Breast Cancer Prediction Using Machine Learning

Dr. R. Arun
P.S. R Engineering College

ABSTRACT

Breast cancer is an illness that we frequently hear about a lotnowadays. It is among the most prevalent illnesses. It impacts women all across the world. As per National Cancer Institute, breast cancer is the 2nd most frequent type of cancer among United States women. Every year, there are approximately 2000+ new cases of breast cancer in men, compared to 2,30,000 new instances in women. It is essentialto diagnose this illness so that women may begin treatment right away. It is ideal for an accurate and prompt diagnosis. This is a crucial stage in recovery and therapy. Mammograms, which are essentially breast X-rays are used to identify breast cancer. It's a device that's used to find and assist in the detection of breast cancer. However, detection is not always simple because of many types of uncertainties in utilizing these mammograms. ML (Machine Learning) method could assist in the detection of breast cancer. These methods may be utilized to provide tools for physicians thatact as an effective method for the early diagnosis and identification of breast cancer, which might considerably increase the survival rates of the patients.

Keywords: Mammograms, Machine Learning, Cancer

INTRODUCTION

Cancer is an illness that arises when mutations or changes take place in the genes that help cell development. These mutations allow the cells to multiply and divide in a very uncontrolled and chaotic manner. These cells continue to multiply and begin to make copies, which leads to their abnormality progressing. A tumour is eventually formed by these abnormal cells. Tumors, unlike other cells, don’t die even though the body doesn’t need them. The term "breast cancer" refers to cancer that starts in the breast cells. This type of cancer could be observed in the breast lobules/ducts. Additionally, fatty tissue and connective fibrous tissue in the breast are also susceptible to cancer. These cancer cells have an uncontrollable nature that allows them to infect other healthy breast tissues and the lymph nodes underneath the arms. There are 2 forms of cancers. Benign and Malignant.

Malignant tumours are cancerous. These cells keep dividinguncontrollably and start affecting other tissues and cells in the body. They distribute to all other body parts and it is hardto cure this type of cancer. Treatment options for various tumor types include chemotherapy, radiation therapy, and immunotherapy. Benign cancer is not cancer. This tumor does not spread to other body parts and is thus considerably less dangerous than a malignant one. In

many cases, such tumors don't require any treatment. Breast cancer is often diagnosed in women over the age of 40. But this illness can affect men and women of any age. It can also occur in families with a history of breast cancer. Statistics show that breast cancer has a high death rate and that it alone is accountable for 15% of all cancer deaths in women globallyand about 25% of all new cancer diagnoses. Scientists know about the dangers of it from very early on, and hence there's been a lot of research put into finding the right treatment for it. Breast cancer can be detected with the use of mammograms, which are essentially breast X-rays. It is a device that aids in the diagnosis & detection of breast cancer. However, identification is difficult owing to several types of uncertainties in utilizing such "mammograms". The result of a mammogram is images that can show any calcifications or deposits of calcium in the breasts. These don't always have to be cancerous. This test could also detect cysts, which are fluid-filled sacs that are common throughout the menstrual cycles of certain women, as well as cancerous and noncancerous lumps.

PROPSED WORK

In the proposed system we plan on using existing data of breast cancer patients which have been collected for a number of years and run different machine learning algorithms on them. These methods will examine the data from datasets to estimate whether or not a patient has breastcancer and it will also tell us if the disease is benign or malignant.
It is done by taking the patient's data and mapping it with the dataset and checking whether there are any patterns found with the data. If a patient has breast cancer, then instead of taking more tests to check whether the cancer is malignant orbenign, ML could be applied to estimate the case on the basisof the huge amount of data on breast cancer. This proposed system helps the patients as it reduces the amount of moneythey need to spend just for the diagnosis.
Also, if the tumor is benign, then it is not cancerous, and the patient doesn't need to go through any of the other tests. Thissaves a lot of time as well.

MODULE DESCRIPTION

1. **Dataset**

The WDBC ("Wisconsin Diagnostic Breast Cancer") dataset which can be found in the University of California, Irvin's ML Dataset Repository will be used for this project. All ML algorithms will be performed on this dataset. The characteristics that make up the dataset have been computed from a digitized picture of a breast mass that was obtained via a FNA ("Fine Needle Aspirate"). These properties of thecell nuclei shown in the image are explained using these features. The dataset comprises 569 data points: 212 for Malignant and 357 for Benign.

2. Artificial Neural Network

Artificial neural networks are a very important tool used in machine learning. This technique, as the name suggests, is inspired by the brain and its activities. They were designed in a way to replicate the way people learn. Neural networks (NNs) normally comprise input as well as output layers, and sometimes a hidden layer containing units that modify the I/P to something that the O/P layer could use. These devices helpin finding different patterns which are too hard for a human to find himself.

MACHINE LEARNING ALGORITHM

1. Decision Tree

This algorithm is applied to estimate the value of a target or output variable based on many input variables. They are a collection of divides and conquer-problem-solving strategies. It takes the shape of a tree-like structure. It starts with root nodes and this splits into sub/child nodes. These branches keep spitting until the outcome isn't reached. It is mostly utilized for classification issues.

Each node inside a Decision tree (DT) indicates a "test" on an attribute (like whether a coin was turned tails or heads), eachbranch indicates the test outcome, and every leaf node exhibitsa class label (choice made after computing every attribute). The classification rules are presented by the paths from the root to the leaf.

DTs have several benefits, including:

- easy to comprehend and interpret. Trees could be visualized.
- Requires minimum data preparation. Other methods generally need data standardization, the formation of dummy variables, and the elimination of blank values. Notably, "missing values" are not accepted by this module.
- The cost of applying the tree (such as data estimation) isproportional to the amount of training data points required.
- Able to manage data that is both numerical and categorical. The analysis of datasets with just one type of variable is often the domain of other methodologies. Capable of handling issues with various outputs.
- Using the "white box" model. Boolean logic makes it easy to describe a condition if it could be seen in a system for a particular circumstance. A black box model like ANN may make it harder to comprehend the results.
- A model can be validated using statistical testing. This makes it easy to take the model's reliability into consideration.

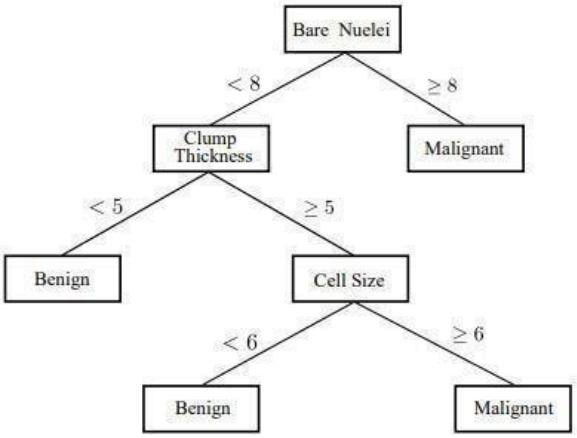

Fig 1 Decision Trees

2. K-Nearest Neighbour Algorithm

This method is among the simplest ML techniques. It is a lazy learning approach used for regression & classification. The objects are classified on the basis of their k-NNs. k-NN only examines the surrounding neighbors, not the underlying data distribution. If k=1, the unknown is essentially allotted to the class of the nearest neighbor. If k is more than 1, the classification is measured by a majority vote using the k-NN prediction outcome. Giving neighbors' contributions weight in such a manner that the neighbors who are close to the average contribute more than the neighbors who are farther away is an effective strategy for both classifications and regression. For example, a popular weighting system is assigning every neighbor a weight of 1/*d*, here *d* indicates thedistance to the neighbors.

A KNN-based classification algorithm's outcome may be identified as the class with the highest frequency among the K most similar occurrences. The prediction belongs to the class that obtains the most votes, with each instance acting as a vote for its class. It is a good notion to select a K value with an oddnumber if you are employing K and you have an even numberof classes (for example, 2) to prevent a tie. Additionally, if you have an odd number of classes, use an even number for K.

Expanding K by 1 and examining the class of the next case inthe training dataset can reliably break ties.

3. Naïve Bayes

It is a classic method for solving classification issues since itis founded on Bayes' theorem. The equation is:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Fig. 2 Navie Bayes Formula

The goal of this exercise is to calculate the probability that event A will occur provided that event B occurs. The naive Bayes classifier incorporates decision-making criteria with the Bayes model, such as the hypothesis that represents the most likely outcomes. The category of supervised learning methods termed naive Bayes techniques is predicated on the naive concept that every pair of features is conditionally independent provided the value of a class variable. It was firstused as a benchmark for text classification tasks and is still inuse today.

4. Forest and Tree Method

Random forests (RF) also recognized as “random decision” forests, are a popular ensemble approach for creating prediction models for classification as well as regression issues. Ensemble techniques integrate many learning models to get better-projected results; for instance, the RF model generates a whole forest of illogical, uncorrelated DTs to identify the optimum solution. RF attempts to minimize the correlation problem by only choosing a subset of the feature space at every split. It aims to set a stopping condition for node splits to de-correlate the trees and prune the trees.

- Classification & regression tasks may be performed with the same RFC (“Random Forest Classifier”) or RF technique.
- The RFC will handle the values that are missing.
- The RFC will not overfit the model if there are more trees in the forest

SYSTEM ARCHITECTURE

As this project does not have any UI, the architecture is the dataset and the features of the dataset. It is trying to understand the dataset and try making the system as simple and easy as possible. The dataset is first split into training as well as testing sets. The training set is first exposed to the ML methods so that the system understands what data give what type of outcome.

After the system is trained, the testing data is used to test whether the system can correctly predict the class of the data.It checks the model's percentage accuracy.

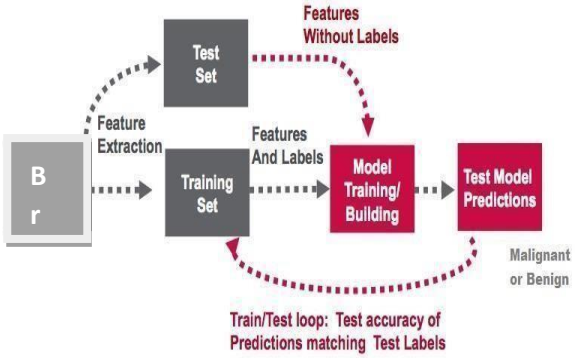

Fig 3 Architecture of System

IMPLEMENTATION AND OUTPUT

1. **Preparation of Data:**

Step 1: The first step in the ML process is to prepare the data.This includes importing all the packages that would assist us to organize as well as visualize the data. The packages used are as follows

```
import pandas as pdimport numpy as npimport matplotlib.pyplot
```

Step 2: After importing all the necessary packages, we need to load the dataset. We use the help of Pandas to load the dataset.

```
data = pd.read_csv('../input/data.csv');"
```

Step 3: We need to drop the dataset's first column which
consists of IDs as this field will not help us in theclassification process. This is done as follows:

```
data.drop(data.columns[[-1, 0]], axis=1,inplace=True
```

Step 4: Count the number of malignant and benign datapoints.

```
"diagnosis_all = list(data.shape)[0]
diagnosis_categories = list(data['diagnosis'].value_counts())
print("\n \t The data has {} diagnosis, {} malignant and {}
benign.".format(diagnosis_all, diagnosis_categories[0],diagnosis_categories[1]))"
```

2. Visualizing The Data

We need to develop data visualizations to decide how to proceed with the ML tools. The Seaborn and the Matplotlib packages will be used for this purpose. We use the features' average values. To simplify some tasks and make the code more readable, we must first isolate those features from the list. “features_mean= list(data.columns[1:11])”

The first method that can be used for visualization is a heat map. A heat map is a two-dimensional depiction of data in which color corresponds to a numerical value. A straightforward heat map gives a fast visual overview of data. Complex data sets may be comprehended through the use of heat maps with increased detail. By charting the distribution of each kind of diagnosis for each of the mean characteristics, we can also observe how malignant and benign tumor cells may or may not have distinct feature values.

We can make use of box plots for visualization. In descriptive statistics, a boxplot or box plot is a visual representation of groups of numerical data by their quartiles. Box-and-whisker plots as well as box-and-whisker diagrams refer to box plots that may also include vertical lines (whiskers) indicating variation outside of the top and bottom quartiles. Individual points can be used to plot outliers. Box plots are non- parametric because they show variation in statistical samples without assuming anything about the distribution they come from. Outliers and the level of dispersion & skewness in the data are shown by the distances between the different box components.

RESULT

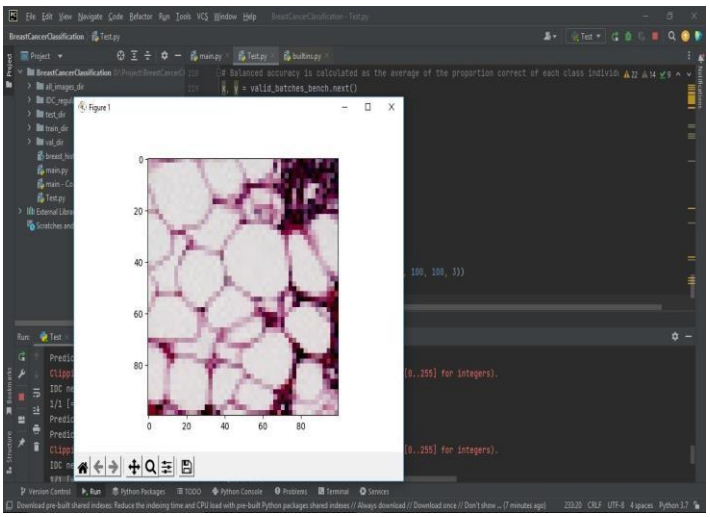

Fig 4 Result

CONCLUSION

In this paper, we have worked to collect the suitable dataset needed to help in this predictive analysis. This dataset is then processed to remove all the junk data. The predictive analysis method is being used in many different fields and is slowly picking up pace. It is helping us by using smarter ways to solve or predict a problem's outcome. Our scheme was developed to reduce the time and cost factors of the patients as well as to minimize the work of a doctor. We have tried to use a very simple and understandable model to do this job. Next, machine learning algorithms should be used in the training as well as the testing data should be used to check if the outcomes are accurate enough.
In the future, we can also use a dataset to predict the re- occurrence of breast cancer after surgery or chemotherapy session. ANNs can be applied to make the prediction better and smarter. Accuracy can be increased by selecting better features.

REFERENCE

[Preeja, 20] Ammu P K and Preeja V. Article: Review on Feature Selection Techniques of DNA Microarray Data. International Journal of Computer Applications 61(12):39-44, January 20.

[Chang, 11] Bing Nan Li, Chee Kong Chui, Stephen Chang,
S.H. Ong, Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, Computers in Biology and Medicine, Volume 41, Issue 1, 2011,

[Konar, 14] Chattopadhyay, P., Konar, P. Feature Extraction using Wavelet Transform for Multi-class Fault Detection of Induction Motor. J. Inst. Eng. India Ser. B 95, 73–81 (2014).

[Li, 11] C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan, and C. Li, "Advances in Research of Fuzzy C-Means Clustering Algorithm," 2011 International Conference on Network Computing and Information Security,
Guilin, 2011.

[Jemal, 18] DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods." 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (2018)

[Cadima, 16] Jolliffe IT, Cadima J. Principal Component Analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;

[Maria, 20] Kalist, V. & Packyanathan, Ganesan & Joseph, Maria & B.S, Sathish & Murugesan, R.. (2020). Image Quality Analysis Based on Texture Feature Extraction Using Second-Order Statistical Approach.

[Varalatchouny, 17] M. Ravishankar and M. Varalatchoumy, "Four novel approaches for detection of the region of interest in mammograms — A comparative study," 2017 InternationalConference on Intelligent Sustainable Systems (ICISS), Palladam, 2017,

[Menalsh, 17] Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of LivestockResearch.

[Soni, 20] Reddy, V. Anji, and Badal Soni. "Breast Cancer Identification and Diagnosis Techniques." Machine Learningfor Intelligent Decision Science. Springer, Singapore, 2020.

Pushparaj, 13] S. Palaniappan and T. Pushparaj, “A Novel Prediction on Breast Cancer from the Basis of Association rules and Neural Network”, 2013.

[Agarwal, 17] S. Singh, D. Srivastava, and S. Agarwal, "GLCM and its application in pattern recognition," 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), Dubai, 2017.