 0.223
compactness_worst	0.027 – 1.058
concavity_worst	0 – 1.252
concave points_worst	0 – 0.291
symmetry_worst	0.157 – 0.664
fractal_dimension_worst	0.055 – 0.207

For selecting the most effective features the "weight by SVM" method is used which shows a better accuracy in classifying the patients in to Malignant and Benign categories. For recognizing the relevant attribute the weights of Information Gain and Gini index are evaluated. Moreover, both feature selection method showed similar result. The Information Gain using weight by SVM is shown in table 2. And the weight using Gini index is shown in table 3. Combining both methods the features with weight higher than 0.6 are considered and the prediction algorithms are applied.

Table 2. Information Gain

Attributes	IG
texture_worst	1.0
radius_worst	0.7624
symmetry_worst	0.7490
area_worst	0.7061
perimeter_worst	0.6845
smoothness_worst	0.6745
radius_se	0.6709
texture_mean	0.6127
concave points_worst	0.5951
concave points_mean	0.5818
area_se	0.5417
concavity_mean	0.5244
area_mean	0.4865
perimeter_se	0.4812
radius_mean	0.4758
concavity_worst	0.4601
perimeter_mean	0.4598
fractal_dimension_mean	0.3652
compactness_se	0.3629
texture_se	0.3128
fractal_dimension_se	0.2311
fractal_dimension_worst	0.2128
smoothness_mean	0.1693
smoothness_se	0.1475
symmetry_se	0.1319
compactness_mean	0.1296
concave points_se	0.0528
concavity_se	0.0203
compactness_worst	0.0191
symmetry_mean	0.0

The weights of the features are displayed in descending order ranging from 0 to 1.Comparing the results of both methods the top features with high confidence which tends closer to 1 are chosen. These features are useful in making decision for patients while diagnosis.

4.1 Comparing the performances of different algorithms

We have used five prediction algorithms i.e. Naive Bayes, Logistic regression, Decision tree, Random forest and Support vector Machine. The performance between these algorithms are compared using the classification accuracy, sensitivity, specificity, classification error, AUC value and the total run time of the model. The comparisons are executed in to two parts, first all algorithms are applied on the dataset with selected features and the second part includes all the features without any elimination. The performance of each algorithm is shown in table 4.

Table 3. Gini Index

Attribute	Gini index
radius_worst	1.0
area_worst	0.9932
perimeter_worst	0.9898
concave points_worst	0.9811
concave points_mean	0.9689
area_mean	0.8601
perimeter_mean	0.8447
radius_mean	0.8421
concavity_mean	0.8170
area_se	0.7665
concavity_worst	0.7165
perimeter_se	0.5443
radius_se	0.5440
compactness_mean	0.4891
compactness_worst	0.4580
concavity_se	0.3310
texture_mean	0.2855
concave points_se	0.2816
texture_worst	0.2728
smoothness_worst	0.2236
symmetry_worst	0.1948
compactness_se	0.1882
smoothness_mean	0.1558
fractal_dimension_worst	0.1316
symmetry_mean	0.1133
fractal_dimension_se	0.0452
symmetry_se	0.0200
fractal_dimension_mean	0.0194
texture_se	0.0030
smoothness_se	0.0

Based on the results in table 4 logistic regression showed a better performance compared to other algorithms with feature selection, which was nearly 97%. All the models with their respective standard error are also shown with accuracy. The accuracy rate of each algorithm is displayed in fig 1. For all the models the sensitivity value is higher than specificity value. On comparing all the models the decision tree algorithm showed considerably lower performance in predicting the patient's classes. Overall performance of logistic regression was good based on all the criteria examined.

Table 4. Performance comparison of the algorithms used

Used feature	Algorithm Used	Accuracy	Sensitivity	Specifi-city	Classific-ation Error	AUC Value	Total run time of the model (sec)
With selected features	Naïve Bayes	95.7± 1.7%	99.0%	90.7%	4.3%	0.998	14
	Logistic regression	96.9± 0.1%	99.2%	93.9%	3.1%	0.998	167
	Decision tree	92.0± 2.7%	97.1%	84.4%	8.0%	0.908	41
	Random forest	92.6± 1.7%	93.9%	90.1%	7.4%	0.985	760
	Support Vector Machine	92.6± 2.7%	96.9%	85.3%	7.4%	0.975	33
All features without feature selection	Naïve Bayes	92.6±3.4%	95.0%	89.9%	7.4%	0.988	13
	Logistic regression	93.8±2.2%	94.8%	92.4%	6.2%	0.952	10
	Decision tree	91.4±2.5%	96.4%	84.2%	8.6%	0.903	4
	Random forest	94.5±3.4%	96.1%	92.0%	5.5%	0.989	78
	Support Vector Machine	95.6±4.7%	98.1%	91.7%	4.4%	0.983	14

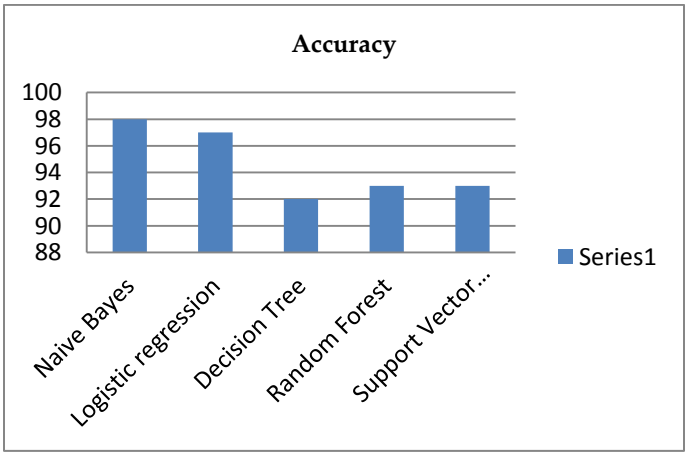

Fig 1. Classification accuracy of different algorithms

As observed in the next part, same algorithms which were directly applied without any feature selection are shown. Result in this part evaluated shows support vector machine as the best algorithm with an accuracy of about 95.6%. Finally on comparing both the results in part 1 and 2, i.e. with feature selection and without feature selection, the part with feature selection has higher accuracy. This shows that selecting the most relevant features increases the prediction accuracy and including all other irrelevant features could lead to false prediction. But the runtime of the model with feature selection are eventually higher than without feature selection. While the record increases the runtime for prediction will also be higher.

ROC curves are useful tools for model selection. ROC curves plot the true positive rate (or sensitivity) versus the false positive rate of one or more classifiers. In fig 2, the ROC curve for all the algorithms can be seen. When the Area Under the Curve value was higher, higher the performance of the algorithm will be. From the diagram Naive Bayes and logistic regression has the greater AUC value which was then followed by Random forest and Support Vector Machine.

Confusion matrices are used in analysing the performance of the prediction models. True positive, true negative, false positive and false negative values for all used algorithms are presented in table 5. While using Naive Bayes algorithm 205+336 patients have been predicted correctly and 23+5 patients were wrongly predicted. This follows the same for all other algorithms, based on this result logistic regression performance was better compared to other models.

Table 5. The confusion matrix of different algorithms using the selected features

Algorithms		true M	true B
Naïve Bayes	pred. M	205	5
	pred. B	23	336
Logistic regression	pred. M	199	6
	pred. B	11	353
Decision tree	pred. M	182	11
	pred. B	34	342
Random forest	pred. M	191	22
	pred. B	22	334
Support Vector Machine	pred. M	177	11
	pred. B	34	347

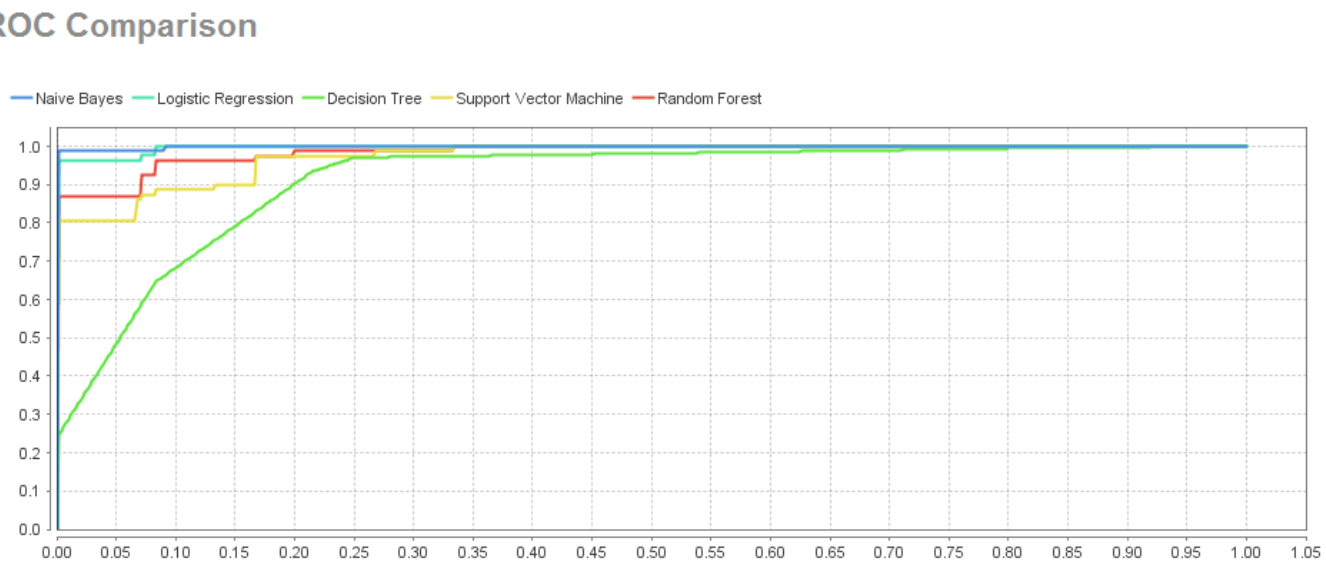

Fig 2. ROC diagram for all the five algorithms used

5 CONCLUSION

Several data mining algorithms have been used on the Wisconsin breast cancer dataset and the results are discussed. Along with the predictive model the feature selection method is also used to improve the prediction accuracy. Algorithms applied with selected feature showed higher accuracy in predicting the classes of the patients whether in Malignant or Benign, than the model applied without feature selection. Logistic regression is assessed as the best model with accuracy of about 97% compared to the other models used. Hence applying this proposed model can identify the state of the patients with high probability. The goal is to increase the accuracy of diagnosis and identifying the most relevant feature which may be useful in the prevention of the disease at its earliest.

REFERENCES

[1] H. A. Abbass, "An Evolutionary Artificial Neural Networks Approach for Breast Cancer Diagnosis," Artificial Intelligence in Medicine, vol. 25, pp. 223-232, July. 2002.

[2] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A Data Mining Approach for Diagnosis of Coronary artery Disease," Computer Methods and Programs in Biomedicine, COMM- 3519, pp. 1-10, July. 2013.

[3] A. N. Arbain, and Y. P. Balakrishnan, "A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data," International Journal of Data Science and Advanced Analytics, vol. 1, no. 1, pp. 1-11, Feb. 2019.

[4] H. Ayatollahi, L. Gholamhosseini, and M. Salehi, "Predicting Coronary artery Disease: a Comparison between Two Data Mining Algorithms," BMC Public Health, vol. 19, no. 448, pp. 1-9, Apr. 2019.

[5] A. Bellaachia, and E. Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," In: Scientific Data Mining Workshop (in conjunction with the 2006 SIAM conference on data mining), Bethesda, Maryland, pp. 20-22, 2006.

[6] V. Chaurasia, and S. Pal, "Data Mining Approach to Detect Heart Diseases," International Journal of Advanced Computer Science and Information Technology, vol. 2, no. 4, pp. 56-66, Jan. 2014.

[7] V. Chaurasia, and S. Pal, "Performance Analysis of Data Mining Algorithms for Diagnosis and Prediction of Heart and Breast Cancer Disease," International Journal of Innovative Computing, Information & Control, vol. 3, no. 8, pp. 1-13, May. 2014.

[8] V. Chaurasia, and S. Pal, and B.B. Tiwari, "Prediction of Benign and Malignant Breast Cancer using Data Mining Techniques," Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119-126, Jan. 2018.

[9] T. L. Daniel, "Data Mining Methods and Models," A John Wiley & Sons, INC Publication, Hoboken, New Jersey, 2006.

[10] S. K. Dehkordi, and H. Sajedi, "Prediction of disease based on Prescription using Data Mining Methods," Health and Technology, pp. 1-8, July. 2018.

[11] J. Han, M. Kamber, and J. Pei, "Data Mining Concepts and Techniques," 3rd edition, Morgan Kaufmann Publishers is an imprint of Elsevier, Waltham, MA, USA, 2012.

[12] D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh, S. Raj, "A Survey on Breast Cancer Prediction Using Data Mining Techniques," Proc. IEEE Conference on Emerging Devices and Smart Systems, pp. 256-258, Mar. 2018.

[13] L. Jena, and N.K. Kamila, "Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-disease," International Journal of Emerging Research in Management & Technology, vol. 4, no. 11, pp. 110-118, Nov. 2015.

[14] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease Analysis using Data Mining Classification Techniques," 6th International conference – Cloud System and Big Data Engineering (Confluence), pp. 300-305, Jan. 2016.

[15] A. K. Mishra, and B. K. Ratha, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," International Journal on Advanced Electrical and Computer Engineering, vol. 3,no. 4, pp. 5-7, 2016.

[16] J. Padmavathi, "Logistic regression in Feature Selection in Data Mining," International Journal of Scientific & Engineering Research, vol. 3, no. 8, pp. 1-4, Aug. 2012.

[17] A. Sanjay, H. V. Nair, S. Murali, and K. S. Krishnaveni, "A Data Mining Model To Predict Breast Cancer Using Improved Feature Selection Method on Real Time Data," International Conference on Advances in Computing, Communications and Informatics, pp. 2437-2440, Sep. 2018.

[18] M. Shouman, T. Turner, and R. Stocker, "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment," 2012 Japan-Egypt Conference on Electronics, Communications and Computers, pp. 173-177, Mar. 2012.

[19] K. Srinivas, B. K. Rani, and A. Govrdhan, "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," International Journal on Computer Science and Engineering, vol. 2, no. 2, pp. 250-255, Mar. 2010.

[20] L. Ya-Qin, W. Cheng, and Z. Lu, "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data," 3rd International Conference on Bioinformatics and Biomedical Engineering, Beijing, China, pp. 11-13, July. 2009.