Breast Cancer Detection using Data Mining and Machine Learning.

Neel Narvankar
Department of Information Technology
Atharva College of Engineering,
Malad (West), Mumbai, India.

Aseema Ajgaonkar
Department of Information Technology
Atharva College of Engineering,
Malad (West), Mumbai,India

Meet Lad
Department of Information Technology
Atharva College of Engineering,
Malad (West), Mumbai, India.

Raj Dedania
Department of Information Technology
Atharva College of Engineering,
Malad (West), Mumbai, India.

Renuka Nagpure
Assistant Professor, Department of Information Technology,
Atharva College of Engineering,
Malad (West), Mumbai, India.

***Abstract* - Breast cancer is seen mostly in women but is not limited to women. Breast cancer is one of the major cause of death among women in India, if cancer can be detected at an early stage it will prove to be advantageous since proper actions and treatment can be done to eliminate it at an early stage. Machine learning and Data mining techniques prove to be very helpful in predicting cancer at an early stage which might be a difficult problem otherwise. Machine learning and Data mining has revolutionized the entire process of how breast cancer can be detected. The main objective of this research is to find out cancer at an early stage by the help of Machine learning and Data mining. The paper also includes a study of currently used systems for a better understanding of the techniques used for detection.**

Keywords:- Breast Cancer, Data Mining, Machine Learning.

I. INTRODUCTION

Breast Cancer is the second most common kind of cancer found in women compared to other cancers. In India, 1 out of 28 women are prone to breast cancer during her lifetime. Women aging from 43 to 46 are prone to breast cancer however, younger women are also detected with this disease. Early detection of this disease can increase the chances of survival of patients. Cancer now causes more deaths than all coronary heart disease or all stroke according to WHO estimates for 2011.

India has a growing rate of breast cancer and statistic show that if the rate continues to increase,by 2030 the number of new cases of breast cancer in India will increase [1] to 200,000 which is a lot.This paper presents a more efficient model of detecting breast cancer so that it can be detected and necessary actions for treatment can be taken.

The tests which are conducted to detect breast cancer can be very expensive and may sometimes incur human errors by doctor. In the proposed system, we detect breast cancer by conducting a test and increasing the accuracy thereby providing a more efficient solution.The process starts by taking X-ray of patient , this X-ray goes through a process of segmentation and adaptive median filters to get the necessary attributes from the input which is the X-ray.The data thus obtained is then run through various classification algorithms which compares the newly obtained data with a pre trained data set,thus classifying the data as benign or malignant.

II. EXISTING SYSTEMS

In [U.Ojha, S.Goel IEEE 2017] [1], data mining algorithms are used to predict all those cases of breast cancer that are recurrent using Wisconsin Prognostic Breast Cancer (WPBC) data-set from the UCI machine learning repository. Then, clustering and classification algorithms are used to find their performance of these models. The clustering algorithms used were K means, EM, PAM and Fuzzy c-means while the four classification algorithms are SVM, C5.0, Naive bayes and KNN.

[Deepika Verma; Nidhi Mishra IEEE 2017] [2], the two disease data-set from UCI machine learning repository are taken on WEKA tool for the classification. First data-set is breast cancer dataset and second one is diabetes data-set. Then they have classified the attributes of disease dataset by applying classification algorithm on WEKA interface. WEKA (classification tool) has three interfaces. Work is done on two interfaces that is WEKA Explorer and WEKA Experimenter interface. On both interfaces they have calculated the classification algorithm accuracy. For evaluating the accuracy of all classification algorithms, The accuracy of the algorithm is determined for every attribute. Performance is determined on the basis FP rate, TP rate, recall, precision etc. Use training set data mode is used for training and testing the data-set.

In [Vikas Chaurasia, Saurabh Pal IJCSE 2014] [3], three data mining techniques are used: RepTree, RBF Network and Simple Logistic. In this paper, they used these algorithms to predict the survivability rate of breast cancer data set. They selected these three classification techniques to find the most suitable one for predicting cancer survivability rate.

III.. PROPOSED SYSTEM

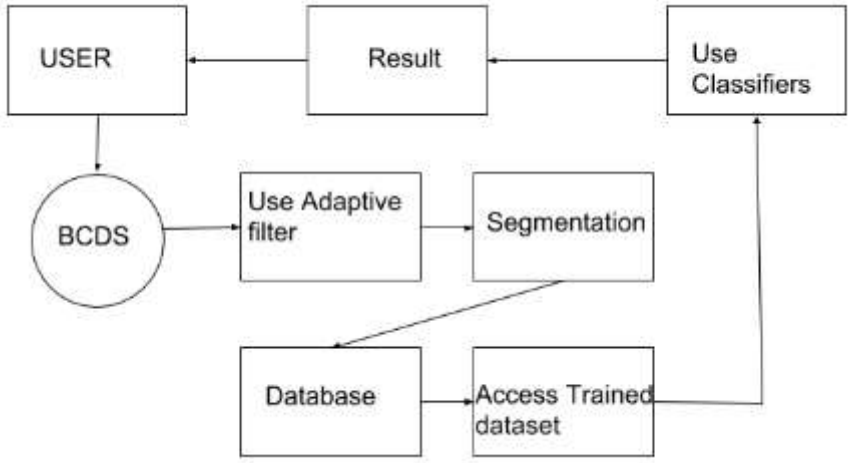

Fig no.1 Block Diagram for proposed system

The proposed system consists of two main layers as seen from above block diagram. In first layer, the first step is taking single image as input from dataset. Adaptive Median Filter is applied to it. After that image segmentation is done using Gaussian Mixture Model (GMM) segmentation. The next step is to check the intensity and extract GLCM features. Once features are extracted then Probability Neural Network (PNN) classifier is trained and applied on the image to predict if cancer is Benign or Malignant. In second layer, input is taken from database (test set) and again all the steps of applying adaptive median filter, image segmentation, feature extraction and lastly using PNN classifier to predict the type of cancer is carried out to predict cancer.

IV. COMPARISON WITH EXISTING SYSTEM

The system thus proposed is different from the current existing systems in many ways which are as follows.The literature surveyed from three paper namely [1] U.Ojha;S.Goel, [2] Deepika Verma;Nidhi Mishra, [3] Vikas Chaurasia;Saurabh Pal.These existing systems use Clustering and classification algorithms to find the performance, some of the algorithms are K means ,EM,PAM and fuzzy c-means these algorithms are used on a training set to get an output of a new entered data point .In the new system we use a different kind of approach where in the user gives X-rays as input to the system and the system processes these images by a sequence of segmentation and adaptive median filter to extract the exact data points and value of the attributes. These newly obtained data point is then run against a previously trained data set. This newly obtained data set is then classified as Benign or Malignant using PNN classifier.This system thus differs from the existing systems as it allows users to give their own input in the form of an X-ray which is then used to collect precise data points.

V. REFERENCES

[1] U.Ojha, S.Goel IEEE 2017 **A study on prediction of breast cancer recurrence using data mining techniques**

[2] Deepika Verma; Nidhi Mishra IEEE 2017 **Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques**

[3] Vikas Chaurasia, Saurabh Pal IJCSE 2014 **Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability**

[4] Mr. Chintan Shah, Dr. Anjali G. Jivani IEEE 2013 **Comparison of Data Mining Classification algorithms for Breast Cancer Prediction**

[5] Shelly Gupta , Dharminder Kumar IJCSE 2011 **Data Mining Classification Techniques Applied for Breast Cancer Diagnosis.**

VI. CONCLUSION

The research helps in determining Breast cancer at an early stage with the help of Data Mining and various Machine Learning algorithms which help in detecting Breast Cancer through various classifiers and helps patients to take necessary actions..The system also provides a different approach than the systems used previously and currently in use.The Project Focuses on using X-ray images as an input and then by using advanced filtering and segmentation to provide necessary results.This system will thus help in reducing the overall deaths caused due to Breast cancer.

.