**SYNOPSIS**

**Report on**

**Performance evaluation of Breast Cancer diagnosis using various Machine learning algorithms**

**by**

**BHARTI GHILDIYAL -**  **2100290140051**

**DIKSHA GUPTA -** **2100290140060**

**KM PURNIMA -** **2100290140075**

**Session:2022-2023 (4th Semester)**

Under the supervision of

**Dr Sangeeta Arora**

### KIET Group of Institutions, Delhi-NCR, Ghaziabad



### Department Of Computer Applications

**KIET GROUP OF INSTITUTIONS, DELHI-NCR, GHAZIABAD-201206**

(march- 2023)

**ABSTRACT**

Breast cancer affects the majority of women worldwide, and it is the second most common cause of death among women. However, if cancer is detected early and treated properly, it is possible to be cured of the condition. Early detection of breast cancer can dramatically improve the prognosis and chances of survival by allowing patients to receive timely clinical therapy. Furthermore, precise benign tumour classification can help patients avoid unneeded treatment.

Machine learning techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer, became a research hotspot and has been proved as a strong technique. In this study,four machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, Naïve Bayes and K-Nearest Neighbours (KNN) are applied for numerical dataset, after obtaining the results, a performance evaluation and comparison is carried out between these different classifiers. This paper study uses Convolution Neural Networks for Image dataset.

These techniques are coded in python and uses numpy, pandas, matplotlib, seaborn libraries. Finally, both image and numerical test data will be used for prediction.

**Keywords:** Machine Learning, Breast Cancer Detection, Logistic Regression

**TABLE OF CONTENTS**

1. Introduction 2
2. Literature Review 3
3. Project / Research Objective 5
4. Research Methodology 6
5. Project / Research Outcome 8
6. Proposed Time Duration 9

7. References 10

**1. Introduction**

Breast cancer has become the most recurrent type of health issue among women especially for women in middle age. Early detection of breast cancer can help women cure this disease and death rate can be reduced. In the present-day scenario, to observe breast cancer mammograms are used and they are known be the most effective scanning technique. The leading risk factor for Breast cancer is simply being a women. Men can also get breast cancer In 2017, the American Cancer Society estimates 2,470 new case of invasive breast cancer will be diagnosed in men in US.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

We will use the Machine Learning Repository for breast cancer dataset. The dataset is taken from Kaggle.

The data set used in this project is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan[4]. The proposed technique indicates the presence of cancer based on regression while new algorithms are available. Model which has been designed for prediction of new data and should give good result in their training and testing phase. Here there are 3 main steps pre processing features, extraction and classification.

Here, we also propose the fully automatic classification and prediction of breast cancer based on image dataset. Using deep learning technique. This learning technique is recognised as best method to predict and classify for image dataset.

**2. Literature Review**

The work in this research is focusing on various models for predicting the time of breast cancer tumor recurrence. Methods include screening the database from GLOBOCAN, CDC, and WHO health repository highlights the lethality of breast cancer, taking thousands of lives each year. However, a timely prediction of cancer can help patients to consult the doctor on time. In the past, various studies have successfully predicted the nature of the tumor to be benign or malignant and if the breast cancer tumor will reoccur or not but, no time-based models have been studied. With the help of Machine Learning, this study shows various prediction models that can be used to predict tumor reoccurrence time as accurately as 1 year. Among the 198 patients analyzed, 40% of the total patients were predicted to have breast cancer tumors reoccurring within 1st year of the diagnosis. The proposed machine learning techniques use various classification models such as Spectral clustering, DBSCAN, and k-means along with prediction models like Support Vector Machines (SVM), Decision trees, and Random Forest. The results demonstrate the ability of the model to predict the time taken by the tumor to reoccur or the time taken by the patient for full recovery with the best accuracy of 78.7% using SVM. This population-based study performed on multivariate real attributed characteristics data can therefore provide the patients a reasonable estimate about their recovery time or the time before which they should consult the doctor.[1]

Priyanka Khanna, Mridu Sahu, Bikesh Kumar Singh, Vikrant Bhatia proposed a model integrating pre-trained Convolutional neural network (CNN) with machine learning for prognosticating pathologic complete response (PCR) using breast cancer dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) prior to commencement of neoadjuvant chemotherapy (NACT). For predicting pathologic complete response (PCR) using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) for breast cancer prior to the start of neoadjuvant chemotherapy, they presented a hybrid model integrating a pre-trained Convolutional neural network (CNN) with machine learning (NACT). In this retrospective study, 64 patients receiving NACT for invasive breast cancer are examined. Deep learning-based pre-trained CNN models ResNet-50 and ResNet-18 were used to extract features from patient visit 1 MRI images (before the initiation of NACT). Mann-Whitney U tests is used to assess features and their relevance (significance level p<0.05 and confidence interval is 95%). Furthermore, features extracted and features selected were independently given as an input to different machine learning classifiers for the prediction of response of NACT. Classification performance was assessed under different data division protocols using accuracy, specificity, sensitivity, and area under the receiver operating characteristic curve (AUROC). The proposed model employing DCE-MRI images acquired before starting chemotherapy has considerable accuracy in classifying PCR and non-PCR patients. The efficacy of the prediction model can improve considerably on the back of a larger dataset.[2]

Breast cancer is the second prevalent type of cancer among women. Breast Ultrasound (BUS) imaging is one of the most frequently used diagnostic tools to detect and classify abnormalities in the breast. To improve the diagnostic accuracy, Computer Aided Diagnosis (CAD) system is helpful for breast cancer detection and classification. Normally, a CAD system consists of four stages: pre-processing, segmentation, feature extraction, and classification. In this paper, the pre-processing step includes speckle noise removal using Speckle Reducing Anisotropic Diffusion (SRAD). The goal of segmentation is to locate the Region of Interest (ROI) and Active contour-based segmentation is used in this work. The texture features are extracted and fed to a classifier to categorize the images as Normal, Benign and Malignant. In this work three classifiers namely K Nearest Neighbours (KNN) algorithm, Decision tree algorithm and Random Forest classifier are used and the performance is compared based on the accuracy of classification.[3]

**3. Research Objective**

Breast cancer is a disease which we hear about a lot nowadays. It is one of the most widespread diseases. There are around 2000+ new cases of breast cancer in men each year, and about 2,30,000 new cases in women every year. Diagnosis of this disease is crucial so that woman can get it treated faster. It is best for a correct and early diagnosis.

The main objective of this research is to help doctors analyze the huge data sets of cancer data and find patterns with the patient’s data and that cancer data available. With this analysis we can predict whether the patient might have breast cancer or not.

Machine learning algorithms will help with this analysis of the data sets. These techniques will be used to predict the outcome. The outcome is based on the performance of ML algo applied on the data set which algo is most accurate in prediction of breast cancer.

This prediction can help doctors prescribe different medical examinations for the patients based on the cancer type. This helps save a lot of time as well as money for the patient.

**4. Research Methodology**

1.Data collection: Two data sets are used one is numerical data set for evaluating the performance of various machine learning algorithms and other one is image data set for building a deep learning model.

(a) Numerical data set: WBCD repository provided the numerical dataset. Features are composed of fine needle aspirate (FNA) of a breast mass. There are 30 features that was extracted to describe characteristics of cell nuclei present in the scanned images. The data set consists of 569 patients ,212 have an outcome of malignancy and 357 are Benign. The classes in the data set are separated into 2 or 4 groups, with 2 corresponding to the benign case and 4 corresponding to the malignant case.

(b) Image data set: The original data set consisted of 162 whole mount slide images of Breast Cancer specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive). Each patch’s file name is of the format: u\_xX\_yY\_classC.png — > example 10253\_idx5\_x1351\_y1101\_class0.png . Where u is the patient ID (10253\_idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC

2. Data cleaning: The term "data pre-processing" refers to the process of converting unstructured data to structured data, as well as resizing and removing undesirable data from a data set. The data set's missing traits are replaced by the mean value. The data is then randomly selected from the data set to ensure that the data is circulated properly.

3.Training and Testing: Training phase extracts the features from the data set,

and the testing phase will deliver new data to be examined to see how well our algorithm works and behaves when it comes to prediction.

4.Proposed convolution neural network for image data set analysis: We are utilizing the CNN algorithm for analysis and prediction in this proposed model.

5.Proposed Method for Numerical Data set: In Numerical data set we are using 4 different algorithms and identifying which algorithm is suitable for the data set and gives most accurate output. Algorithms used are SVM, KNN, Decision Tree and Naïve bayes.

**5. Research Outcome**

The final outcome of the research targeted on correctness of the algorithms in the training data set. This work chose Support Vector Machine (SVM), Logistic Regression, KNN, Naïve Bayes algorithm for test; the performance analysis of all the algo comes as outcome and after comparing the performance and accuracy of the algorithms we will have which one is the most accurate and best to detect breast cancer.

To facilitate the ease of interface the GUI is developed using Flask framework to connect the front end to the back-end model to process and provide prediction. Medical practitioners can enter input values manually using patient records and on submission the record is classified as malignant or benign. Also, image can be uploaded which then will be process by the model built and the prediction is made.

**6. Proposed Time Duration**

| **Task name** | **week 1** | **week 2** | **week 3** | **week 4** | **week 5** | **week 6** | **week 7** | **week 8** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Planning** |  |  |  |  |  |  |  |  |
| **Research** |  |  |  |  |  |  |  |  |
| **Design** |  |  |  |  |  |  |  |  |
| **Implementation** |  |  |  |  |  |  |  |  |
| **Follow up** |  |  |  |  |  |  |  |  |

7. **References**

[1] Gupta(2022) ‘Prediction time of breast cancer prediction’*. Cancer Treatment and Research Communication.*

https://[www.sciencedirect.com/science/article/pii/S2468294222000922](http://www.sciencedirect.com/science/article/pii/S2468294222000922)

[2] Priyanka Khanna, Mridu Sahu, Bikesh Kumar Singh, Vikrant Bhateja, **’**Early Prediction of Pathological Complete Response to neoadjuvant Chemotherapy in breast cancer MRI images using combined Pre-trained Convolutional neural network and machine learning**’.***Measurement***(**2022**)**

https://[www.sciencedirect.com/science/article/pii/S0263224122014658](http://www.sciencedirect.com/science/article/pii/S0263224122014658)

[3] S. Pavithra, R. Vanithamani, J. Justin,‘Computer aided breast cancer detection using ultrasoundimages’,*MaterialsToday:Proceedings*,(2020),

https://[www.sciencedirect.com/science/article/pii/S2214785320362532](http://www.sciencedirect.com/science/article/pii/S2214785320362532)

[4] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data