Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm

Dr.M.Vasumathi Devi[1], M.Sravani[2], K.Ramya[3], N.Bindulakshmisai[4], V.Parameshwari[5], Department of Computer Science and Engineering, Vignan's Nirula Institute of Technology and Science for Women, Pedapalakaluru, AP, India.

ABSTRACT

According to Breast Cancer Institute (BCI), Breast Cancer is one of the most dangerous types of diseases that are very effective for women in the world. As per clinical expert detecting this cancer in its first stage helps in saving lives. As per cancer.net offers individualized guides for more than 120 types of cancer and related hereditary syndromes. For detecting breast cancer mostly machine learning techniques are used. In this paper we proposed adaptive ensemble voting method for diagnosed breast cancer using Wisconsin Breast Cancer database. The aim of this work is to compare and explain how ANN and logistic algorithm provide better solution when its work with ensemble machine learning algorithms for diagnosing breast cancer even the variables are reduced. In this paper we used the Wisconsin Diagnosis Breast Cancer dataset. When compared to related work from the literature. It is shown that the ANN approach with logistic algorithm is achieved 98.50% accuracy from another machine learning algorithm.

Keyboards: Breast Cancer, Neural Network, Logistic, Machine Learning Algorithm, WDBC Dataset.

I. INTRODUCTION

The most dangerous disease in the world is cancer in which breast cancer is the dangerous for women. Many women die every year because of breast cancer. Detecting the breast cancer manually takes a lot of time and it is difficult for the physician to classification. So the detecting the cancer through various automatic diagnostic techniques is very necessary. There are various method and algorithm are available for detecting breast cancer such as Support Vector Machine, Naïve Bayes, KNN and Convolution Neural Network is the latest algorithm in deep learning that is also used for classification. CNN and deep learning algorithm mainly used for images classification and object detection. In this paper we use UCI open database for training and testing purpose in which two classes of Tumor are available, one is Benign Tumor and the other is malignant in which benign Tumor is non-cancerous and the malignant is a cancer Tumor. Many reasecher are still performing research for detecting and diagnosing cancer in an early stage. Because the early stage cancer is not a so panful and expensive for complete its treatment and many researcher are still trying to developing a proper diagnosis system for detection the Tumor as early as possible. So the treatment can be started earlier and the rate for resolution may increase. This work main aimis comparatively study of various machines learning algorithm with Artificial Neural Network.

The rest of the paper is organized as follows: Section 2 presents the literature review of the proposed work. Section 3 includes the architecture of the proposed work. Section 4 describes the methodology which is used for the proposed work. Section 5 describes feature selection process for the proposed work. In section 6 we discuss the model implementation of proposed work. Section 7 discusses the results and Section 8 concludes the proposed work.

II. RELATED WORK

Many machine learning algorithms are available for prediction and diagnosis of breast cancer. Some of the machine learning algorithm are Artificial Neural Network (ANN), Naïve Bayes, Support Vector Machine (SVM), K- Nearest Neighbour (KNN) and Convolutional Neural Network etc. Many researcher have done research in breast cancer by sing various dataset suchas using Mammogram images as dataset, SEER dataset, Wisconsin Dataset and also dataset from various hospitals. By using these dataset authors extract and select various features and complete your research. These are some important research review.

1. The Author Moh'dRasoul used DWT tool for image filtering and BPNN for processing and achieved 93.7% accuracy
2. The author Clemen Deng used WHAVE algorithm with Wisconsin Breast Cancer Database and achieved 99% accuracy
3. The author Ashwaq Qasem used marker Controller Watershed algorithm and achieved 95% accuracy
4. The author AlirezaIsareh used signal to noise ratio with SVM and achieved 98.80% accuracy [4]. The author Junaid Ahmad used Adaptive Resonance Theory with UCI database and got 84.21% accuracy
5. By B.M. Gayathri work on comparative study about Relevance vector machine and achieved 97% accuracy
6. The author Ms.H.R. Mhaske used KNN and SVN with 150 images database and achieved 8090% accuracy
7. The author S. Aruna used Naïve Bayes and SVM with UCI database and achieved 68-79% accuracy
8. By Yohannes Tsehay developed a weakly supervied computer aided detection system that was used biopsy for learn data
9. By Sudarshan Nayak used Naïve BayesandSVMandgot98%accuracy
10. The Ryota Shimizu used Deep learning and neural network and achieved 90% accuracy
11. Many other authors also did research for detection and diagnosis breast cancer using various machines learning algorithm.

III. PROCESS FLOW DIAGRAM

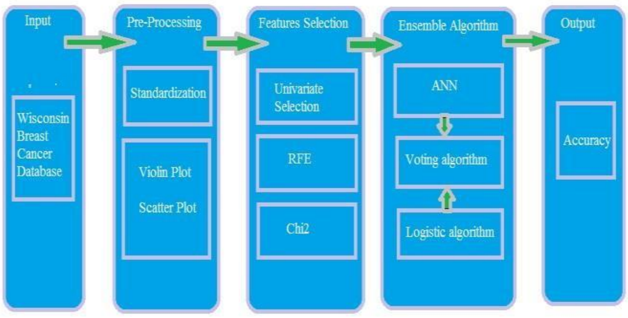


Fig. 1. Process Flow Diagram

Fig1 shows the process flow diagram or proposed work. First we collected the Wisconsin Breast Cancer Database from UCI website then pre-processed the dataset and select 16 important features. For feature selection we used Recursive feature Elimination Algorithm using Chi2 method and get 16 top features. After that applied ANN and Logistic algorithm individually and compute the accuracy. Finally, we used proposed Ensemble Voting method and compute best method for diagnosis breast cancer disease.

IV. EXPERIMENT AND METHODOLOGY

In this paper we have using ensemble method for diagnosis breast cancer with neural network and logistic algorithm. All process consist main three parts: Pre-processing data, features selection and voting models. In this work we have used BCI dataset having 569 rows and 30 column of dataset. In experiment part we have first evaluated the features from default dataset. For features selection we have used Univariate Features selection method and Recursive Features Selection method with Cross ValidationMethod.

A. Pre ProcessingData

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking in certain behaviours and likely to contain many errors. Data pre- processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. For pre- processing we have used standardization method to pre-process the UCI dataset.

B. Standardization Method:

In this method the dataset is a common requirement for many machine learning estimators. In this paper we have created different data visualization for data pre-processing. First we have count malignant and benign from all dataset and plot in graph format. The Fig 2 shown the total malignant and benign for breast cancer diagnosis from UCI dataset.

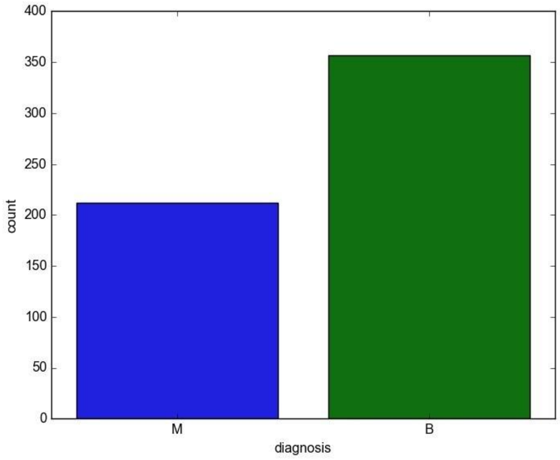


Fig. 2. Number of Malignant and Benign

In second stage we have created a Violin plot for dataset compression. It shows the distribution of quantitative data across several levels of one categorical variables such that those distribution can be compared. In proposed work we have created top 16 feature violin plot for comparison. The fig 3 shows the comparison of features.

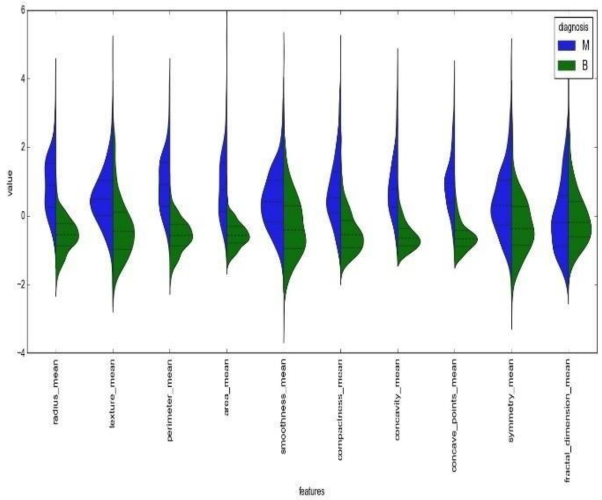


Fig. 3. Comparison of top 10 Features

Then we draw a scatter plot with non-overlapping points. This gives a better representation of the distribution of values. The fig 4 shows the scatter plot of top 16 features cancer dataset. Univariate feature selection algorithm used chi2 method for computing chi-square stats between each non-negative features and classes.

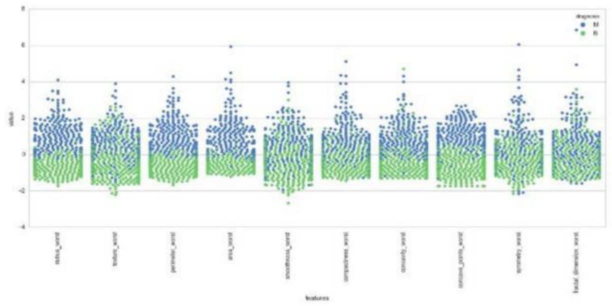


Fig. 4. Scatter plot with non-overlapping points

Here we have created a relationship scatter plot between dataset features for more understanding. The fig 5 showed the relationship scatter plot of between dataset features.

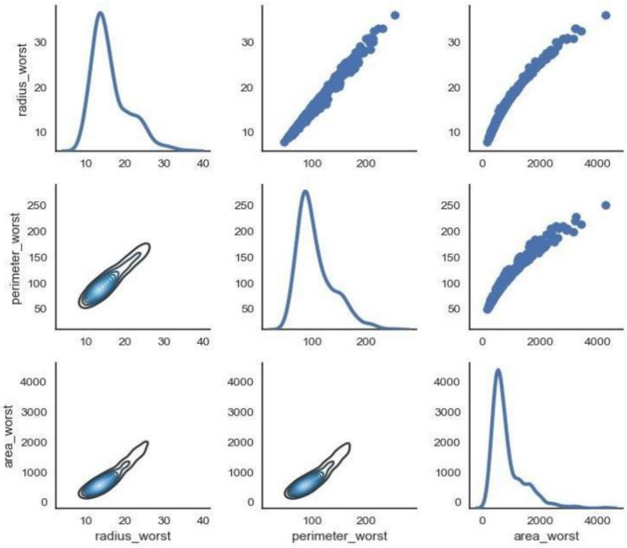


Fig. 5. Relationship scatter plot between dataset

The subscript "c" are the degrees of freedom. "O" is your observed value and "E" is your expected value.

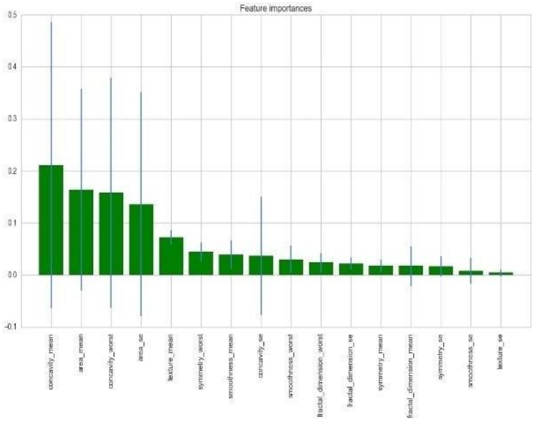


Fig 6. Top 16 Features Using Univariate Feature Selection

Then we have applied recursive feature elimination algorithm with cross validation on 16 top important features from UCI Dataset shows in Fig 7.

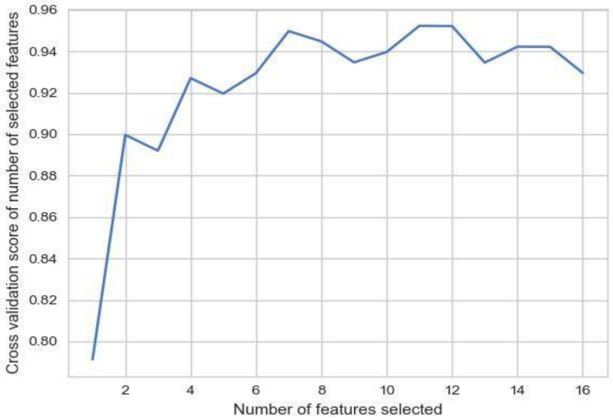


Fig. 7. Top 16 Features using Cross Validation

V. FEATURES SELECTION PROCESS

In machine learning and statistics, feature selection, also known as variable selection, attribute selection, is the process of selection a subset of relevant features for use in model construction. There are many methods for feature selection, we are using univariate feature selection method in this paper.

a. Univariate Feature Selection

In proposed work we have used Univariate feature selection method for examines each feature individually to determine the strength of the relationship of feature with the response variable. This method are simple to run and understand are in general particularly good for gaining a better understanding of data. After run this method we have got 16 top features for breast cancer diagnosis. Fig 6 shows all 16 features for breast.

b. Neural Network

In proposed method we have used neural network with the logistic algorithm. Both algorithm provide individual accuracy of UCI dataset then we have applied voting on both algorithm result. In proposed method we have used following parameters for neural network implementation.

- Lbfgs: It is optimizer in the family of quasi-Newton methods.
- Hidden layer: we have used 15 neurons in hidden layer.
- Activation Relu: The rectified linear unit functions.

c. Performance Evaluation Parameters

The following evaluation parameters used

- False Positive (FP): An input without breast cancer is incorrectly diagnosed as havingcancer.
- False Negative (FN): An input with breast cancer is incorrectly diagnosed as having nocancer.
- True Positive (TP): Its means patient having a breast cancer.
- True Negative (TN): Its means patient having no cancer.

d. Logistic Regression

The logistic regression formula is derived from the standard linear equation for a straight line. The standard linear formula is transformed to the logistic regression formula.

VI. MODEL IMPLEMENTATION

Ensemble algorithm for combine these results and an compute the final accuracy.

f (z)=1+ $e-z$ (2)

Precision: Precision is the number of correct results divided by the number of all returned results.

□□

Precision=__________________ (3)

□□+□□

Recall: Recall is the number of correct results divided by the number of results that should has beenreturned.

□□

Recall=__________________ (4)

□□+□□

F1-Score: A measure that combines precision and recall is the harmonic mean of precision andrecall.

□□□□□□□□□* □□□□□□

F =2*__________________________ (5)

□□□□□□□□□+□□□□□□

In this stage we have first implement logistic algorithm on these dataset and the implement Neural Network algorithm individual then we are implement Voting

Accuracy:

□□+□□

Accuracy=__________________ (6)
□□+□□+□□+□□

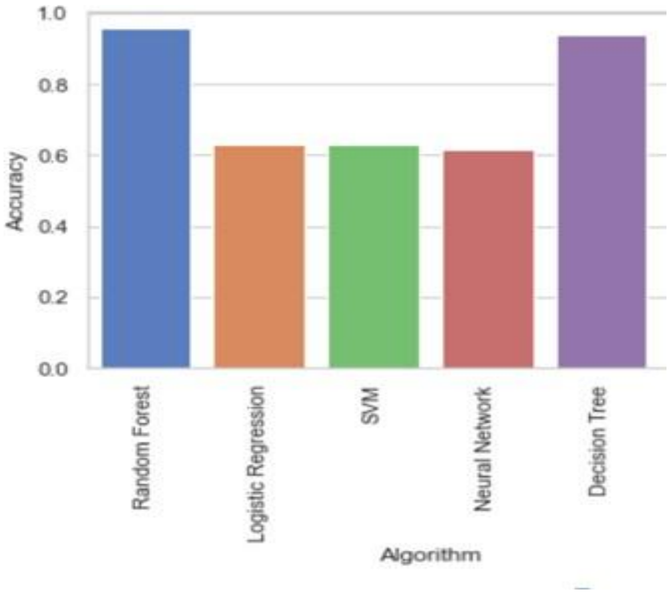
TABLE I. CLASSIFICATION REPORTFOR VOTING ALGORITHM

	precision	recall	f1-score	support
Benign	0.99	0.99	0.99	74
Malignant	0.98	0.97	0.97	40
Avg / total	0.98	0.98	0.98	114

The logistic regression function is useful for predicting the class a binomial target feature.

VII. RESULTS AND DISCUSSION

In this paper we proposed Ensemble Machine Learning algorithm with Logistic and Neural Network for diagnosis and detection of breast cancer. We have used standardization method for pre-processing breast cancer dataset then we have applied Univariate Features Selection algorithm. Univariate Feature Selection algorithm used chi2 method for selection Best 16 Features from UCI dataset. After collect final 16 features from univariate Feature Selection algorithm we implement logistic and neural network algorithm on these 16 features and final applied voting algorithm on result and achieved 98.50% accuracy. Wisconsin Breast Cancer Dataset have contain 699 rows with features categories 30 features. After applied Univariate Feature Selection method top 16 features are decided from final model implementation. Because large features are effect on cost of model implementation. Achieved accuracy is good from individual achieved accuracy from both machine learning algorithm.



VII. CONCLUSION AND FUTURE WORK

This work is the proposed an ensemble machine learning method for diagnosis breast cancer, in which we can see in the table and graph that proposed method is showing with the 98.50% accuracy. In this paper we used only 16 features for diagnosis of cancer. In future we will try on all features of UCI and to achieve best accuracy. Our work proved that neural network is also effective for human vital data analyzation and we can do pre-diagnosis without any special medical knowledge.

REFERENCES

1. M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp.35-39.
2. C. Deng and M. Perkowski, "A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection," 2015 IEEE International Symposium on Multiple-Valued Logic, Waterloo, ON, 2015, pp.115-120.
3. A. Qasem et al., "Breast cancer mass localization based on machine learning," 2014 IEEE 10th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, 2014, pp. 31-36.
4. A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," 2010 5th International Symposium on Health Informatics and Bioinformatics, Antalya, 2010, pp. 114- 120.
5. J. A. Bhat, V. George and B. Malik, "Cloud Computing with Machine Learning Could Help Us in the Early Diagnosis of Breast Cancer," 2015 Second International Conference on Advances in Computing and Communication Engineering, Dehradun, 2015, pp.644-648.
6. B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp.1-5.
7. H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Bengaluru, 2013, pp.1-5.
8. S. Aruna, S. P. Rajagopalan and L. V. Nandakishore, "An algorithm proposed for Semi-Supervised learning in cancer detection," International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011), Chennai, 2011, pp. 860- 864.
9. Y. Tsehay et al., "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, 2017, pp.642-645.
10. S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona,2017,pp. 13-14.
11. doi: 10.1109/CEM.2017.7991863
12. S. Mythili and A. V. S. Kumar, "CTCHABC- hybrid online sequential fuzzy Extreme Kernel learning method for detection of Breast Cancer with hierarchical Artificial Bee," 2015 IEEE International Advance Computing Conference (IACC), Banglore, 2015, pp.343-348.
13. R. Shimizu et al., "Deep learning application trial to lung cancer diagnosis for medical sensor systems," 2016 International SoC Design Conference (ISOCC), Jeju, 2016, pp.191-192.
14. S. Kim, S. Jung, Y. Park, J. Lee and J. Park, "Effective liver cancer diagnosis method based on machine learning algorithm," 2014 7th International Conference on Biomedical Engineering and Informatics, Dalian, 2014, pp.714-718.
15. S. J. Savari Antony and S. Ravi, "Development of efficient image quarrying technique for Mammographic image classification to detect breast cancer with supervised learning algorithm," 2013 International Conference on Advanced Computing and Communication Systems, Coimbatore, 2013, pp.1-7.
16. S. K. Wajid, A. Hussain, K. Huang and W. Boulila, "Lung cancer detection using Local Energy-based Shape Histogram (LESH) feature extraction and cognitive machine learning techniques," 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), Palo Alto, CA, 2016, pp.359-366.