International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in www.ijarst.in
ISSN: 2457-0362

BREAST CANCER DETECTION USING MAMMOGRAM FEATURES USING RANDOM FOREST ALGORITHM

Sai Prathyusha, P.Pedda Sadhu Naik
Dr.Samuel George Institute of Engineering and Technology, Markapur,A.P

ABSTRACT

Breast Cancer is one of the most dangerous diseases for women. This cancer occurs when some breast cells begin to grow abnormally. Machine learning is the subfield of computer science that studies programs that generalize from past experience. This project looks at classification, where an algorithm tries to predict the label for a sample. The machine learning algorithm takes many of these samples, called the training set, and builds an internal model. This built model is used to classify and predict the data. There are two classes, benign and malignant. Random Forest classifier is used to predict whether the cancer is benign or malignant. Training and testing of the model are done by Wisconsin Diagnosis Breast Cancer dataset.

Keywords: Breast cancer, Random forest algorithm, Wisconsin Diagnosis Breast Cancer dataset.

1. INTRODUCTION

Breast cancer is one of the most dangerous and common reproductive cancers that affect mostly women. Breast tumour is an abnormal growth of tissues in the breast, and it may be felt as a lump or nipple discharge or change of skin texture around the nipple region. Cancers are abnormal cells that divide uncontrollably and are able to invade other tissues. Cancer cells have the ability to spread to other parts of the body through the blood and lymphatic systems. It is the leading cause of death among middle aged and older women. According to cancer statistics, breast cancer is the second most common and the leading cause of cancer deaths among women, second only to lung cancer. Around 1 in 36 (3%) women die due to breast cancer. It has become a major health issue in the past 50 years, and its incidence has increased in recent years in Malaysia, breast cancer is the most frequent type of cancer among women. It has an incidence rate of about 26% (more than 4400 women) among cancer affecting women. Around 40% of the women who suffered from breast cancer in Malaysia have died (IARC). Hence, determining the right decision from a right diagnosis is crucial.

In today's world with the advent of personalized medicine, it increases the workload and complexity of the doctors in cancer diagnosis. Radiologic and pathology are the key players in making decision for cancer diagnosis. Based on the radiology diagnosis, the results will be submitted to pathology for further diagnosis. Pathology and radiology form the core of cancer diagnosis, yet based on our observation at our studied hospital and under current process of diagnostic medicine, the communication among them remained on papers. That paper contains their respective report of the case on the same patient. This scenario is in parallel with what James et al. had highlighted in their paper. The working flows of both specialties remain ad hoc and occur in separate silos with no direct linkage between their case accessioning and/or reporting systems, even when both departments belong to the same host institution. Since both radiologists' and pathologists' data are essential to make correct diagnoses and appropriate patient management and treatment decisions, the isolation of radiology and pathology work flows can be detrimental to the quality and outcomes of patient care. These detrimental effects underscore the need for pathology and radiology work flow integration and for systems that facilitate the synthesis of all data produced by both specialties. With the enormous technological advances currently occurring in both fields, the opportunity has emerged to develop an integrated diagnostic reporting system that supports both specialties and, therefore, improves the overall quality of patient care. In this, we focused on breast cancer diagnostic for data collected from Kaggle. Hence, breast radio-pathological correlation is essential. The covered topics would include radio-pathological correlation with recent imaging advances such as machine learning with use of technical method such as mammography. As a standard, the current diagnostic screening consists of a mammography to identify suspicious regions of the breast, followed by a biopsy of potentially cancerous areas.

Problem Definition

One of the main problems of the treatment of breast cancer is the difficulty of early detection and the lack of data set. Moreover, the breast of the woman differs in both shape and density from that of their counterparts around the world. breast cancer is usually detected in late stages, when it has impairs the function of more vital organ systems and becomes widespread throughout the body.

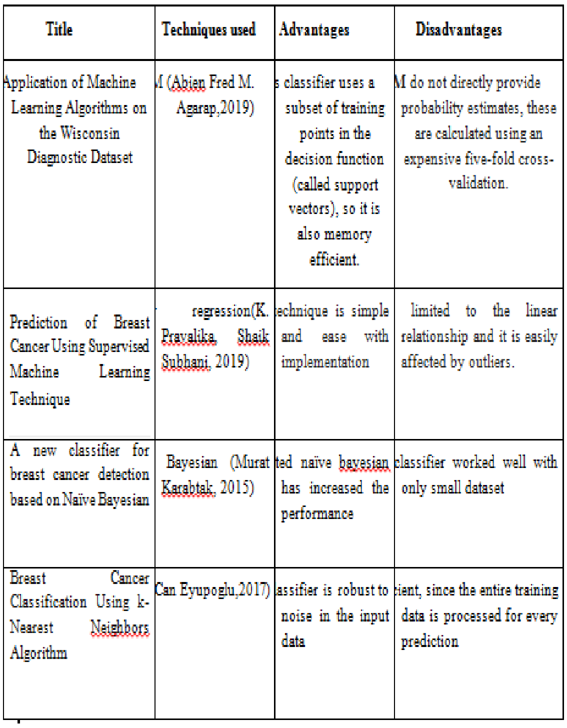
Early detection of diagnostic techniques is the focus of this study because they are of utmost importance. There are many differences between cancer cases, even of the same organ. Different cases of breast cancer, is one of the main reasons that makes treatment so difficult. This is due to the fact that different physical, anatomical or physiological nature. Accurate diagnosis, nowadays, are supposed to be sophisticated to corroborate the outstanding medical techniques for the benefit of the patients. The study has made use of artificial intelligence to process the dataset in general especially testing training.

Objectives of the project

- To detect whether the person's cancer is curable or non-curable.
- To display output (Benign / Malignant)
- To increase the accuracy rate based on the performance of the system

2. LITERATURE SURVEY

Breast cancer is the leading cause of the death among the women. Mammography is the best diagnostic technique for the breast cancer. But not all breast cancer can be seen by mammogram. Although breast cancer can be mortal, people have the highest chances to survive if cancer could be detected at the early stages. But there are certain limitations of the segmentation technique it is difficult to find the effected region perfectly. The proposed work deals with an approach for extracting the malignant masses in the mammography image for the detection of earlier breast cancer. The steps involved are removal of noise from the background information, thresholding and retrieving the largest region of interest, performing morphological operations and extracting the ROI and identifying the malignant masses from the image. Various pre-processing techniques used are Initial cropping, Intensity adjustment, CLAHE (Contrast limited adaptive histogram equalization), Noise reduction, Remove background information, Thresholding, Elimination of noise by locating connected segment larger area, Erosion, Perform subtraction, Removing connected components corresponding to the pectoral muscles.



Title	Techniques used	Advantages	Disadvantages
Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset	I (Abien Fred M. Agarap,2019)	s classifier uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.	M do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.
Prediction of Breast Cancer Using Supervised Machine Learning Technique	regression(K. Pravalika, Shaik Subhani, 2019)	echnique is simple and ease with implementation	limited to the linear relationship and it is easily affected by outliers.
A new classifier for breast cancer detection based on Naïve Bayesian	Bayesian (Murat Karabtak, 2015)	ted naïve bayesian has increased the performance	classifier worked well with only small dataset
Breast Cancer Classification Using k-Nearest Neighbors Algorithm	Can Eyupoglu,2017)	assifier is robust to noise in the input data	ient, since the entire training data is processed for every prediction

Table 1: Comparitive analysis

EXISTING SYSTEM

Drawbacks of existing approach:

- The results indicate only the presence of tumor and it doesn't provide any information whether the cancer present in the patient is benign(curable) or malignant(non-curable).
- Less number of features taken
- Low accuracy

PROPOSED SYSTEM

In proposed breast cancer detection system, we are using Random forest algorithm for testing and training the model. In proposed system, it also results whether the cancer present in the patient is benign(curable) or malignant(non-curable).

Architecture

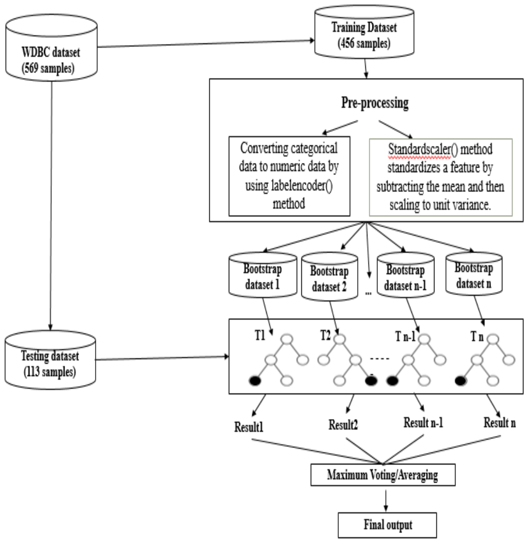


Fig 1. Carcinoma Prediction architecture

The dataset is divided into two parts as training and testing datasets in 80:20 ratio. Now the training dataset undergoes pre-processing. The labelEncoder is used to convert the Y class into 0's and 1's. 0 indicates benign and 1 indicates malignant. To this data, StandardScaler is used to standardise the features to unit variance. Now the random forest classifier splits the dataset into 100 subsets randomly. Each subset is fed to a decision tree. So, each decision tree will individually give separate output. Likewise, we will have 100 outputs arise from 100 decision trees. Now, based on there outputs, the majority label is given as output to the user.

BREAST CANCER DETECTION SYSTEM

a. DATA COLLECTION

The dataset used detecting breast cancer is the Breast Cancer Wisconsin (Diagnostic) Data Set.

This dataset contains 569 records of and 32 features (including the Id and diagnosis). The features of the diagnostic collection describe characteristics of the cell nuclei present in a digitized image of a fine needle aspirate (FNA) of a breast mass [34]. Every cell nucleus is defined by ten traits and for every trait the mean, the standard error and the worst (mean of the three largest values) are computed, resulting in a total of 30 features for each image

Features are nothing but the characteristics of the image such as the ID number, Class, adius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst. The features are fed into the machine learning model. This dataset can be found on UCI Machine Learning Repository.

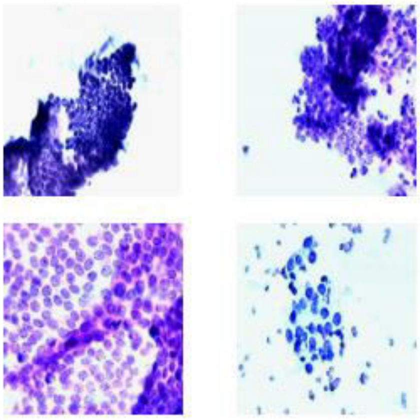


Fig 2. Benign cancer cell

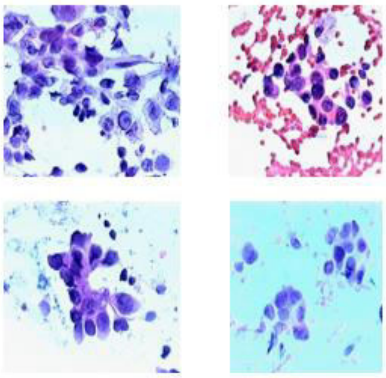


Fig 3. Malignant cancer cell

PROCESS OF BREAST CANCER DETECTION SYSTEM

The Breast cancer detection system is trained using supervised learning approach in which it takes feature values which are taken from mammogram image. The system includes the training and testing phase followed by dataset collection, feature selection, classification. Training and Testing of model is done by using Random Forest algorithm.

Feature selection:

Initially, during the training we have train the system to classify the dataset in different classes based on their label. Then next during the testing period the user will give mammogram feature values as input to the system.

SelectFromModel() is a built in function in SK-learn which is generally used to consider fewer and more prominent features and omit the rest. But we took all the 32 features into account. Also, in decision tree algorithm calculating nodes and forming the rules will happen using the information gain and gini index calculations. Whereas in random forest algorithm, Instead of using information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly.

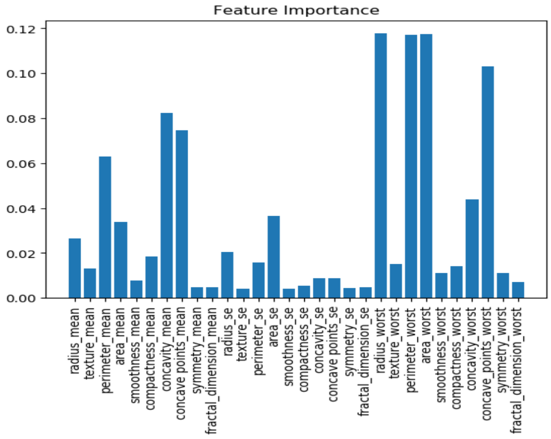


Fig.4 Histogram of feature importance

CONCLUSION

There are many methods such as SVM, KNN and many more for detecting the breast cancer. But our proposed system is developed using random forest algorithm, by using this method we improved the prediction rate of cancer properly which leads to increase of accuracy rate. To built breast cancer detection system we have used Wisconsin Diagnosis Breast Cancer dataset.

Reference

1. Bin Dai; Rung-Ching Chen; Shun-Zhi Zhu; Wei-Wei Zhang, “Using Random Forest Algorithm for Breast Cancer Diagnosis” 2018 International Symposium on Computer, Consumer and Control (IS3C), 6-8 December 2018.

2. T.M. Kolb, J. Lichy, J.H. Newhouse, "Comparison of the performance of screening mammography physical examination and breast US and evaluation of factors that influence them: an analysis of 27825 patient evaluations", Radiology, vol. 225, no. 1, pp. 165-75, 2002.

3. Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., & Bray, F.(2015). Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer. GLOBOCAN

4. Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569).