Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier

Anisha P R¹, Kishor Kumar Reddy C², K Apoorva³ and Meghana Mangipudi C⁴

¹,²,³,⁴Stanley College of Engineering and Technology for Women Chapel Road, Abids, Hyderabad, India.

kishoar23@gmail.com

Abstract. Breast Cancer is one of the most dreadful diseases and is a potential cause of death in women. Late prediction of Breast Cancer may greatly reduce survival chances, and as a solution to that the automatic disease detection system aids the medical field to diagnose and analyze, which offers rapid response, reliability, effectiveness as well as decrease the risk of death. In this paper, we explain how breast cancer can be predicted using a Machine Learning Technique named Random Forest Classifier. This classifier structures the data into numerous trees and obtains a final result i.e., whether a person is at risk of having breast cancer or not. This model has an accuracy of 98%.

Keywords: Breast Cancer, Machine Learning, Random Forest, Classification, Prediction ROC, AUC, Confusion Matrix

1. Introduction

Breast cancer is one of the most dangerous cancers among women. It represents the majority of new cancer cases and cancer-related deaths according to global statistics [1]. It has to be considered as an important health issue in the present world.

Fig 1. Breast Cancer Mortality in US Women (2012-2016)
As observed in figure 2, the percentage of women affected presently is higher between the young age range of 20-50 when compared to 25 years ago [2]. The two figures above help us understand the rising trend of the incidence of breast cancer and the urgent need to come up with a solution to reduce these percentages. Diagnosis at an early stage can improve survival chances. In such cases, an accurate diagnosis is extremely important. Machine Learning (ML) helps us to analyze complicated datasets and works as a great choice for predicting the risk [3].

Some common risk factors playing a big role are family history, personal history, genetic factors. Mammograms can help identify lumps in different parts of the breast. The causes of breast cancer are not well known but it has been observed that the risk increases with age. For example, if we consider personal history, a woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast. A woman has a higher risk of breast cancer if a close female relative suffered from it. Genetic factors include genetic mutations which are the changes to BRCA1 and BRCA2 genes. Other gene changes may also raise breast cancer risk as well. childbearing and menstrual history [4].

The goal of this project is to use the concept of machine learning to predict the risk of breast cancer in patients by considering various parameters like family history, lumps, genetic factors, age, gender and various other body fluids like glucose, insulin, HOMA, leptin, adiponectin, resistin, msp.1 [5]. It classifies the patients into those who may have a risk of getting breast cancer and those may not have the risk. We use the Random Forest Classifier for prediction [6].

2. Data Description

2.1. Data Collection

The data used in this project has been collected from various sources on the internet. The dataset being used in this model consists of 14 columns out of which the 13 columns are considered as the main parameters and the final column consists of the prediction values. The dataset consists of age, gender, BMI, glucose, HOMA, insulin, adiponectin, leptin, resistin, MCP.1, Family History, Genetic Factors, Lumps and Position. The risk of breast cancer is more likely to occur in women and the risk usually increases with age. High levels of glucose, BMI, low insulin can also increase risk. HOMA controls insulin resistance. An excess of the other compounds secreted in the body like leptin, adiponectin, resistin and MCP.1 may also have an effect. Family history can also play a big role in determining risk.
Genetic factors include genetic mutations which are the changes to the BRCA1 and BRCA2 genes. Lumps are tumors which can be found in the breasts. Lumps are usually found by analyzing mammograms. The position of the lump is also taken in the dataset. A preview of the dataset we used in this model is given below. The dataset we use is in csv.

2.2. Exploratory Data Analysis

Data visualization is an extremely necessary skill in machine learning. It is responsible for providing a qualitative understanding of the given data. The visualizations we use in this project are a heat map and a pair plot of all the parameters being considered.

Exploratory Data Analysis (EDA) is used to find patterns and relationships in data. There are many methods which can be used to perform EDA but, in this project, we have chosen to use the pair plot and the heat map. A pair plot allows us to visualize the relationships between each parameter and is very useful when we need to identify trends.

Heat maps are very convenient to use when understanding complex data sets. Heat map is a 2D representation of data in which values can be represented by different color schemes. Heatmaps are mainly used to find strong relationships between various parameters as they are a great indicator of correlation [7].

The heat map we use to represent data in this model is given below.

![Heat Map](image-url)
We can understand the following from the heat map. The colors indicate the correlation. Darker the color, higher the correlation. Similarly, lighter the color, lower the correlation. Another method of visualization used in our model is the pair plot between all the parameters.

Pair plot is used to figure out the best set of features to explain the relationship between two variables. Every parameter is plotted with the other remaining parameters. For example, age is plotted with age, gender, genetic factors, lumps, family history. This is repeated for every parameter. As we are considering five parameters, we obtain a total of twenty-five graphs. A legend is also added for better clarity [8].

The pair plot figure is given below:

![Pair Plot](image)

**Fig. 4. Pair Plot**

3. Model Building

3.1. Data Preprocessing

Step 1 in model building is data preprocessing. Data preprocessing is the conversion of unstructured data into structured data.

The steps in data preprocessing are
- Reading the dataset
- Check for missing values and fill with required data
- Splitting data into dependent and independent data.
- Label encoding [9]
- One hot encoding (Binarization) [10]
- Splitting the independent and dependent values into train and test sets.
3.2. Proposed Algorithm – Random Forest Classifier

Random forest classifier initially selects a random subset of data and generates numerous decision trees. It then summarizes the votes from different decision trees and then takes the decision to decide the final classification of the test object [11].

A single decision tree has more probability to lead to an error, but when many decision trees are involved in the classification process, we observe that the error reduces and accuracy increases.

This algorithm uses the concept of weights when considering the effect of each output/decision from any of the decision trees. A tree with high error is given low weight and tree with low error is given high weight [12].

Parameters for Random Forest Classifier:
- n_estimators: Specifies the number of trees we use in the algorithm.
- criterion: “Gini” and “Entropy” are the two criterions that we can use.
- min_samples_split: Refers to the minimum number of working set size at node that is required to split. Default value is taken to be 2.

3.3. Evaluation Metrics

Evaluation metrics are responsible for explaining the performance of a model.

It is extremely important to check the accuracy of the model before computing the predicted values. Different types of metrics can be considered while evaluating the models [13].

We draw the Reverse Operating Characteristics (ROC) to figure out the accuracy. The Area Under Curve (AUC) gives us the accuracy. AUC is drawn by plotting the False Positive Rate (FPR) against the True Positive Rate (TPR) [14]. The ROC of this model is given below.

As observed in figure, the value of AUC is 0.98 indicating that the accuracy of the model is 98% which is very desirable.

Another way of evaluation metrics is using the confusion matrix.

A confusion matrix is an M by M matrix, where M is the number of classes we are trying to predict. In this model we have two classes, one class represents the patients at the risk of breast cancer and the other class represents the patients not at the risk of breast cancer. Therefore, the confusion matrix we obtain in a 2x2 matrix [15].
Confusion matrices are typically used with class output models. According to the confusion matrix, the correct or positive values are more than the wrong or negative values making the model more accurate [16].

The confusion matrix in this model is

\[
\text{array([[23, 0],
           [ 1, 20]], dtype=int64)}
\]

In our confusion matrix, TP=23, FP=0, FN=1, TN=20

4. Results and Conclusion

| S.no | Algorithm                                | Accuracy |
|------|------------------------------------------|----------|
| 1    | Logistic Regression                       | 95.8     |
| 2    | Decision Tree                            | 97.2     |
| 3    | K-Nearest Neighbors                      | 95.4     |
| 4    | Naive Bayes                              | 50.3     |
| 5    | Support Vector Classification (using RBF) | 95.4     |
| 6    | Random Forest Classification             | 98.0     |

**Table 1.** Algorithms tested for the dataset

The Random Forest classification algorithm is implemented to train and test the model. The prediction can be 1 or 0. If the prediction obtained is 1, the result displayed on the user interface is ‘Person is at risk of being diagnosed with breast cancer in the future’. If the prediction obtained is 0, the result displayed on the user interface is ‘Person is not at risk of being diagnosed with breast cancer in the future’. Thus, classification takes place.

The false positive rate (FPR) is plotted against the true positive rate (TPR) to attain the Reverse Operating Characteristics (ROC). The area under this curve we draw is called the AUC (Area Under Curve). After testing this model, the AUC was found to be 0.98 which indicates that this model provides an accuracy of 98% which is desirable.

To improve the model, more parameters can be included in the dataset to make the model more comprehensive and useful. For example, the readings obtained in the mammograms can also be added to analyze the physical dimensions of the lumps and give more information about the stage of breast cancer giving us an extra layer of classification. To improve the model further, every parameter can be given a weightage and the percentage of risk can be calculated to give a more numerical prediction.

Using the algorithm of Random Forest Classifier, classification between patients at risk and those who are not at risk is done. A prediction is made and displayed. The patients are classified and a final accuracy of 98% is obtained.

References

[1] DeSantis, C., Ma, J., Bryan, L. and Jemal, A., 2014. Breast cancer statistics, 2013. CA: a cancer journal for clinicians, 64(1), pp.52-62.

[2] Agarwal, G. and Ramakant, P., 2008. Breast cancer care in India: the current scenario and the challenges for the future. Breast care, 3(1), pp.21-27.

[3] Islam, M.M., Haque, M.R., Iqbal, H. et al: Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290 (2020). https://doi.org/10.1007/s42979-020-00305-w

[4] Velentgas, P. and Daling, J.R., 1994. Risk factors for breast cancer in younger women. Journal of the National Cancer Institute. Monographs, (16), pp.15-24.

[5] Ramik Rawal: Breast Cancer Prediction using Machine Learning, Journal of Emerging
Technologies and Innovative Research (JETIR), vol. 7, Issue 5, 13-24(2020).

[6] Joseph A. Cruz, David S. Wishart.: Applications of Machine Learning in Cancer Prediction and Prognosis, Sage Journals, vol. 2, (2006).
https://doi.org/10.1177%2F117693510600200030

[7] Wilkinson, L. and Friendly, M., 2009. The history of the cluster heat map. The American Statistician, 63(2), pp.179-184.

[8] https://seaborn.pydata.org/generated/seaborn.pairplot.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[11] Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

[12] Azar, A.T., Elshazly, H.I., Hassanien, A.E. and Elkorany, A.M., 2014. A random forest classifier for lymph diseases. Computer methods and programs in biomedicine, 113(2), pp.465-473.

[13] Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Huang, S., Brooks, M., Lee, M.J. and Asadi, H., 2019. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. American Journal of Roentgenology, 212(1), pp.38-43.

[14] Marzban, C., 2004. The ROC curve and the area under it as performance measures. Weather and Forecasting, 19(6), pp.1106-1114.

[15] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

[16] Luque, A., Carrasco, A., Martin, A. and de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, pp.216-231.