Breast Cancer Detection

**Abstract:**

Breast cancer is the most common reason for deaths due to cancer. It is very necessary to detect cancer at early stages. There are various Machine Learning techniques available for the purpose of diagnosis of breast cancer data. This paper presents a Machine Learning model to perform automated diagnosis for breast cancer. This method employed CNN as a classifier model and Recursive Feature Elimination (RFE) for feature selection. Also, five algorithms SVM, Random Forest, KNN, Logistic Regression, Naïve Bayes classifier have been compared in the paper. The performance of the system is measured on the basis of accuracy and precision. Activation function such as ReLu have been used to predict the outcomes in terms of probabilities.

**Keywords:**

Breast Cancer, Dataset, CNN, KNN, Naïve Bayes, Random Forest, SVM, Logistic Regression

**Motivation:**

As many people in this world are affected because of the breast cancer ,so it causes physical and mental health problem to the patients, so to make them fully cure, detection is the first process .To make them fully cure and happy it motivates us.

**1.Introduction:**

Breast cancer (BC) is one of the most prominent type of cancer among women all around the world, according to a research conducted by World Health Organization (WHO). BC is a leading causes of death among women all around the world. Breast cancer also has an exceedingly high rate of cancer fatalities in India which is around 14% and is the most common cancer among women. BC. affects about 5% of Indian women, but it affects about 12.5 percent of women in Europe and the United States. It confirms that women in Malaysia who have breast cancer present at a later stage of the disease than women in other countries. Breast cancer is in most of the cases easy to diagnose if any particular symptom appears. Some women with breast cancer, on the other hand, face no symptoms. Thus, periodic breast cancer screening is crucial for early detection.

Doctors say that breast cancer happened due to abnormal growth of cells in the breast and these cells spread in size like Meta Size from breast to lymph nodes or the other parts of the body also. Hence it is necessary to detect and stop the growth of these unwanted cells as early as possible to avoid the next phase consequences. If a tumor is diagnosed then the first step taken by the doctor is, they check whether the tumor is Benign or Malignant. Because the treatment and prevention methods of both the tumors are different. Benign cells are neither cancerous and nor spread but Malignant cells are cancerous and can spread to other parts of bodies. The problem with this disease is, there is no such proper diagnostic machine is present to detect cancer in the early phase so the person can start the treatment as early as possible and try to stop the growth of unwanted cells or tumors.

Early diagnosis of any disease is often curable with a touch amount of human effort. Most people fail to detect their disease before it becomes chronic. It leads to an increase in the death rate around the world. Breast cancer is one of the diseases that could be cured when the disease is identified at earlier stages before it is spreading across all the parts of the body.

The lack of prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, time requires developing the technique which gives minimum error to increase accuracy. The available tests to detect breast cancer such as mammogram, ultrasound, and biopsy were time consuming, so there was a need for a computerized diagnostic system in which Machine Learning methodology was used. This methodology includes algorithms that help for the classification of the tumor and detect the cells more accurately and take less time as well

**2.Literature Review:**

Arpita Joshi and Dr. Ashish Mehta, compared the classification results obtained from the techniques i.e. KNN, SVM, Random Forest, Decision Tree (Recursive Partitioning and Conditional Inference Tree). The dataset used was Wisconsin Breast Cancer dataset obtained from UCI repository. Simulation results showed that KNN was the best classifier followed by SVM, Random Forest and Decision Tree.

Kalyani Wadkar, Prashant Pathak and Nikhil Wagh, did a comparative study on ANN and SVM and integrated various classifiers like CNN, KNN and Inception V3 for better processing of the dataset. The experimental results and performance analysis concluded that ANN was a better classifier than SVM as ANN proved to have a higher efficiency rate.

K. Anastraj, Dr. T. Chakravarthy, K. Sriram , have performed a comparative analysis between different machine learning algorithms: back propagation network, artificial neural network (ANN), convolutional neural network (CNN) and support vector machine (SVM) on the Wisconsin Breast Cancer (original) datasets. Deep and convolutional neural network with ALEXNET was used for feature extraction are analysis the benign and malignant tumor. The simulation results concluded that support vector machine is the best approach and had given better results (94%).

**3. Methodology:**

**Logistic Regression**

In linear regression, the linear regression hyperplane that is obtained cannot be used to predict the dependent variable by using the independent variable. Hence, when there is categorical data, logistic regression is used. Logistic Regression predicts whether something is true or false instead of predicting something continuous. It is used for classification. The sigmoid function is used to convert the independent variable into an expression of probability which ranges from 0 and 1 concerning the dependent variable. The ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular Machine Learning algorithm. A drawback of Logistic Regression is the assumption of linearity between the dependent and independent variables.

**Decision Tree**

Decision Trees (DTs) are one of the most helpful administered learning calculations out there. Rather than solo realizing (where there is no result variable to direct the growing experience and information is investigated by calculations to track down designs), in administered learning your current information is as of now marked and you know which conduct you need to foresee in the new information you get. This is the sort of calculations that independent vehicles use to perceive walkers and items, or associations exploit to gauge client’s lifetime esteem and their beat rates. Decision Trees are Machine Learning calculations that continuously partition informational collections into more modest information bunches considering an unmistakable element, until they arrive at sets that are sufficiently little to be portrayed by some mark. They expect that you have information that is marked (labelled with at least one names, like the plant name in pictures of plants), so they attempt to name new information considering that information.

**Random Forest Classifier**

Random forest is an administered learning algorithm. It is an assortment of Decision Trees. Decision Tree is various levelled in nature in which nodes address specific circumstances on a specific arrangement of highlights, and branches split the decision towards the leaf nodes. Leaf decides the class marks. Decision Tree can be developed either by utilizing Recursive Partitioning or by Conditional Inference Tree. Recursive Partitioning is the bit-by-bit process by which a Decision Tree built by either parting or not parting every node. We can say that the tree is advanced by parting the source set into subsets in view of a property estimation test. The recursion ends if subset at a node has a gross similar value to the objective variable. Contingent Inference Tree is a factual based approach that involves non parametric tests as dividing models that is rectified for different testing to avoid over fitting. Random Forest is reasonable for high layered information displaying as it can deal with missing qualities, persistent, all out and parallel information however for very informational indexes, the size of the trees can take up a ton of memory. It can will quite often over-fit, so there is a need to tune the hyper-boundaries

**Our Observations:**

The decision tree has 93.85% accuracy.

The logistic regression has the 96.49% accuracy.

The Random forest classifier has the 97.36% accuracy.

**Conclusion:**

The decision tree has the highest accuracy than logistic regression and the random forest classifier. So decision tree will be the great technique for the classification of the cancer tumor (Benign or Malignant).

**References:**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/#:~:text=Afshar%20et%20al.,%25)%20and%20accuracy%20(84%25>)

<https://www.ijert.org/research/breast-cancer-detection-using-machine-learning-techniques-IJERTV10IS070064.pdf>

Mullooly M, Ehteshami Bejnordi B, Pfeiffer RM, Fan S, Palakal M, Hada M, et al. Application of convolutional neural networks to breast biopsies to delineate tissue correlates of mammographic breast density. Npj Breast Cancer 2019;5:43. <https://doi.org/10.1038/s41523-019-0134-6>.

Rauschert S, Raubenheimer K, Melton PE, Huang RC. Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. Clin Epigenetics 2020;12:51. https://doi.org/10.1186/s13148-020-00842-4.