PREDICTION OF BREAST CANCER USING MACHINE LEARNING TECHNIQUES

[1] Medisetty Hari krishna, [2] Dr. Kunjam. Nageswara Rao
[1]M. Tech Student, [,2]Professor
[1,2]Department of Computer Science and Systems Engineering [1,2]Andhra University College of Engineering, Visakhapatnam, AP, India
[1]hrkrishna98@gmail.com, [2]kunjamnag@gmail.com

Abstract

Breast Cancer is one of the leading diseases that cause high rate of deaths in every year. Early prediction of breast cancer plays a prominent role in medical field to overcome the patient deaths. The main aim of the paper is to predict the breast cancer by using Machine learning techniques. The Breast cancer dataset is taken from the UCI (University of California, Irvine) Repository which is having 32 attributes (i.e... ID number, Radius, Texture, Perimeter, Area and Smoothness etc). The data which is taken from UCI repository is preprocessed to remove the redundant data and replace the null values. The machine learning supervised algorithms SVM(Support Vector Machine) classifier, Random Forest, Gradient boosting, Naive Bayes, Cart Model, Neural Network and Linear Regression trained with preprocessed data to classify the patient's data into two categories Benign and Malignant here Malignant indicates which is a cancer and Benign which is non-cancerous. Among these Support Vector Machine gives good accuracy rate by comparing with other machine learning algorithms.

Keywords: Breast Cancer, Ensemble, Machine learning, Neural Network, Random Forest, SVM, Supervised.

1. Introduction

Breast cancer is one of the most severe cancer, has taken hundreds of thousands of lives every year. Early prediction of breast cancer plays a major role in successful treatment and save thousands of patient's lives every year. However, the conventional approaches limited in providing such capability. The recent breakthrough of machine learning and data mining techniques has opened a new door for health care diagnosis and prediction. By utilizing these advancements for predicting breast cancer risk based on a labeled dataset. Particularly, utilizing the clinical data as well as the associated symptoms of patients to develop predictive models which can classify patients into different breast cancer categories, i.e., benign or malignant.

Many people suffered from Breast Cancer. In general Breast cancer occurs mainly in women, but men can get it, too. Many people do not realize that men have breast cancer tissue and this can develop breast cancer. Breast cancer is starts when cells in the breast begin to grow out of control. Breast cancer develops in breast cells mostly in lobules or the duct of the breast.

Breast cancer is the most frequent diagnosed cancer in women and it is in second place in death rates after lung cancer. Cancer occurs when mutations takes place in cell that regulates cell growth. In early stages breast cancer may not produce any symptoms, depends upon the type of cancer, symptoms will vary. The advents of machine learning algorithms have opened a door for medical diagnosis and prediction. By utilizing the features of machine learning techniques to improve the prediction model for better prediction of breast cancer.

2. Literature Review

Classification is one of the most important and essential task in machine learning and data mining. Many researchers have conducted on different medical datasets to classify the type of cancer using data mining and machine learning approaches. Some of them show good classification accuracy.

Compare various machine learning algorithms to evaluate the prediction accuracy. Those Algorithms are Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN). Aims to evaluate the efficiency of each algorithm in terms of specificity, precision, sensitivity and accuracy [1].

Description of the practical application of data mining methods for assessment of survival rate and disease weakening for breast cancer patients. A relative study of Machine Learning models carried out [2].

The set of four biomarkers of Breast Tumors used to predict the Breast Cancer using Multi-layer perception. Predictive potential of markers are used to define the state of nodal involvement in Breast Cancer. Two methods of outcome evaluation like stratified and simple k-fold cross validation (CV) studied to assess their accuracy and reliability for neural network validation. Output accuracy, sensitivity and specificity used to select the best validation technique instead of evaluating the network outcome for different combinations of markers [3].

A system developed that can classify "Breast Cancer Disease" tumor using neural network with Feedforward, Backpropagation Algorithm to classify the tumor from a symptom that causes the breast cancer disease. Cost-effective and easy–to-use systems for supporting clinicians is developed using this model. Experimental indicated show that the concise models extracted from the network achieves high accuracy rate on the training data set and on the test dataset [4].

Early diagnosis needs an accurate diagnosis procedure used by physicians to classify whether the tumor is benign or malignant. To classify whether it is benign or malignant tumor, Classification algorithms used. Comparing the results of supervised learning classification algorithms and combination of these algorithms using voting classifier technique is done. Voting is one of the ensemble approaches where we can combine multiple models for the better classification [5].

Classify the breast tumor by SVM classifier. The image segmentation used to fragment the breast tissue corresponding to the tumor and the discrete wavelets transform (DWT) used as a feature extraction method to extract various features from the segmented images. SVM classifier to classify the breast tissue equivalent to the features and achieved an accuracy of 88.75% [6].

Pattern recognition method used to classify masses for micro calcification and abnormal severity such as benign or malignant from mammographic image. Wavelet analysis as a feature extraction technique and fuzzy-Neuro as a classifier to achieves a better classification rate [7].

The pattern recognition task used to predict the breast cancer, by considering Color Wavelet Features as feature extraction technique from segmented histopathology image. The SVM classifier used to build a breast cancer prediction model which gives an accuracy of 98.3% [8].

The main objective of this project is to develop an approach for modeling lung cancer survival based on Adaptive neuro-fuzzy inference System where we combine the reasoning capabilities of fuzzy logic and learning capabilities of neural network in order to give enhanced prediction capabilities, as compared to using a single methodology alone[9].

3. Proposed Methodology

The data present in the Wisconsin breast cancer dataset from the UCI Machine Learning Repository is in structured format. But before performing breast cancer prediction on the dataset by remove the null values and redundant attributes from the dataset for further processing. The following steps are involved in breast cancer Prediction.

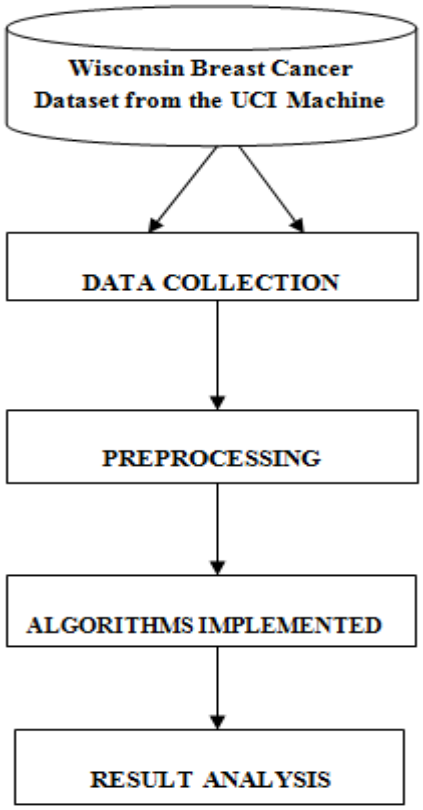

Fig.1. System Architecture

Data Collection

Data collected from Wisconsin breast cancer dataset from the UCI Machine Learning Repository. Dataset consists of 32 columns, with the first column being the ID number, the second column being the diagnosis result (benign or malignant), and followed by the mean, standard deviation and the mean of the worst measurements of ten features. The features together with description are listed in Table1.

Table.1. Description of ten features used in the dataset

Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	The total distance between the snake points constitutes the nuclear perimeter
Area	Number of pixel on the interior of the snake and adding one-half of the pixel in the perimeter
Smoothness	Local variation in radius length, quantified by measuring the difference between the length of a radial line and the mean length of lines surrounding it.

Compactness	Perimeter ^2 / area
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	The length difference between lines perpendicular to the major axis to the cell boundary in both directions.
Fractal dimension	Coastline approximation. A higher value corresponds to a less regular contour and thus to a higher probability of malignancy.

Data preprocessing

After data collected from the UCI Machine Learning Repository, need to pre-process the data. Here the dataset balanced but you need to remove Null values and redundant attributes from the dataset before applying algorithms. Some of such operations listed below

1) R function used to check and remove null values in the dataset.
2) Before applying methods on dataset we must remove the column i.e., ID_Number column from the dataset by using R function

Training and Testing Machine Learning Classifier

After selecting the features, by choosing a machine learning classifier for Breast cancer prediction. The data classified into train data and test data. Train data used to train the classifier and its performance measured by using test data.

In this paper by using Machine learning Techniques to predict the Breast cancer.

Machine learning is a part of Artificial Intelligence and it enables the Systems to learn themselves automatically and to improve their performance through experience without any instructions by programmer. Machine learning algorithms are not computer programs but it creates the rules. The basic of Machine learning process is to give training data to learning algorithm. The learning algorithm generates a new set of rules, based on conclusion of data. This is generating a new algorithm.
Machine learning algorithms categorized as three types

(i) Supervised learning,

(ii) Unsupervised learning

(iii) Reinforcement learning

This paper proposed supervised machine learning algorithms i.e., Support vector machine, Random Forest and Naive Bayes classifiers to classify the breast cancer. We use training dataset to build the model and validating on test dataset.

Naive Bayes is one of the supervised learning classifier and it uses the Bayes theorem to build the model. This algorithm can work efficiently for large data sets. Bayes theorem described by;

$$P(M/N) = (P(N/M)\ P(M)/P(N)$$

Random forest is a versatile simple to use machine learning algorithm. Random Forest creates a forest and makes it somehow random. Random Forest is an ensemble of decision trees. Random Forest algorithms are useful for both classification and regression. Gradient Boosting is one of the most powerful Machine learning Technique to build predictive model and it trains the model sequentially.

Support Vector Machine is mostly used for classification problems. This model uses nonlinear mapping to transform the original training data into higher dimension to search linear optimal separating hyper plane. This is also a supervised machine learning algorithm which is used for both classifications as well as regression challenges. Support Vector Machine classifier uses a kernel function to map a low dimensional feature space to a higher dimensional space to separate classes.

4. Results

Machine Learning algorithms are tested against the Wisconsin Breast Cancer data to find the performance of algorithms. The models have given best results to classify the disease and are given in table 2.

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 69  0
         M  2 42

               Accuracy : 0.9823
                 95% CI : (0.9375, 0.9978)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9625
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9718
         Pos Pred Value : 0.9545
         Neg Pred Value : 1.0000
             Prevalence : 0.3717
         Detection Rate : 0.3717
   Detection Prevalence : 0.3894
      Balanced Accuracy : 0.9859

       'Positive' Class : M
```

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 70  3
         M  1 39

               Accuracy : 0.9646
                 95% CI : (0.9118, 0.9903)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9235
 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.9286
            Specificity : 0.9859
         Pos Pred Value : 0.9750
         Neg Pred Value : 0.9589
             Prevalence : 0.3717
         Detection Rate : 0.3451
   Detection Prevalence : 0.3540
      Balanced Accuracy : 0.9572

       'Positive' Class : M
```

```
          Reference
Prediction  B  M
         B 68  2
         M  3 40

               Accuracy : 0.9558
                 95% CI : (0.8998, 0.9855)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9057
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9524
            Specificity : 0.9577
         Pos Pred Value : 0.9302
         Neg Pred Value : 0.9714
             Prevalence : 0.3717
         Detection Rate : 0.3540
   Detection Prevalence : 0.3805
      Balanced Accuracy : 0.9551

       'Positive' Class : M
```

Fig.2 (a). Confusion Matrix and Performance Metrics of SVM classifier, Random Forest, Gradient boosting

In Fig.2 (a) shows that Support vector machine, Random Forest and Gradient boosting algorithms gives the accuracies 98.23%, 96.46% and 95.58% respectively.

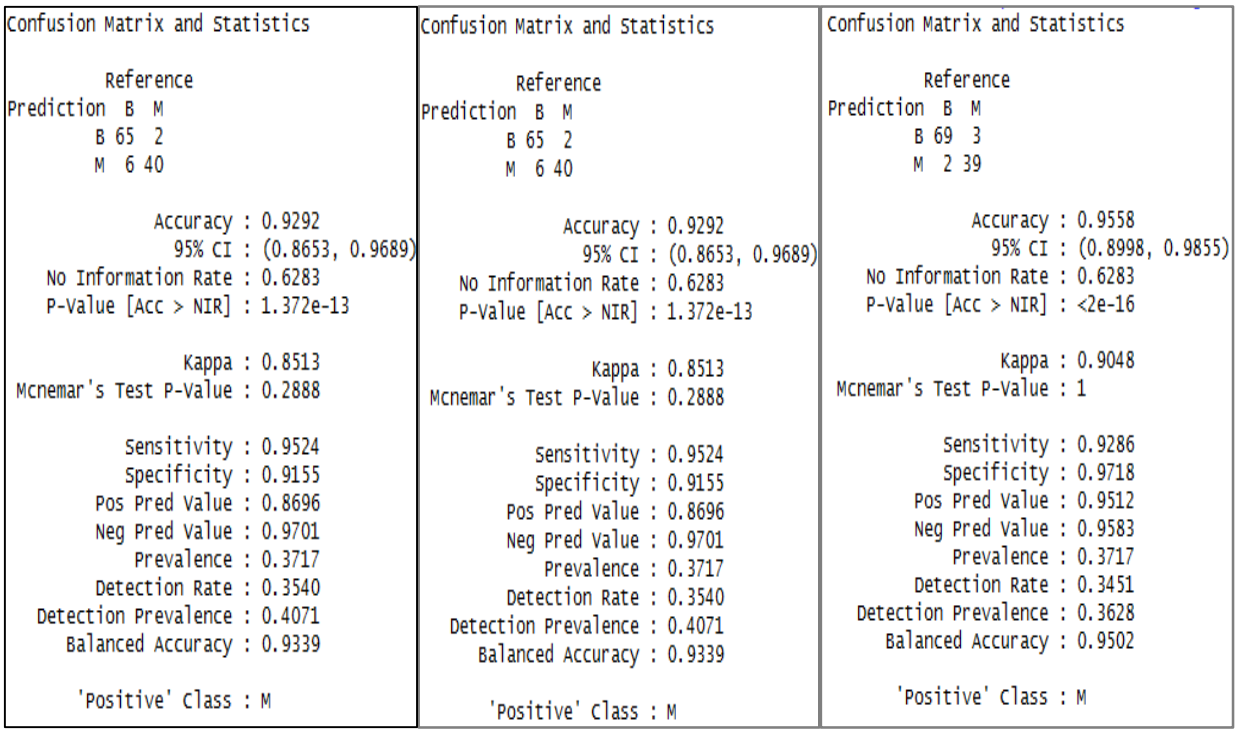

Fig.2 (b). Confusion Matrix and Performance Metrics of Naive Bayes, Cart Model, Neural Network

In Fig.2 (b) shows that Naïve Bayes, Cart Model and Neural Network algorithms gives the accuracies 92.92%, 92.92% and 95.58% respectively.

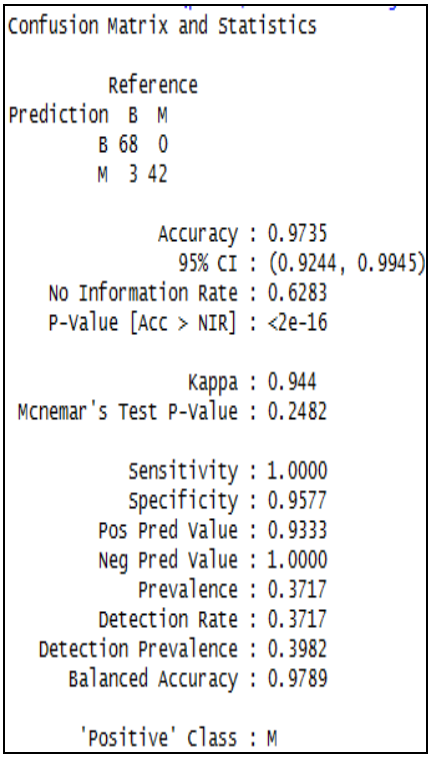

Fig.2(c). Confusion Matrix and Performance Metrics of Linear Regression

In Fig.2(c) shows that Confusion Matrix and Performance Metrics of Linear Regression algorithms with an accuracy 97.35%.

Figures.2 (a), (b), (c) shows the confusion matrix and Performance Metrics of the Support vector classifier, Random Forest, Gradient Boosting, Naive Bayes, Cart Model, Neural Network, and Linear Regression algorithms in that we clearly observe Support vector machine outperforms other algorithms

The formulae's for calculating sensitivity, specificity and accuracies of algorithms are shown as;

The sensitivity of the confusion matrix (Support Vector Machine)

$$\text{Sensitivity} = \frac{TRUE\ POSITIVES}{TRUE\ POSITIVES+FALSE\ NEGATIVES} = \frac{42}{42} = 1$$

Now specificity using confusion matrix (Support Vector Machine)

$$\text{Specificity} = \frac{TRUE\ NEGATIVES}{TRUE\ NEGATIVES+FALSE\ NEGATIVES} = \frac{69}{71} = 0.9718$$

Finally substitute both sensitivity and specificity to get the accuracy (Support Vector Machine)

Accuracy of Support vector Algorithm is

$$= SENSITIVITY * \left(\frac{POS}{POS+NEG}\right) + SPECIFICITY * \left(\frac{NEG}{POS+NEG}\right)$$

$$= 1 * \left(\frac{42}{113}\right) + 0.9718 * \left(\frac{71}{113}\right)$$

$$= 0.9823$$

Table.2. Performance of Algorithms

S. No	Methods	Sensitivity	Specificity	Accuracy
1.	Support Vector Machine	100	97.18	98.23
2.	Random Forest	92.86	98.59	96.46
3.	Gradient Boosting Tree	95.24	95.77	95.58
4.	Naive Bayes	95.24	91.55	92.92
5.	Cart Model	95.54	91.55	92.92
6.	Neural Networks	92.86	97.18	95.58

7.	Linear Regression	100	95.77	97.35

The Table.2 clearly indicates that Support Vector Machine had proven its performance given by more accurate results in classifying the cancer disease.

Figure.3. shows comparison of different algorithms along with its accurate results

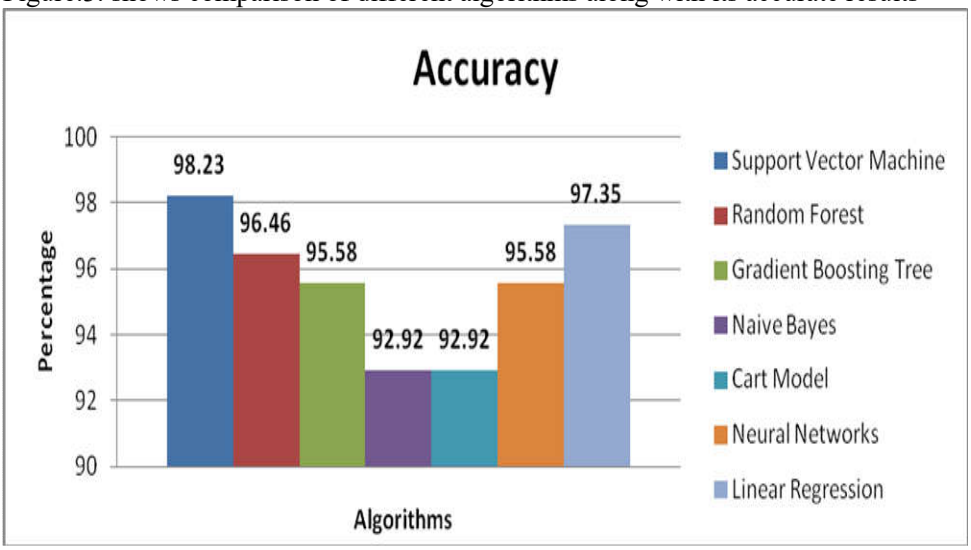

Fig.3. Comparison of different algorithms

In the Figure.3. Shows seven different algorithms in a bar graph representation in that clearly observe Support vector machine outperforms other algorithms

5. Conclusion

To analyze the medical data by various data mining and machine learning techniques are available. A major challenge in data mining and machine learning areas is to build an accurate and computationally efficient classifiers for Medical applications. During this study, by utilizing four main algorithms: Support vector classifier, Random Forest, Gradient Boosting, Naive Bayes, Cart Model, Neural Network and Linear Regression algorithm on the Wisconsin Breast Cancer (original) datasets. we tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. Support vector reaches the accuracy of 98.23% and outperforming other algorithms. In support vector has proven its efficiency in Breast Cancer prediction and diagnosis, it achieves the best performance in terms of precision and low error rate.

6. References

[1] HibaAsria, HajarMousannifb, Hassan Al Moatassime, Thomas Noeld, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis “ , The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016)

[2] Bojana R. AndjelkovicCirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D,” Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients”, Filipovic, Member, IEEE

[3] S.A. Mojarad, s.s.Dlay, W.L.Woo, and G.V. Sherbet”, Breast Cancer Prediction and Cross Validation Using Multilayer Perceptron Neural Networks”, 978-1-86135-369-6/101, 2010 IEEE

[4] Muhammad Sufyian Bin MohdAzmi, ZailaniAbuCobb, "Breast Cancer Prediction Based On Backpropagation Algorithm", 978-1-4244-8648-9/10/2010 IEEE.
[5] karthikkumar, sainikhilreddy, ksumangali, "Prediction of Breast Cancer using Voting Classifier Technique", 978-1-5090-5905-8/17/2017 IEEE
[6] Rejani YIA, Selvi ST (2009) Early detection of breast cancer using SVM classifier technique. International Journal on Computer Science and Engineering 1(3):127-130.
[7] Mousa R, Munib Q, Moussa A (2005) Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. Expert Systems with Applications 28(4):713-723.
[8] Mohammad R. Mohebian , Hamid R. Marateba , MarjanMansourian, Miguel Angel Mañanas, FariborzMokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning", M.R. Mohebian et al. / Computational and Structural Biotechnology Journal 15 (2017) 75–85
[9] Plachikkad.A. Rehana Bhadhrika1, Sri Kunjam Nageswara Rao , "Lung Cancer Survival Modelling using Adaptive Neuro-fuzzy Inference System", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015.
[10] Uma Ojha, Dr. SavitaGoel," A study on prediction of Breast Cancer recurrence using Data Mining Techniques", 978-1-5090-3519-9/17/2017 IEEE
[11] Young Ryu,Chandrasekaran, Varghese, Jacob, "Breast cancer prediction using the isotonic separation technique", Y.U. Ryu et al. / European Journal of Operational Research 181 (2007) 842–854
[12] Ismail Saritas, "Prediction of Breast Cancer Using Artificial Neural Networks", Springer Science+Business Media, LLC 2011.
[13] SoltaniSarvestani, A. A. Safavi, N.M. Paraneh, M.Salehi,"Predicting Breast Cancer Survivability Using Data Mining Techniques", 978-1-4244-8666-3/10/2010 IEEE.
[14] RunjieShen, Yuanyuan Yang, Fengfeng Shao, "Intelligent Breast Cancer Prediction Model Using Data Mining Techniques",978-1-4799-4955-7/14 2014 IEEE.
[15] Ir CATH Tee, Ali H. Gazala, "A Novel Breast Cancer Prediction System", 978-1-61284-922-5/11/2011 IEEE.
[16] Tuan Tran and Uyen Le, "Predicting Breast Cancer Risk: A Data Mining Approach", 978-981-10-4361-1_37.
[17] Deepika Verma and Nidhi Mishra, "Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques" , 978-1- 5386-1959-9/17/$31.00 ©2017 IEEE
[18] Gopal K. Dhondalay, Dong L. Tong and Graham R. Ball, "Estrogen receptor status prediction for breast cancer using artificial neural network", 978-1-4577-0308- 9/11/$26.00 © 2011 IEEE
[19] Rashmi G D, A Lekha and Neelam Bawane," Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset", 978-1-4673-9563-2/15/$31.00©2015 IEEE
[20] Shabina Sayed, Shoeb Ahmed and Rakesh Poonia," Holo entropy enabled decision tree classifier for breast cancer diagnosis using wisconsin (prognostic) data set", 978-1-5386-1860-8/17/$31.00 ©2017 IEEE
[21] Yifan Zhang, William Yang and Dan Li ," Toward precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis" 978-1-5090-3050-7/17/$31.00 ©2017 IEEE
[22] Dongdong Sun , Minghui Wang and Huanqing Feng, "Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: Supervised feature extraction and classification for breast cancer prognosis prediction", 978-1-5386-1937-7/17/$31.00 ©2017 IEEE.