Breast Cancer Prediction

Anurag Khanra
Computer Science
PES University
Bangalore, India
anurag.khanra306@gmail.com

Hanuraag Ravilla Baskaran
Computer Science
PES University
Bangalore, India
rbhanuraag01@gmail.com

Harshavardhan Veeranna Navalli
Computer Science
PES University
Bangalore, India
harshavardhannavalli3@yahoo.com

Prakruti Pruthvi Kumar
Computer Science
PES University
Bangalore, India
ppk20142001@gmail.com

***Abstract*—Various studies have been conducted on the use of different methods of data mining methods and different classifiers as well. These methods have been discussed in the following pages, with reviews of different papers on the subject.**

***Keywords*—breast cancer, prediction, machine learning, classification, tumor,**

I. Introduction

The given problem we are trying to solve is the accurate detection of breast cancer. This is particularly important as breast cancer is the most invasive cancer for women and is the second highest cause of death for women all over the globe. The development of data mining methods and the development of more efficient methods of prediction, classification and clustering helps with accurate and much faster detection of breast cancer for patients. This helps curb misdiagnoses, which saves valuable time for treatment. Hence, computer aided diagnosis (CAD) has started to play a huge role in the timely detection, and by extension, the prevention of this disease.

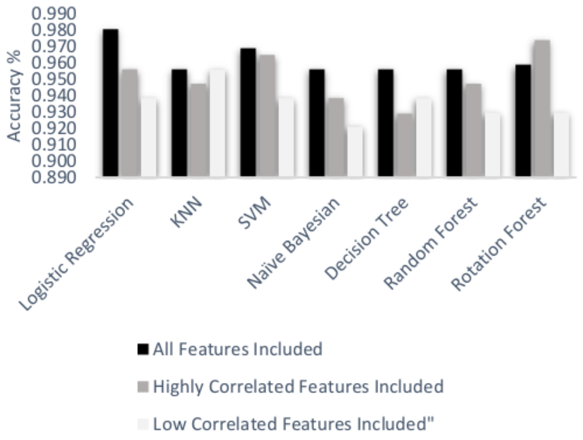
II. Literature Review

This paper looks into various classification models to evaluate their performance in classifying the nature of breast tumors. In this particular study, 5 models are chosen, which are Decision Trees (DT), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Neural Network (NN). The R language is the tool used to implement the given algorithms, which are also classic algorithms in the field of Machine Learning and Artificial Intelligence. The databases used are the Breast Cancer Coimbra Dataset (BCCD) and also the Wisconsin Breast Cancer Database (WBCD), the latter of which is considered the definitive dataset for analysis of breast cancer. This research is of particular importance as the accuracy of the model chosen can literally make the difference between life and death for some patients, as a misdiagnosis could lead to mistreatment and for people to lose valuable time to curing.

The given dataset is divided into the test data set and the training data set, with a 3:7 ratio, i.e., 30 percent test data and 70 percent training data. RF is employed to build the classification model, and a baseline is registered. Following this, the other models are compared (SVM, DT, LR and NN) to the baseline set by the RR model. The baseline set by the measure of accuracy and the F-measure metric. The higher F-measure metric signifies a higher efficiency of the model. When the results are plotted onto a histogram, the RF model shows the highest F-measure metric, and shows the highest accuracy as well. For this reason, RF is chosen as the primary analytical model. To verify this, the ROC curve is plotted, and the AUC is taken, which verifies RF model is the most accurate classifier. [1]

The given dataset is divided into the test data set and the training data set, with a 3:7 ratio, i.e., 30 percent test data and 70 percent training data. RF is employed to build the classification model, and a baseline is registered. Following this, the other models are compared (SVM, DT, LR and NN) to the baseline set by the RR model. The baseline set by the measure of accuracy and the F-measure metric. The higher F-measure metric signifies a higher efficiency of the model. When the results are plotted onto a histogram, the RF model shows the highest F-measure metric, and shows the highest accuracy as well. For this reason, RF is chosen as the primary analytical model. To verify this, the ROC curve is plotted, and the AUC is taken, which verifies RF model is the most accurate classifier. [2]

They analyzed the Wisconsin's breast cancer dataset and classified the tumors as benign or malignant by using different machine learning techniques. They believe that identifying the tissue type (benign or malignant) is important as the both of them have different treatments. After cleaning and visualization, they generate 3 data sets, first having all the features, second having highly correlated features and third having features with low correlation. The Machine Learning algorithms used on these datasets to classify the tissue type were logistic regression, kNN, support vector machines, naive Bayes, decision tree, random forest and rotation forest. Accuracy results were obtained for each algorithm on all the different generated datasets and results were compared. It was observed that logistic regression gave the highest accuracy among all algorithms. The graph below shows that Logistic Regression had a 98% accuracy. [3]



In this paper, the researchers have concentrated on creating a prediction system that can predict the incidence of breast cancer at an early stage by analysing the smallest set of attributes that has been selected from a clinical set. They achieved classification accuracy of 99.28% by comparing actual to predicted values. They have used Correlation-Based measures to get a good feature subset that contains features highly correlated with predictions of the class. The performance of the proposed system was evaluated by considering the actual and the predicted classification. They compared the results achieved with different classifiers used, such as KNN Accuracy, Linear regression Accuracy and SVM Accuracy. According to their paper, they concluded that KNN classifiers yield the highest classification accuracy when used with most predictive variables. The proposed system reduces the cost of treatment and improves the quality of life by predicting breast cancer at an early stage of development. [4]

In this paper the researchers have chosen Colon cancer as their main issue, this is due to very less presence of early symptoms. The main method chosen here is Naive Bayes classification. This is a technique prediction based on simple probabilistic and on the application of the Bayes theorem. This model has an accuracy classification of 95.24% achieved. They have chosen a dataset from Al - Islam hospital Bandung Indonesia to establish Naive Bayes. The weakness of Naive Bayes is that the assumption of independence between attributes reduces accuracy, because there is some part of the data where the attributes are related to each other. [5]

They initially summarised all 6 kinds of breast cancer and proceeded to discuss the various machine learning models which included but are not limited to Logistic Regression, ANN, KNN, Decision Tree, Naïve Bayes, SVM. The paper also contains a survey in Breast Cancer with various data on Ductal Carcinoma in Situ (DCIS) breast cancer cases, Breast Cancers deaths, etc. Authors used the Wisconsin original dataset that was collected from UCI machine learning repository. The dataset used had 10 attributes with 458 benign and 241 malignant patients, three major matrices were designed on the basis of two classes: actual healthy and actual not healthy. Comparison of five nonlinear machine learning algorithms including Multi-Layer Perceptron (MLP), K Nearest Neighbours (KNN), Classification and Regression Tree (CART), Support Vector Machines (SVM) and Gaussian Naive Bayes was done for breast cancer detection. To predict the breast cancer on tumour cells, authors used the deep learning technique with different activation functions: Tanh, Rectifier, Maxout and Exprectifier, to provide the comparative analysis with machine learning algorithms such as Naive Bayes, Decision Tree, Support Vector Machine (SVM) and Random Forest. They also discussed the overview of how different papers were selected to review breast cancer predictions. After summarizing the different ML/DL methods they provided a Comparative review of techniques for breast cancer prediction. From the given chart it was inferred by us that Multi layered Perceptron and Support Vector Machines seemed to have the highest accuracy. Post that there was a Comprehensive review of major machine techniques (in terms of breast cancer prediction) which discussed the benefits and limitations of various models. [6]

III. SUMMARY

We have taken 6 research papers in a bid to understand the different methods used and the different classifiers used during implementation. The first paper concluded, using proper testing and training datasets that Random Forest as a classifier was the most accurate classifier among 5 classifier algorithms. The second paper suggested using artificial neural networks as the classifier using dominance-based feature filtering approach. The third paper suggests using Logistic Regression has a 98% accuracy, and should be used as the classifier. The 4th paper suggests using K-Nearest Neighbors as the suggested classifier, while the 5th paper suggests using Naive-Bayes classifier, but also points out the weakness that it assumes there is independence between attributes, which reduces accuracy. According to the last paper, Multi-Layer Perceptron have the highest accuracy, at 99.6%.

IV. PROBLEM STATEMENT

Breast cancer prediction - classifying the tumour between malignant and benign.

We intend to address the issue of the timely detection of the tumour, and based on various parameters such as the dimensions of the tumour, the surface of the tumour and other factors to predict whether or not the tumour is malignant or benign.

V. REFERENCES

[1] ‘Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction’, Yixuan Li, Zixuan Chen, 2018

[2] ‘Breast Cancer Prediction Using Dominance-based Feature Filtering Approach: A Comparative Investigation in Machine Learning Archetypes’, Noreen Fatima, Li Liu, Sha Hong, Haroon Ahmed, 2020

[3] Muhammet Faith Ak, ‘A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications' , 26 April 2020.

[4] Madhu Kumari and Vijendra Singh ‘Breast Cancer Prediction System’, 2018.

[5] Nafizatus Salmi and Zuherman Rustam, Naive Bayes Classifier Models for predicting colon cancer, 2019

[6] Noreen Fatima, Li Liu, Sha Hong and Haroon Ahmed, ‘Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis’, 2020