Original Article

Model To Detect Breast Cancer Based On Patient Symptoms

Mr. Satish Dekka[1*], Dr.K. Narasimha Raju[2], Dr. D. ManendraSai[3], Mr. Mohammad Rafi[4]

[1*]Associate Professor, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology (A), Vizianagaram, JNTUK-AP
[2]Associate Professor, Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering. JNTUK-AP
[3]Associate Professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women. JNTUK-AP
[4]Lecturer, Computer Science and Engineering Department Yanbu University College, Yanbu Kingdom of Saudi Arabia

***Corresponding Author: -** Mr. Satish Dekka

*Associate Professor, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology (A), Vizianagaram, JNTUK-AP
DOI: 10.47750/pnr.2022.13.S05.439

Abstract

The number of medical data warehouses is expanding quickly these days. As a result, it is difficult for us to predict or analyse these facts in order to uncover hidden knowledge that is valuable. For forecasting medical analysis, many machine learning methods and tools are employed. The most prevalent and well-known malignancy, particularly among women, is breast cancer. It ranks among the leading global causes of death. The sole remedy is early detection, which lowers the mortality rate from breast cancer. Breast cells can develop into cancer, which is known as breast cancer. Breast cancer has recently become a highly serious disease, not just in India but also in other nations. The primary goal of this research is to diagnose breast cancer patients as early as possible. Three machine learning approaches Decision Tree, Support Vector Machine, and Logistic Regression are employed for the early detection and prevention of breast cancer patients. These techniques help reduce waiting times and human and technical errors in breast cancer diagnosis. By employing these methods, we can increase the number of lives saved and decrease the death rate by maximising early diagnosis of breast cancer. The likelihood that an infection will be successfully treated depends on precisely identifying and locating it as soon as possible using logistic regression and SVM. A significant obstacle to the diagnosis of breast cancer is the classification of the appropriate machine learning technique. Thus, in order to analyse risk levels that contribute to prognosis, we developed a model for a breast cancer early prediction system. Doctors can diagnose breast cancer using this experimental study, and patients can benefit from early therapy to prolong their lives.

Keywords: Machine learning, Breast Cancer, Classification techniques.

1. INTRODUCTION

Medical data archives are getting bigger and bigger all the time these days. Due of this, it is difficult for us to predict or analyse these facts in order to uncover hidden knowledge that is useful. It is possible to anticipate medical analyses using a variety of machine learning tools and approaches. In instance, women are more likely to be diagnosed with breast cancer than any other type. It is a leading cause of death in the entire world. Early detection is the only way to stop breast cancer deaths. The cells of the breasts are where breast cancer develops. Today, not only in India but also in other nations, breast cancer has emerged as a very serious disease. Early diagnosis of breast cancer patients is the main goal of this article. Three machine learning techniques Decision Tree, Support Vector Machine, and Logistic Regression—are used to diagnose breast cancer earlier and with less human and technical error. They help reduce waiting times for patients and improve early prevention and diagnosis of the disease. The accuracy of the early detection and correct location of the infections utilising this logistic regression and SVM techniques determines its cure rate and expectation. One of the biggest challenges in the diagnosis of breast cancer is identifying the best machine learning technique. As a result, we developed a model for a system that can predict breast cancer before it spreads, and it also analyses risk factors that affect prognosis. To diagnose breast cancer and to assist patients in receiving early treatment to prolong their lives, this experimental study of a paper is highly helpful to doctors.

2. LITERATURE SURVEY

After reviewing the different literature showed that there have been several studies on the early detection and prevention of breast cancer using data mining techniques such as decision tree [1, 2]. Sahar A. Mokhtar et al [3] have studied decision tree, artificial neural network and support vector machine classification models for the prediction of the severity of breast cancer. The performances of the three models have been evaluated using the statistical measures (accuracy, sensitivity, specificity) and found that Support vector machine model performance is better than the other two models on the prediction of the severity of breast cancer. In the study of Pend Harker patterns in breast cancer, data mining tool is

valuable in identifying patterns in breast cancer cases that can be used for diagnosis, prognosis, and treatment purposes [4]. Rajshree Dash et al [5] was proposed a hybridized K-means clustering algorithm to improve the efficiency of the original K-means clustering by applying the PCA on original data sets. It is a new approach to identifying cluster centres and the steps of assigning data points to appropriate clusters. There was a given data set that partitioned into k clusters. After the experimental result, it shows that the proposed algorithm provides better efficiency and accuracy comparison to the original k-means algorithm with reduced time. Zakaria Suliman zubi et al [6] used different data mining techniques such as neural networks for detection and classification of lung cancers. In the Modern medical sciences, there should be developed new blood test technique that can easily detect the eight types of common cancers in a single test which is known as SEEK cancer. In this type of blood test detects tiny amounts of DNA and proteins released into the bloodstream from cancer cells. This type of blood test technique indicates the presence of ovarian, liver, stomach, pancreatic, oesophageal, bowel, lung or breast cancers. This blood test is known as a liquid biopsy, which is different from a standard biopsy, where a needle is put into a solid tumour to confirm a cancer diagnosis. SEEK Cancer is also far less invasive. It will be helpful for the early diagnosis of cancer and more chance of cure with the modern medical medicines and surgery. Decision trees are classification algorithms that are becoming more powerful and popular with the advancement of the data mining. To construct the DT, mathematical algorithms (e.g., information gain, Gini index, and Chi-squared test) are used to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups [7]. Random Forest is a bagging algorithm that successfully applied at highly quantified models [8]. Due to overfitting problems in the decision tree, Random Forests builds hundreds or may thousands of trees. That are different from each other, it uses random samples with replacement [29]. On the average, 33% of the rows will be left out of each sample [10]. Each tree classifies its observations, and at the end majority votes [11], decisions are chosen. Random Forest can also be used in the unsupervised mode for assessing proximities among data points [12] Support Vector Machine (SVM) is the idea of a hyperplane that divides a dataset into two classes in the best possible way. SVM has been used in different types of problems and that have already been successful in pattern recognition in bioinformatics, cancer diagnosis [13]. The separating Hyperplane can be linear or non-linear. For linear classification, SVM computes the linear decision function in the central gap of the two classes by classifying all the training data points and placing the decision function as far from the given data points as possible. SVM also uses a non-linear mapping technique to transform the original training data into a higher dimension. To minimizing the classification errors, perform the classification for separate the classes [14].

3.METHODOLOGY

Modern screening methods and treatments have decreased the fatality rate from breast cancer in recent years [15, 16]. Different sorts of scientific study offer information about the illness in various but complementary ways. Through the use of several advanced techniques and computational methodologies, breast cancer treatment has been made more effective. Consequently, this cancer's fatality rate has dropped in recent years. A. Typical Breast Cancer Research Methods Laboratory tests, observational studies, and clinical diagnosis are the most often used investigational techniques for breast cancer. Hypothesis testing is done under controlled circumstances in laboratory experiments. that produces accurate findings but is constrained by the regulated setting. Observational studies look at a population's characteristics and determine the relationship between the variables and the result. It doesn't always prove the link between the cause and the effect. Conduct a medical investigation involving humans as part of the clinical diagnosis. This demonstrates that there is a causal connection between the factors and the results. The discovery of medications and the treatment of breast cancer have both made extensive use of clinical diagnostics [17]. Analysis of the medical prognosis medical prognosis is a scientific subject that assesses the likelihood of disease recurrence and forecasts a patient's or group of patients' survival [18]. Medical prognosis analysis is used to estimate the health of patients. These projections can be used to guide treatment planning based on anticipated results. The advancement of medical sciences and its existence is made possible by the fields of knowledge discovery and data mining techniques. These techniques are more potent than conventional statistical techniques. In The likelihood that a breast cancer patient would survive after being diagnosed with the disease has increased due to recent advancements in early detection and prevention [19]. In this study, we investigate three machine learning approaches for breast cancer detection studies as shown in the figure 1. The goal of this study is to create prediction models that can identify and stop breast cancer in its earliest stages. The collected data is used for the pre-processing of data to eliminate the further burden during the construction of the model. The entire data is split as training set and testing set to build the classification model and to evaluate the model. The There are different machine learning techniques are used to build the classification model. The model is tested for important metrics such as Accuracy and AUC.

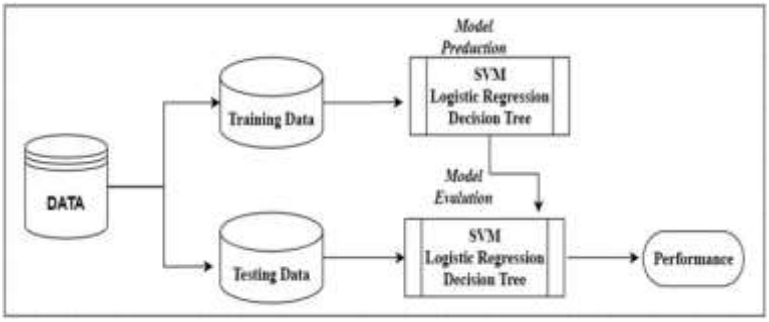

Figure 1: Proposed Model for the classification of patients with Breast Cancer.

4. EXPERIMENTAL SETUP AND EVALUATION

High-performance programming language for technical computing is MATLAB. Several tests are carried out using various machine learning technique in a MATLAB environment. Data is retrieved from the dataset made available by the Kaggle. Variables that influence the identification of the breast were selected, focusing on symptoms and diagnosis and disregarding identification and treatment variables. This led to a selection of 10 variables associated with the symptoms or signs that a patient may present, and 1 variable associated with the diagnosis that allows the identification of the type of Breast Cancer. Table 1 presents a list of the 10 identified variables and their description. In this work, DT, LR, and SVM three machine learning techniques are used to early prevention and detection of breast cancer to find which method performs better.

Table 1: List of identified variables.

Sno	*Variables*	*Description*
1	*radius*	*(Mean of distances from center to points on the perimeter)*
2	*texture*	*(Standard deviation of gray-scale values)*
3	*perimeter*	
4	*area*	
5	*smoothness*	*(Local variation in radius lengths)*
6	*compactness*	*(Perimeter^2 / area - 1.0)*
7	*concavity*	*(Severity of concave portions of the contour)*
8	*concave points*	*(Number of concave portions of the contour)*
9	*symmetry*	
10	*fractal dimension*	*("Coastline approximation" - 1)*

The type of variable states the different values that it can assume and can be either continuous, e.g., radius, or texture, e.g., smoothness. The variable “Type” indicates the diagnosis issued by the treating physician based on the symptoms and medical record of the patient, with the possibility of presenting of one of the following classifications:

1. Invasive ductal carcinoma
2. Invasive lobular carcinoma
3. Inflammatory breast cancer
4. Paget's disease of the breast
5. Angiosarcoma of the breast
6. Phyllodes tumors
7. Ductal carcinoma in situ (DCIS)
8. Lobular carcinoma in situ (LCIS)
9. HER2 status
10. Triple-negative breast cancer
11. Metastatic breast cancer
12. Breast cancer in men

4.1 Performance metrics

There are different performance metrics that are used to evaluate the performance of our designed model are accuracy, AUC and RoC.

Accuracy: it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

AUC: The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).

$$AUC = \frac{TP+TN}{2 \times (TP+TN+FP+FN)}$$

The dataset is divided into two sets: training and testing, the first corresponding to 80% of the data (320) and the second to 20% (80).

5.RESULTS AND DISCUSSION

Experiments are conducted with different models – "Decision trees, Support vector machines, Logistic Regression" and observed their comparative analysis. The figures indicates the confusion matrix of the three models, AUC curves. And the parallel coordinates the maximum accuracy for 10 variables is obtained by using a Support vector machine, which results in an accuracy of 97.7%. This means that the classification of Breast Cancers by the Support vector machines coincides with that issued by the treating physician in 97.7% of the 400 cases comprising the test set.

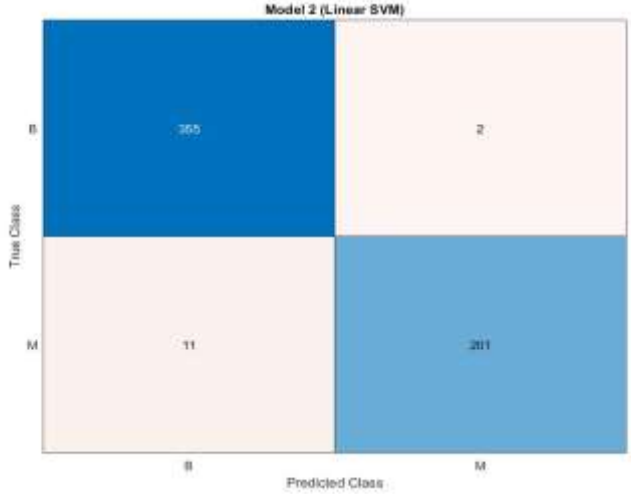

Figure 2: Running scenarios of Prediction model 1 Confusion Matrix for SVM

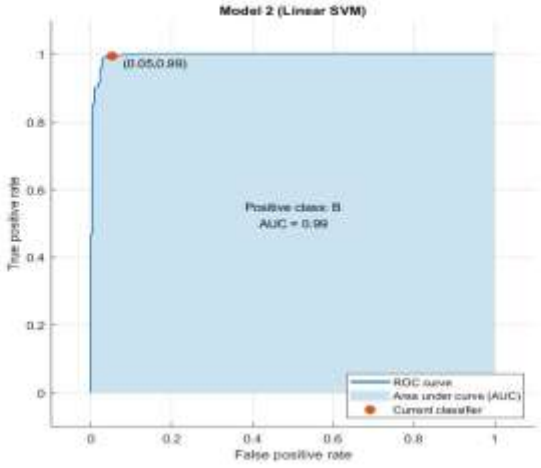

Figure 3: Running scenarios of Prediction model 1 ROC Curve for SVM

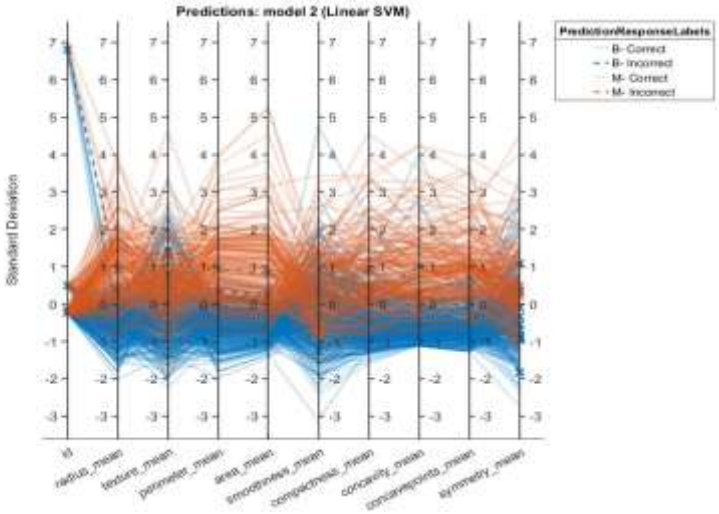

Figure 4: Running scenarios of Prediction model 1 Parallel Coordinators for SVM

The Decision Tree models show accuracies and precisions >92%, highlighting values obtained with this model for 10 variables which reaches values 92.6% in both metrics, thereby indicating adequate classification on comprising the dataset.

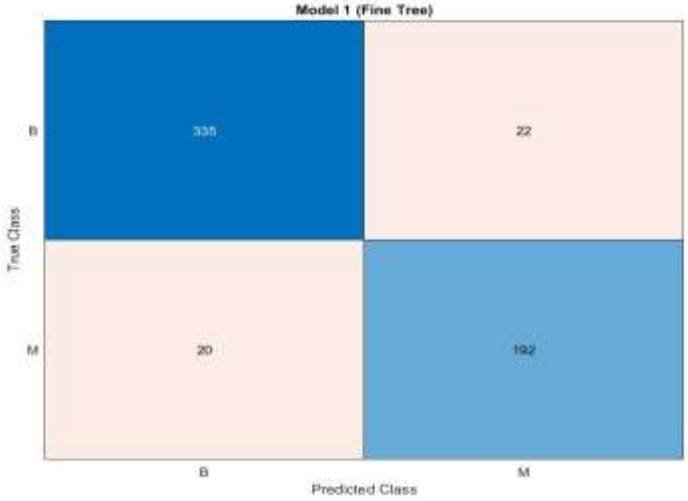

Figure 5: Running scenarios of Prediction model 1 Confusion Matrix for Decision tree

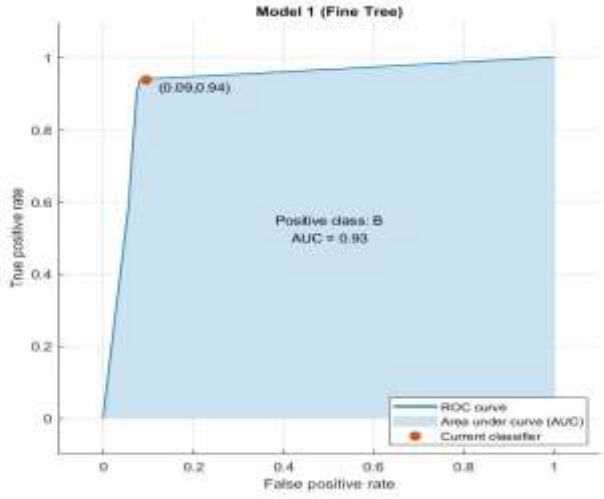

Figure 6: Running scenarios of Prediction model 1 ROC Curve for Decision tree

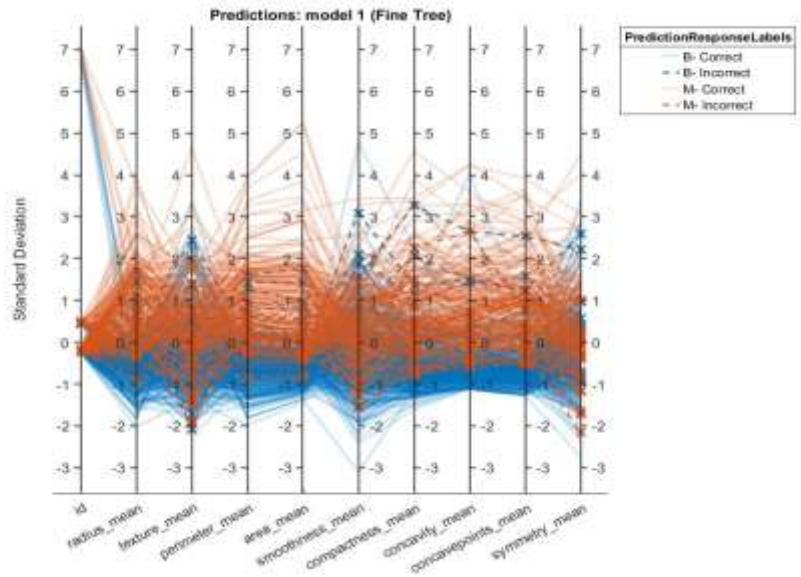

Figure 7: Running scenarios of Prediction model 1 Parallel Coordinators for Decision tree

The Logistic Regression models show accuracies and precisions >94%, highlighting values obtained with this model for 10 variables which reaches values 94.9% in both metrics, thereby indicating adequate classification on comprising the dataset.

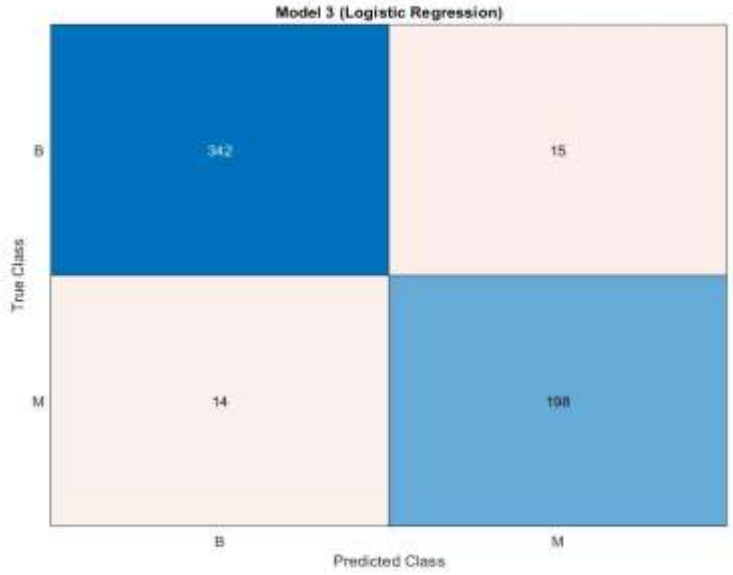

Figure 8: Running scenarios of Prediction model 1 Confusion Matrix for Logistic Regression

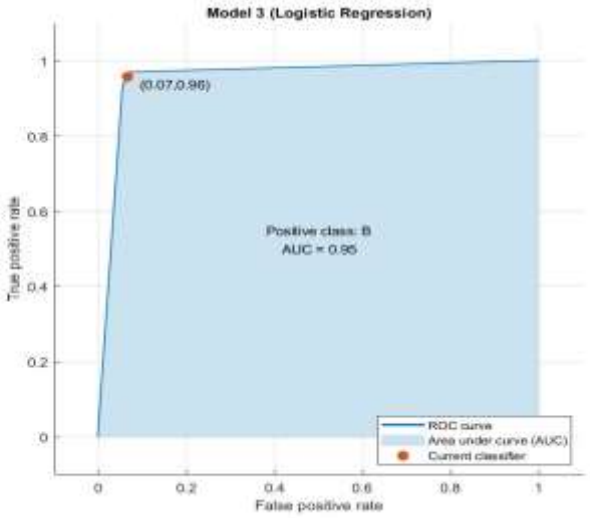

Figure 9: Running scenarios of Prediction model 1 ROC Curve for Logistic Regression

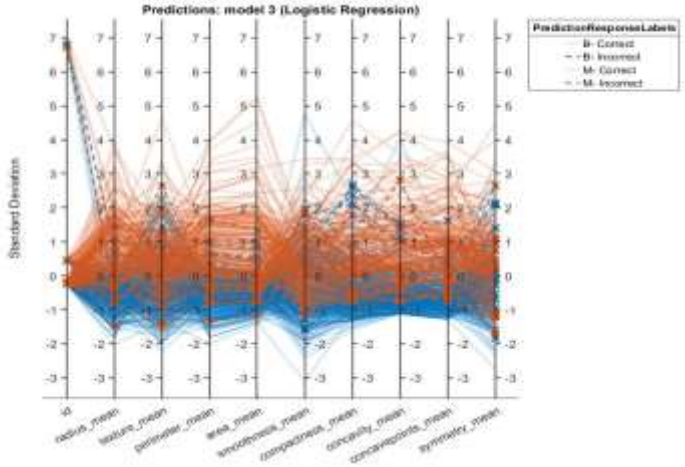

Figure 10: Running scenarios of Prediction model 1 Parallel Coordinators for Logistic Regression

Comparison of Machine Learning Algorithms

For the sake of comparison, the performance of the classification obtained by the best model proposed here is compared with the results of the previously published work for classification of Breast Cancer. Table 2 presents the results of the performance measures for various configurations and classification models. Precision, a metric calculated independently for each of the classes, is taken as the weighted average of the seven Breast Cancer classes.

Table 2: Comparative Analysis

Complete model with 10 variables					
S. No		**Accuracy**	**Precision**	**Training Time:**	**Prediction Speed:**
1	Decision trees	92.6%	0.94	12.847sec	~ 3100 obs / sec
2	Support vector machines	97.7%	0.97	18.893 sec	~ 3100 obs / sec
3	Logistic Regression	94.9%	0.94	5.0831 sec	~ 5400 obs / sec

CONCLUSION AND FUTURE SCOPE

The early detection of breast cancer will easily prevent with the help of the different medical therapy. In this work, we analyzed breast cancer at early stage with the help of the machine learning techniques and compared the performance of the DT, LR and SVM. It is observed that the Support vector machines performance is better than the other techniques to predicting the cancer at early stage. In future work, the current work can further be studied with deep learning techniques.

REFERENCES

1. Zhou ZH, Jiang Y (2003) Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans Inf Technol Biomed 7: 37-42.

2. Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 34:113-127.
3. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org
4. Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N,Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.
5. Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66
6. Zakaria Suliman zubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" Journal of Software Engineering and Applications, February 2014, 7, 69-77
7. Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.
8. Ziegel, E. R. (2012). The Elements of Statistical Learning.Technometrics.
9. Kotsiantis, S. B. (2013). Decision trees: a recent overview.Artificial Intelligence Review, 39(4), 261-283.
10. Montano-Gutierrez, L. F., Ohta, S., Kustatscher, G., Earnshaw, W. C., & Rappsilber, J. (2016). Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data, 050302.
11. Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests. Bioinformatics, 32(6), 859-866.
12. Afanador, N. L., Smolinska, A., Tran, T. N., & Blanchet, L.(2016). Unsupervised random forest: a tutorial with case studies. Journal of Chemometrics, 30(5), 232-241.
13. Cristianini N, Shawe-taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, London: Cambridge University Press.
14. Joachims T (1998) Making large-scale support vector machine learning practical. Advances in Kernel Methods:Support Vector Learning. MIT Press, Cambridge, MA, 169-184.
15. Warren J. Cancer death rates falling, but slowly. WebMD medical news; 2003 (http://aolsvc.health.webmd.aol.-com/content/ Artcile/73/ 82013.htm).
16. Progress shown in death rates from four leading cancers (http://cancer.gov/newscenter/pressreleases/2003Report Release).
17. The ABCs of breast cancer types of research studies(http://www.komen.org/bci/abs/chap_01.asp).
18. Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform 2001; 34:428—39.
19. Brenner H, Gefeller O, Hakulinen T. A computer program for period analysis of cancer patient survival. Eur J Cancer 2002;38(5):690—5.