Breast Cancer Prediction using Machine Learning

Parth Panchal[1] Siddhant Mishra[2] Harsh Panchal[3] Ashish Yadav[4]
[1,2,3,4]Student
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Thakur Polytechnic, Maharashtra, India

Abstract— Approximately about 1 in 8 U.S. women, roughly about 12.4% will develop invasive breast cancer over the course of her lifetime, according to the study by non-profit organization breastcancer.org [1]. It becomes absolutely necessary to find some countermeasures that can detect such cancer in early stages so as to prevent it from becoming fatal.
Key words: Breast Cancer Prediction, Machine Learning

I. INTRODUCTION

Artificial Intelligence, which is also referred as AI refers to the cognitive ability of non-human objects such as computers to think and take decisions similar to a human being. The method by which computers are provided with this ability is known as Machine Learning or ML. With the help of ML, the computers can be instructed to do a particular task but without being explicitly programmed by the developers to do so. The Machine Learning trend has its influence in almost any background today. Whether a company needs to analyze its data or to predict the stock market price, Machine learning can serve the purpose quite efficiently. Adding to its increasing influence, Machine Learning has also changed the perspective of Medical Sciences. The main aim of this project is to develop a machine learning model which can take all the input patients' data and apply algorithm in order to take decisions whether the cancer is recursive or non -recursive, either malignant or benign. Thus, by utilizing machine learning, the model can be trained to learn the various parameters which decide whether cancer is positive or negative, recursive or non-recursive.

II. TECHNOLOGY

The most important aspect of machine learning is the data that it is fed with. In short, the accuracy of the machine learning model depends upon the training data. The Breast Cancer dataset used for this model is obtained from Wisconsin [2] Breast Cancer dataset. The features used to train the model are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. It describes characteristics of the cell nuclei present in the image.

A. The Attribute's in the Dataset

1) The ID numbers
2) Diagnosis (M = malignant, B = benign)
3) It uses Ten real-valued features that are computed for each and every cell nucleus:
4)
   a) The radius (that is; mean of distances from center to points on the perimeter)
   b) The tumor texture (standard deviation of gray-scale values)
   c) The perimeter of the tumor cells
   d) The adjoining area
   e) The smoothness (local variation in radius lengths)
   f) The compactness (perimeter^2 / area - 1.0)
   g) The concavity (severity of concave portions of the contour)
   h) The concave points (number of concave portions of the contour)
   i) The symmetry
   j) The fractal dimension (coastline approximation)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were calculated for each image, resulting in total of 30 features. For instance, field number 3 is Mean Radius, field number 13 is Radius SE, and field number 23 is Worst Radius. All above feature values are recoded with a total of four significant digits. Class distribution: is 357 benign, 212 malignant

III. WORKING OF PROPOSED MODEL

Initially the model will require a cancer dataset that must be give as input into the machine learning algorithm. There are two ways to do this, the dataset can be uploaded as a text file and also the data can be fetched directly from the server. Once the cancer data is provided to the model, its training can begin. The next step which will be performed is clustering [4]. Clustering is the process of grouping similar data into a single group or a cluster. For Clustering, the Nearest Neighbor Algorithm [5] is used by the model. The clustering process will group all the patients with recurrence cancer into a single group and patients with non-recurrence cancer into another group. Once clustering is done, the next step is Classification [6]. The classification process identifies the severity of the cancer that is, whether it is a stage 1 or stage 2 or stage 3 cancer. The classification process is done by Naive Bayes Classification Algorithm [7]. Once the model applies the classification and clustering techniques, the result of the model will be displayed with the help of a graph which will clearly predict Brest Cancer probability with the help of features that are specified during its training phase.

IV. BACK-END USED BY THE MODEL

XAMPP is an abbreviation of Cross-Platform (X), Apache (A), MariaDB (M), PHP (P) and Perl (P). It is a simple and easy to use, lightweight Apache distribution model that makes it extremely easy for developers and programmers to create a local web server for testing and deployment their projects. All the training cancer data is stored on XAMPP localhost server.

V. FRONT-END USED BY THE MODEL

The model uses Java programming language to develop the GUI. The Java AWT and Swing packages are used to develop the various GUI components such as frame, buttons, scrollbars, graphs and windows. All the algorithms and program logic are implemented using Java.

VI. OTHER COMPONENTS REQUIRED

- Java JDK (Development Kit) 1.8+
- Java JRE (Runtime Environment)
- PhpMyAdmin
- Personal Computer
- XAMPP localhost server

VII. FUTURE SCOPE

With the help of such prediction model, it will be easy to study and detect the breast cancer elements at a much early stage, thus the chances of fatal deaths can be reduced to great extents. Also, the development of more and more accurate machine learning algorithms will drop the tolerance or error level, which will ensure precise results.

VIII. FEATURES

1) The model will input the patient's breast cancer report data.
2) The algorithm will perform analysis of the provided data (clustering and classification).
3) A large amount of data can be analyzed in the graphical form which makes it much easier.
4) Breast Cancer can de detected at a much earlier phase using such model.

IX. CONCLUSION

In this review, we discussed the terminologies of ML and outlined their application in breast cancer prediction. Majority of the studies that have been put forth the last years focus on the development of predictive models using supervised ML methods and classification algorithms that aim to predict valid disease outcomes. Based on the studies and analysis of their results, it is prominent that the integration of multidimensional data, combined with the application of different techniques and measures for feature selection and classification can provide effective tools for research in the cancer domain.

REFERENCES

[1] https://www.breastcancer.org/
[2] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
[3] https://en.wikipedia.org/wiki/Breast_cancer
[4] https://en.wikipedia.org/wiki/Clustering
[5] https://en.wikipedia.org/wiki/Nearest-neighbor_chain_algorithm
[6] https://en.wikipedia.org/wiki/Clustering
[7] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

