An Efficient Cancer Prediction System using Ensemble Methods

Sopna P, Sowmiya E, Sanjana M, Sujatha R

Abstract*: *Breast cancer is the most dreadful disease in the world in past few decades. Many women in the world has been affected due to this horrible disease and died. Breast cancer occurs in breast cells, the fatty tissue or the fibrous connective tissue in the breast. Breast cancer is malignant tumors tend to become progressively worse leading to death. Factors such as age genetic mutations and a family's reordered history in breast cancer can increase the risk of breast cancer. Two types of tumors: Benign: this tumor type is not dangerous for a human body and rarely causes human death. Malignant: this tumor type is more dangerous and causes human death, it is called breast cancer. Machine learning was the boon technique in the fields of the medical industry. By the development of machine learning and data analytics a decision making tool can be made which helps in early detection and diagnosis of cancer tumor in women. This concept is to study and develop a decision based tool to eradicate breast cancer. The prediction system makes use of the ensemble algorithms to detect the cancer at earlier stage. It also differentiates the type of cancer from which the patient is being affected with effective accuracy.

KEYWORDS: Decision Support Tools, Machine Learning, Prognosis, Diagnosis.

I. INTRODUCTION

According to World Health Organization (WHO) cancer is the most tragic disease in the world in which an abnormal cell invades the part of a body and made it to become failure. It is a group of disease under which several abnormal cell causes different cancers. Breast cancer is one of the major kind of cancer which has a high mortality rate than other cancer cells. As of statistics around the world 10 - 15% of the patients with breast cancer where reported to die before the deductions and treatments. Day by day the risk of life of breast cancer patients increased although several deductions and treatment technologies were implemented. Because all these treatment technologies were less accurate and inappropriate. So it doesn't work on all kind of patients since each patient's intensity of tumors varies from stage to stage. Breast cancer is the most dreadful disease in the world in past few decades. Many women in the world has been affected due to this horrible disease and died. Many image processing techniques are made in the medical industry for early detection of tumors in breast cancer. In the past few decades several algorithms using machine learning and data mining are developed which are highly accurate in detecting the tumors of breast cancer. The detection process involves three stages preprocessing, feature extraction, and classification. The preprocessing is the first stage which helps in improving the visibility and makes us to know about the intensity of the tumor. Feature extraction is the most important step which helps to distinguish between benign and malignant tumors. After these stages based on the extraction the classification was made.

II. RELATED WORKS

Different transform-based texture analysis techniques are applied and studied to convert the image into a replacement kind victimization the abstraction frequency properties. The normally used techniques involve riffle remodel, quick Fourier remodel (FFT), Dennis Gabor transforms, and singular price decomposition (SVD) to scale back the spatiality of the feature illustration, principal part analysis (PCA) will be used. Several works have tried to automatic designation of carcinoma supported machine learning algorithms and computer science. As an example, Malek et al. [3] suggested riffle technique for options extraction and mathematical logic for classification. Sun et al. [4] analyzed the matter by scrutiny options choice strategies, whereas Zheng et al combined K-means algorithmic rule and a support vector machine (SVM) for carcinoma designation and classification. Many works supported agglomeration and classification are conducted and studied. Another approach, introduced by Aličković and Subasi, used a genetic algorithmic rule for feature extraction and rotation forest as a classifier. Finally, a recent work by Bannaie was supported the improved resonance imaging (DCE-MRI) technique to get relevant and economical data. The contribution of the authors of this paper focuses on the preprocessing stage. Strategies delineated within the literature for carcinoma designation are thought-about as semi-automatic strategies.

III. EXISTING SYSTEM

The existing system diagnoses the prostate cancer using neural networks which recognizes patterns. Initially, the dataset for the prostate cancer is collected from various patients and preprocessing is done. The noise and bias is identified and removed using mean-mode normalization and roughest algorithm is used to select the most optimized features using the probability .It is done using the radbas initiation work. The elements thus obtained are used to highlight the learning neural systems. Finally, the efficiency of the proposed system is evaluated using the results which are bases on mean square mistake rate and exactness. The framework shows 99.3% exactness which is the highest one. This exactness was examined using the training and testing prostrate biomedical data examination process.

Revised Manuscript Received on July 22, 2019.

***Sopna P**, Department of Computer Science & Engineering,Sri Krishna College of Technology, Coimbatore, India. Email: 16tucs223@skct.edu.in

Sowmiya E, Department of Computer Science & Engineering,Sri Krishna College of Technology, Coimbatore, India. Email: 16tucs224@skct.edu.in

Sanjana M, Department of Computer Science & Engineering,Sri Krishna College of Technology, Coimbatore, India. Email: 16tucs210@skct.edu.in

Sujatha R, Department of Computer Science & Engineering, Sri Krishna College of Technology,Coimbatore,India. Email: r.sujatha@skct.edu.in

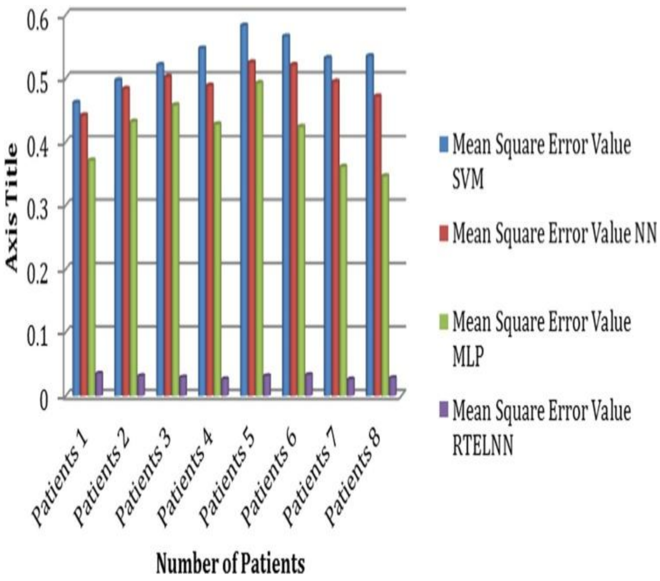


Fig.3.1 Accuracy with training set

IV. PROPOSED SYSTEM

The theme of the proposed system is to make sure that a model need to be developed on the basis of machine learning to diagnosis the breast cancer also to ensure such models that can be treated for breast cancer. The ensemble learning is used for the optimization of the algorithm by combining multiple machine learning algorithms into a single model. This model uses the feature selection algorithms such as recursive feature selection to arrive at the best features of the model and these algorithms are used to detect whether it is B(Benign) or M(Malignant).

V. MODULES

Fig 5.1, explains about the block diagram of proposed system. In that certain classification algorithms are applied after the preprocessing techniques. After getting the analysis report, final predicted output will be provided as a result.

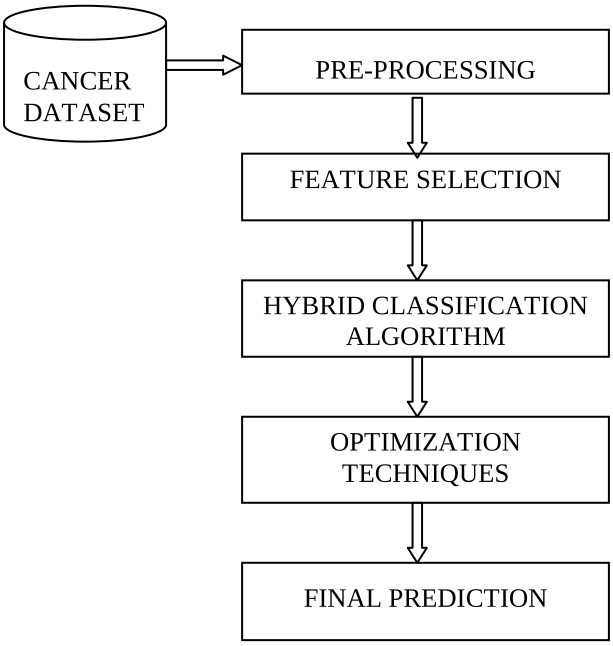


Fig.5.1 Block Diagram

A. DATA COLLECTION

Public dataset. Source: UCI dataset of breast cancer detection. 570 rows and 32 attributes. Target variable: Two (B or M). Format: CSV. 31 numerical variable (independent variable) and one string variable (target variable).

B. PRE – PROCESSING

Data contains corrupted original data and the missing values like observations of process that were not recorded. Handling missing data is more important and necessary as many machine learning algorithms do not support data with missing values. Handling the missing values is one of the greatest challenges, because making the right decision on how to handle it. Fig 5.2.shows the checking of missing values and fig 5.3, fig 5.4 and fig 5.5 are showing the Diagnosis and distribution of a variable and correlation of a variable.

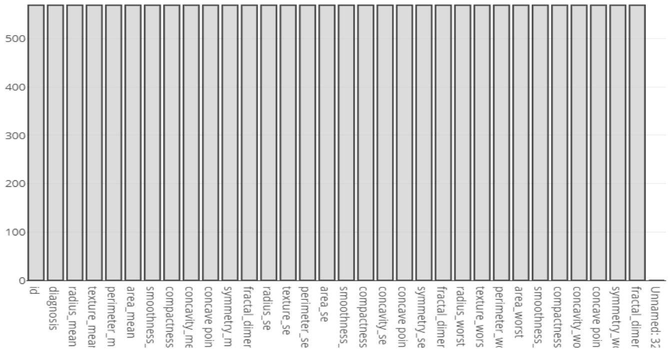


Fig.5.2 Missing values –check

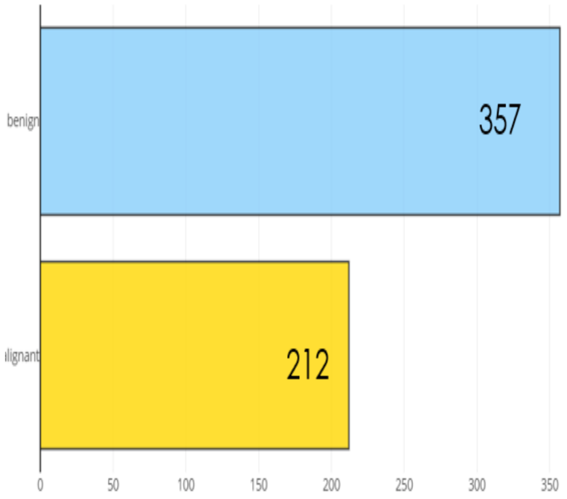


Fig.5.3 Count of diagnosis variable

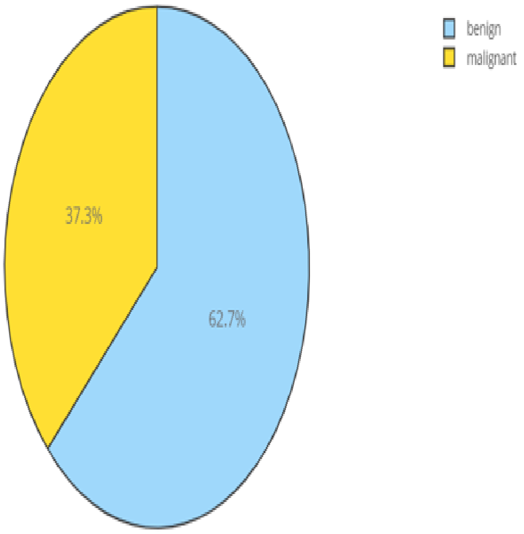


Fig.5.4 Distribution of variable

C. CORRELATION BETWEEN VARIABLES

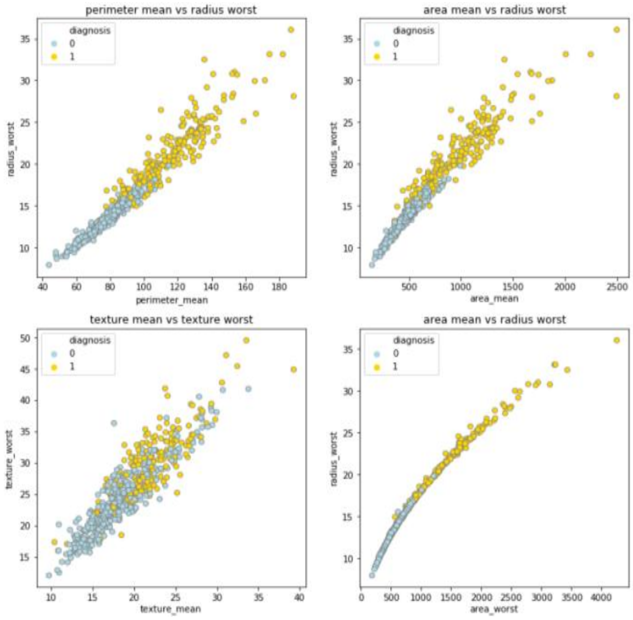


Fig.5.5 Correlation between variable

D. PRINCIPLE COMPONENT ANALYSIS

The third step of preprocessing involves the principal component analysis which leads to a new set of variables. These newly transformed variables are referred as principal components. These components are orthogonal where the retention of variation decreases on moving down. Hence the component at the beginning retains the maximum variation. They are the eigenvectors of covariance matrix and so called orthogonal.

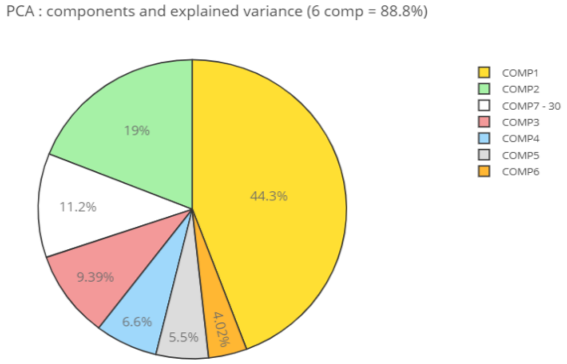


Fig.5.6 PCA components

VI. IMPLEMENTATION

A. LOGISTIC REGRESSION

Logistic regression is one of the machine learning algorithm. The probability of a categorical dependant variable is calculated by logical regression. The dependent variable will be coded as 1 or 0 in logistic regression. And the variable 1 denotes yes and success. Variable 0 denotes no and failure. Grid-search finds the optimal hyper parameters of a model which results in the most 'accurate' predictions. Grid search builds a model for every combination of hyper parameters and evaluates each model.

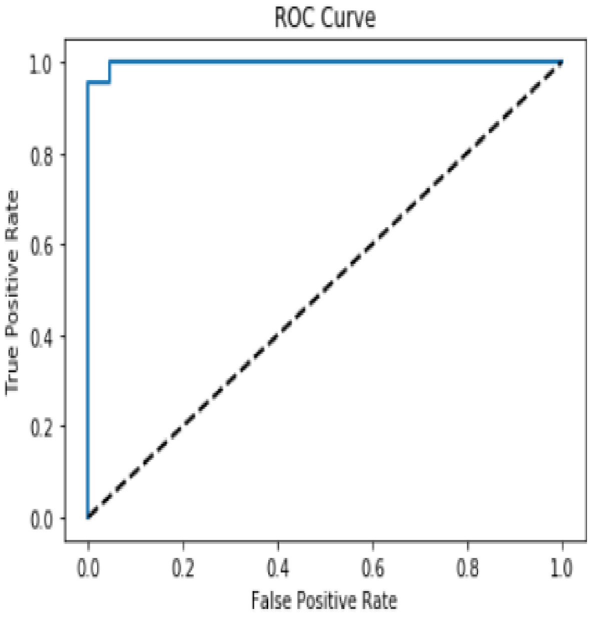


Fig.6.1 ROC values

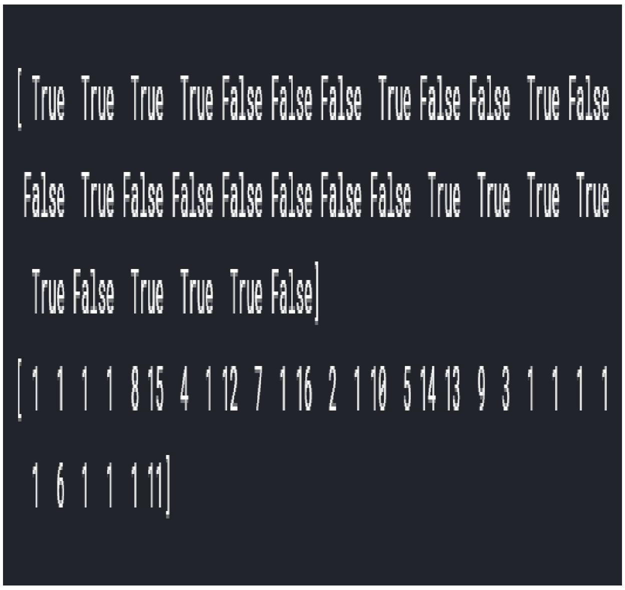
B. RECURSIVE FEATURES ELIMINATION

Recursive Feature Selection (RFE) is a feature selection method that is used to fit the model and eliminates the least important features until the specified number of important features are reached. Features are grouped by feature specification and build a model for computation purpose.

Retrieval Number: D7682049420/2020©BEIESP
DOI: 10.35940/ijeat.D7628.049420

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

In the results the least ranked features gets removed and higher ranked features gets grouped. After grouping, a model is rebuilt and computation is made. The results are stored as predictor subset which acts as a reference. These references help in tuning the RFE. The final model will be trained by using the optimal subset. The selection system is resampled, together with the quantity of the nearest associates or the quantity of weight decay in a neural network. The resampling system includes the function choice recurring and the outside resamples are used to estimate the correct subset size.



```
[ True  True  True  True False False False  True False False  True False
 False  True False False False False False False  True  True  True  True
  True False  True  True  True False]
[ 1  1  1  1  8 15  4  1 12  7  1 16  2  1 10  5 14 13  9  3  1  1  1  1
  1  6  1  1  1 11]
```

Fig.6.2 Result of Recursive features elimination

C. VOTING AND RESULTS

Voting is one of the ensemble algorithms which involves combining multiple models into a single one. The voting classifier averages the predictions of multiple sub-models to make the predictions of new data. The final output of the predictions is identified with one that receives more than half votes. If it doesn't happen then the ensemble method is determined to not make suitable prediction .

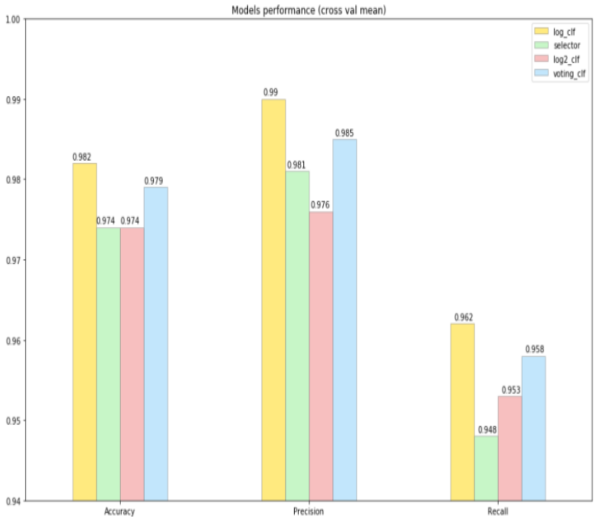


Fig.6.3 Models Performance Accuracy, Precision, Recall

VII. CONCLUSION

Machine learning is one of the most advanced area in computer engineering helps in several modern medical technologies. As the world knows, breast cancer is the deadliest cancer with high mortality rate, to treat that we already have several methods, but this ensemble algorithm helps in early detection and treatment of breast cancer tumors by distinguishing them between Benign and Malignant. The paper attempts to explain, compare and assess the performance of different machine learning algorithms that are being applied to cancer prediction and prognosis. Our approach utilizes ensemble feature selection method to improve the classification accuracy and reliability to deliver a prediction system that can be used for future diagnosis with increased accuracy level.

REFERENCES

1. Conference on Machine Learning, volume 96, pages 148–156, 1996.
2. D.H. Wolpert. Stacked generalization. Neural Networks, 5(2):241–259, 1992.
3. C. M. Bishop. Pattern Recognition and Machine Learning, chapter 1.3. Springer Science+Business Media, 2006.
4. L.Breiman. Random forest. Machine Learning, 45:5–32, 2001.http://scikit-learn.org/stable/.

AUTHORS PROFILE

Sopna P doing her Bachelor of Engineering in Sri Krishna College of Technology. Her area of interests is Neural Networks and Machine Learning. She published her papers in various journals.

Sowmiya E doing her Bachelor of Engineering in Sri Krishna College of Technology. Her area of interests are Neural Networks and Machine Learning. She published her papers in various journals.

Sanjana M doing her Bachelor of Engineering in Sri Krishna College of Technology. Her area of interests is Neural Networks and Machine Learning. She published her papers in various journals.

Ms. Sujatha R, B.Tech(IT), M.E(CSE), working as an Assistant Professor in the Department of Computer Science and Engineering at Sri Krishna College of Technology, Coimbatore. She is pursuing her research work on Deep Learning using Convolution Neural Network in agriculture. She has published 3 Scopus Indexed Journals and 2 IEEE journals. She is also a Member in IAENG and CSIR.