

# Breast Cancer Prediction using Supervised Machine Learning Algorithms

Mamta Jadhav<sup>1</sup>, Zeel Thakkar<sup>2</sup>, Prof. Pramila M. Chawan<sup>3</sup>

<sup>1</sup>B.Tech Student, Dept of Computer Engineering, VJTI College, Mumbai, Maharashtra, India

<sup>2</sup>B.Tech Student, Dept of Computer Engineering, VJTI College, Mumbai, Maharashtra, India

<sup>3</sup>Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

\*\*\*

**Abstract** - Breast Cancer is the most leading malignancy affecting 2.1 million women each year which leads to greatest number of deaths among women. Early treatment not only helps to cure cancer but also helps in prevention of its recurrence. And hence this system mainly focuses on prediction of breast cancer where it uses different machine learning algorithms for creating models like decision tree, logistic regression, random forest which are applied on pre-processed data which suspects greater accuracy for prediction. Amongst all the models, Random Forest Classification leads to best accuracy with 98.6%. These techniques are coded in python and uses numpy, pandas, seaborn libraries.

**Index Terms:** Decision Tree, Logistic Regression, Random Forest Classification, Numpy, Pandas, Seaborn.

## 1. INTRODUCTION

According to World health organization, Breast cancer is the most frequent cancer among women and it is the second dangerous cancer after lung cancer. In 2018, from the research it is estimated that total 627,000 women lost their life due to breast cancer that is 15% of all cancer deaths among women. In case of any symptom, people visit to oncologist. Doctors can easily identify breast cancer by using Breast ultrasound, Diagnostic mammogram, Magnetic resonance imaging (MRI), Biopsy. Based on these test results, doctor may recommend further tests or therapy. Early detection is very crucial in breast cancer. If chances of cancer are predicted at early stage then survivability chances of patient may increase. An alternate way to identify breast cancer is using machine learning algorithms for prediction of abnormal tumor. Thus, the research is carried out for the proper diagnosis and categorization of patients into malignant and benign groups.

Three types of tumors are as follows:-

a. Benign tumors are not cancerous they cannot spread or they can grow very slowly. And if doctors remove them, then they cannot cause any harm to the human body.

b. In Premalignant tumors the cells are not cancerous but they have potential to become malignant.

c. Malignant cells are cancerous and they can spread rapidly in body.

In machine learning, cancer classification can be done using benign or malignant cells could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed. And thus, our aim is to develop a prediction system that can predict chances of breast cancer on a huge data.

## 2. LITERATURE SURVEY

Data mining is been applied on medical data of the past and current research papers. Thorough study is done on various base reports. Jacob et al. [1] have compared various classifier algorithms on Wisconsin Breast Cancer diagnosis dataset. They came across that Random Tree and SVM classification algorithm produce best result i.e. 100% accuracy. However they mainly worked on 'Time' feature along with other parameters to predict the outcome of non-recurrence or recurrence of breast cancer among patients. In this paper, "Time" feature has not been relied upon for prediction of recurrence of the disease. Here, prediction is based on "Diagnosis" feature of WBCD dataset.

Chih-Lin Chi et al. [2] used the ANN model for Breast Cancer Prognosis on two dataset. They predicted recurrence and non-recurrence based on probability of breast cancer and grouped patients with bad (<5 years) and good (>5 years) prognoses. Delen et al. [3] used the SEER dataset of breast cancer to predict the survivability of a patient using 10-fold cross validation method. The result indicated that the decision tree is the best predictor with 93.6% accuracy on the dataset as compared to ANN and logistic regression model.

## 3. PROPOSED SYSTEM

### 3.1 Problem Statement

"To identify breast cancer symptoms at early stage to save someone's life by using data mining techniques and machine learning models on WBCD dataset."

### 3.2 Proposed Methodology

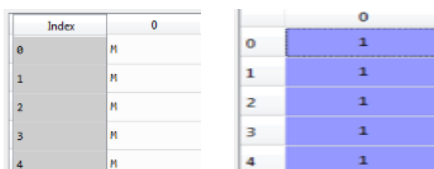
We obtained the breast cancer dataset of Wisconsin Breast Cancer diagnosis dataset and used jupyter notebook as the platform for the purpose of coding. Our methodology involves use of supervised learning algorithms and classification technique like Decision Tree, Random Forest and Logistic Regression, with Dimensionality Reduction technique.

#### 1. Data Processing:

Our dataset may be Incomplete or have some missing attribute values, or having only aggregate data. So, there is a need to pre-process our medical dataset which has major attribute as id, diagnosis and other real valued features which are computed for each cell nucleus like radius, texture, parameter, smoothness, area, etc.

#### 2. Categorical Data

Categorical data are variables that contain label values rather than numeric values. So, here we have represented benign cells as value 0 and malignant cells as value 1.



Index	diagnosis	0
0	M	
1	M	
2	M	
3	M	
4	M	

0	1
0	1
1	1
2	1
3	1
4	1

Fig -1: Data Conversion

#### 2.1 Splitting the dataset:

The data we use is usually split into training data and test data. In our project 75% data is trained data and 25% data is test data.

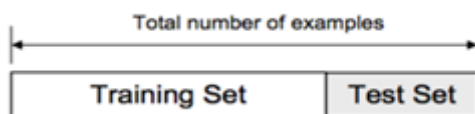


Fig -2: Data Splitting

#### 3. Feature Scaling

Generally, dataset contains features which highly varies in magnitudes, units and range. So there is a need to bring all features to the same level of magnitudes. This can be achieved by scaling.

#### 4. Model Selection

This is the most important phase where algorithm selection is done for the developing system. Data Scientists use various types of Machine Learning

algorithms which can be classified as: supervised learning and unsupervised learning.

For this Prediction System, we only need Supervised Learning.

#### 4.1 Supervised Learning:

Supervised Learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised systems provide the learning algorithms with known quantities to support future judgments.

#### 1. Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction in machine learning. The Decision trees algorithm consists of two parts: nodes and rules (tests). A Decision tree is like tree structure, where node denotes a test on an attribute, Branch represents an outcome of the test, and each leaf node holds a class label.

For detecting breast cancer, its leaf nodes are categorised as benign and malignant. And then certain rules are established to check tumor is benign or malignant.

#### 2. Random Forest Classification

Random forest algorithm is a supervised classification algorithm. In this classifier, the higher the number of trees in the forest gives the high accuracy results.

a. Random forest algorithm can use for both classification and the regression task and can handle the missing values too. For this dataset, we have already handled missing values of attributes. Random forest classifier doesn't over fits the model, if it includes many trees.



Fig -3: Random Forest Classifier

b. It can also work on categorical values too. In this case, we had categorical data as B & M representing benign & malignant which is further converted to numeric data as 0 & 1 respectively.

### 3. Logistic Regression:

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. It is generally used when the dependent variable is binary and provides a constant output.



Fig -4: Graph for logistic regression function.

Estimated efficiency of Logistic model will be the 95% which is obviously lower than Random Forest.

### 3.3 Proposed System Architecture

As shown in below diagram, we first collected the data. To create a machine learning models, collecting appropriate data is very essential. After collection of data, Cleaning needs to be done for removal of unwanted observations and for deleting duplicate or irrelevant values from dataset. Above mentioned Models are used in this projects which predicts the chances of breast cancer.

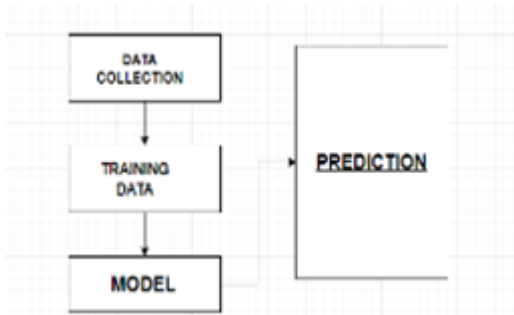


Fig -5: Work flow

### 4. CONCLUSION

In this paper, different types of models are reviewed and their accuracies are computed and compared with each other, so that the best cancer prediction model can be used by doctors in real life to identify breast cancer relatively faster than previous methods. Above examined writing study, proposed that the Random Forest Classification algorithm is proficiently utilized and efficient for detection of breast cancer as compared to Decision tree and Logistic Regression algorithms.

### REFERENCES

[1] Shomona G. Jacob, R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data through Data Mining Techniques", *Proceedings of*

*the World Congress on Engineering and Computer Science 2012*, vol. I, October 2012.

[2] C.L. Chi, W. N. Street, W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets", *American Medical Informatics Association Annual Symposium*, pp. 130-134, Nov. 2007.

[3] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113127, 2004.

[4]<https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>

[5]<https://www.ijitee.org/wpcontent/uploads/papers/v8i6/F3384048619.pdf>

[6]<https://ieeexplore.ieee.org/document/7943207>

### BIOGRAPHIES



Mamta Jadhav is currently pursuing B. Tech in Computer Engineering, from Veermata Jijabai Technological Institute (VJTI), Matunga, Mumbai. She has completed her Diploma in Computer engineering from K J Somaiya polytechnic Vidyavihar, Mumbai.



Zeel Thakkar is currently pursuing B. Tech in Computer Engineering, from Veermata Jijabai Technological Institute (VJTI) Matunga, Mumbai. She has completed her Diploma in Computer engineering from K J Somaiya polytechnic, Vidyavihar, Mumbai.



Pramila M. Chawan is working as an Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She has done her B.E. (Computer Engg.) and M.E (Computer Engineering) from VJTI COE, Mumbai University. She has 27 years of teaching experience and has guided 75+ M. Tech. projects and 100+ B. Tech. projects. She has published 99 papers in the International Journals, 21 papers in the National/ International conferences/ symposiums. She has worked as an Organizing Committee member for 13 International Conferences, one National Conference and 4 AICTE workshops. She has worked as NBA coordinator of Computer Engineering Department of VJTI for 5 years. She had written

proposal for VJTI under TEQIP-I in June 2004 for creating Central Computing Facility at VJTI. Rs. Eight Crore (Rs. 8,00,00,000/-) were sanctioned by the World Bank on this proposal.