Machine Learning Techniques for Prediction and Classification of Breast Cancer

Monalisha Pattnaik
Umrah Naushad
Deepti Rani Pattanaik
Sambalpur University
(monalisha_1977@yahoo.com)
(umrahnaushad@suniv.ac.in)
(97deeptirani@gmail.com)
Satabhisa Debata
Gangadhar Meher University
(dsatabhisa@gmail.com)
G Mariam Rao
Utkal University
(mariamrao94@gmail.com)

ISBN: 978-1-943295-22-7

Breast cancer continues to be a menace in the healthcare industry, grappling not only women but also a significant number of males. Every form of cancer begins with a tumour and can be classified into two types like malignant and benign. It is the malignant that turn out to be cancerous, growing abnormally. This study aimed at predicting and classifying the tumours into the said two groups. Related to this, machine learning techniques like decision tree and linear discriminant analysis are applied to identify 6 risk factors out of 30 different variables. The outcomes of this study are backed by substantial significance.

Keywords: Breast Cancer, Diagnosis, Decision Tree, Linear Discriminant Analysis, Risk Factors

1. Introduction

Breast cancer (after lung cancer) poses as a menace to a majority of women’s health (Yarabarla, 2020). According to World Health Organization, about 2.3 million women were diagnosed with breast cancer with as many as 685,000 reported deaths globally. Even though there have been measures taken by numerous organizations in high-income nations, the age-standardized death rate from breast cancer has shown a decline of 40% between the 1980s and 2020, it continues to grapple the healthcare industry. However, a meagre 2-4 per cent annual reduction in breast cancer mortality has also been attained by the countries that have been successful in doing so (WHO Report, 2023). Breast cancer affects one in every thirty-nine women and is a deadly disease with a high death rate—it accounts for 2.5% of all fatalities (Munshi et al., 2024). Since breast cancer can spread to other parts of the body if left untreated, early detection and treatment are essential. While there have been countless efforts have been made in terms of diagnosis, response and treatment, there still lags extensive prognosis. Up to 80% more people have an increased chance of survival with early detection and treatment (Munshi et al., 2024). With recent advancements in statistics and specifically, machine learning, new areas for the arena around healthcare (Islam et al., 2020). Machine learning algorithms have so far shown reliable results in terms of identification and prediction of breast cancer at the very onset (Singh, 2020). Tumour development is the first stage of all cancers. Two types of tumours can be identified: benign and malignant. Cancer is the term used to describe the malignant ones. Thus, cancer is a particular kind of disease that results from cells growing in a malignant way. The world's health continues to face a serious threat from this aberrant and uncontrollable cell proliferation. It is one of the most difficult diseases to diagnose, treat, and manage since it is not only complex but also heterogeneous (Yarabarla et al., 2019). Depending on where the malignant tumour originated, there are over 100 different varieties of cancer, each with its distinct traits, causes, and risk factors. The study aims at identifying, predicting and categorising the tumours into the said two groups’ i.e. benign and malignant. In line with the mentioned aim, this study uses machine learning techniques, preferably, Decision Tree and Linear Discriminate Analysis to analyze 6 risk factors out of 31 different variables. The variables such as mean of concave points, worst of concave points, mean concavity, mean of texture of area of the tumour cells and the like are taken into account throughout the study. The results show an increasing trend in the amount and quality of research, demonstrating the usefulness of AI as a supportive tool to clinical reasoning for more accessible and reasonably priced healthcare (Moore, 1988). Since MRI screening can identify pre-invasive tumours, premalignant lesions, and pre-invasive malignancies, it is an excellent method of patient follow-up, particularly for high-risk populations (Ahmad et al., 2022). The study's ensemble approach works exceptionally well, displaying a 97.66% accuracy rate, highlighting the enormous potential of AI and ML in breast cancer prediction. Better patient outcomes and physician judgement are promised by the model's enhanced capabilities(A. Sharma et al., 2024).The TPE-optimized Borderline-SMOTE Light GBM model performs exceptionally well as compared to previous studies (Omotehinwa et al., 2023). The Voting classifier fared better than the other classifiers, as seen by its high accuracy of 98.77%. Having the lowest mistake rate increases its classification efficacy for breast cancer (Uddin et al., 2023). Moreover, a Flask and React website was developed, utilising the best approach for improved usability and accessibility (Uddin et al., 2023).The incorporation of tumour size, age, lymph node metastases, and clinical stage III into the decision support tool significantly improved the prediction of breast cancer recurrence (Sharma et al., 2022). The Decision Tree and XGBoost classifiers achieved the highest accuracy of any model, at 97%. The classification of breast cancer by the XGBoost classifier was remarkably successful, with an Area Under the Curve (AUC) score of 0.999 (Nemade & Fegade, 2022).

2. Methodology

2.1 Decision Tree Analysis

Decision trees are very popular tools for classification and prediction problems. A decision tree is a classifier which recursively partitions the instance space or the variable set. Decision trees are represented as a tree structure where each node can be classified as either a leaf node or a decision node. A leaf node holds the value of the target attribute, while a decision node specifies the rule to be implemented on a single attribute-value. Each decision node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes-values. Each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. After implementing the rule on the decision node, a sub-tree is an outcome. Each of the leaf nodes holds a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path.

2.2 Linear Discriminant Analysis

Discriminant analysis is used to distinguish distinct sets of observations and allocate new observations to previously defined groups. For example, if a study was to be carried out in order to investigate the variables that discriminate between fruits eaten by (1) primates, (2) birds, or (3) squirrels, the researcher could collect data on numerous fruit characteristics of those species eaten by each of the animal groups. Most fruits will naturally fall into one of the three categories. Discriminant analysis could then be used to determine which variables are the best predictors of whether a fruit will be eaten by birds, primates, or squirrels. Discriminant analysis is commonly used in biological species classification, in medical classification of tumors, in facial recognition technologies, and in the credit card and insurance industries for determining risk. The main goals of discriminant analysis are discrimination and classification. The assumptions regarding discriminant analysis are multivariate normality, equality of variance-covariance within group and low multicollinearity of the variables. Through the analysis of 32 risk factors associated with breast cancer, we aim to develop a predictive model for enhanced diagnosis.

3. Data Specification

Cancer is a complex disease, with tumors categorized as either benign or malignant. This study specifically focuses on breast cancer, one of the most commonly diagnosed forms. The following discussion centers around the application of machine learning techniques in the classification and prediction of breast cancer at its very onset.

Breast Cancer Wisconsin data set from the *UCI Machine learning repo* is used to conduct the analysis. Decision tree (DT) analysis and a linear discriminatant function (LDF) are constructed to predict and classify new observations. Using this dataset of 30 of independent variables measuring the size and shape of cell nuclei, goal is to create a model that will allow us to predict whether a breast cancer cell is benign or malignant. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Our dataset consists of 569 observations and 32 variables. There is an ID variable, a diagnosis variable revealing if they were benign or malignant and 32 variables. There is an ID variable, a diagnosis variable revealing if they were benign or malignant, and 30 measurement variables detailing the size and shape of the cell nuclei. The diagnosis, a categorical variable, is our response variable and the 30 measurement variables, all of which are continuous, are our potential explanatory variables for our model.

Also can be found on UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. The attributes information are:

1. ID Number
2. Diagnosis (M = Malignant=1, B = Benign=0)
3. Radius_Mean
4. Texture_Mean
5. Perimeter_Mean
6. Area_Mean
7. Smoothness_Mean
8. Compactness_Mean
9. Concavity_Mean
10. Concave.Points_Mean
11. Symmetry_Mean
12. Fractal_Dimension_Mean
13. Radius_Se
14. Texture_Se
15. Perimeter_Se
16. Area_Se
17. Smoothness_Se

18. Compactness_Se
19. Concavity_Se
20. Concave.Points_Se
21. Symmetry_Se
22. Fractal_Dimension_Se
23. Radius_Worst
24. Texture_Worst
25. Perimeter_Worst
26. Area_Worst
27. Smoothness_Worst
28. Compactness_Worst
29. Concavity_Worst
30. Concave.Points_Worst
31. Symmetry_Worst
32. Fractal_Dimension_Worst

The primary objective of this study is to predict breast cancer by incorporating decision tree analysis and linear discriminate analysis. Further, the classification of breast tumors is done into benign and malignant categories. Additionally, our research extends to predicting the tumor's nature and, consequently, the rate of survival. Statistical properties of the sample data are given in Table 1 below.

Table 1 *Descriptive Statistics of Control Variables of Dataset of Breast Cancer Patients*

Sl. No.	Input (Control) and Output (Response) Variables of 569 no. of Patients	Variable Type	Min	Max	Mean	First Quartile	Median	Third Quartile	Standard Deviation
1	Radius mean	Numerical	6.98	28.11	14.13	11.7	13.37	15.78	3.52
2	Texture mean	Numerical	9.71	39.23	19.29	16.17	18.84	21.80	4.30
3	Perimeter mean	Numerical	43.79	188.50	91.97	75.17	86.24	104.10	24.29
4	Area mean	Numerical	143.5	2501.0	654.90	420.3	551.1	782.70	351.91
5	Smoothness mean	Numerical	0.05	0.16	0.10	0.09	0.10	0.11	0.001
6	Compactness mean	Numerical	0.02	0.35	0.10	0.06	0.09	0.13	0.005
7	Concavity mean	Numerical	0.00	0.43	0.09	0.03	0.06	0.13	0.007
8	Concave points mean	Numerical	0.00	0.20	0.05	0.02	0.03	0.07	0.004
9	Symmetry mean	Numerical	0.11	0.30	0.18	0.16	0.18	0.20	0.003
10	Fractal dimension mean	Numerical	0.05	0.1	0.06	0.06	0.06	0.07	0.001
11	Radius se	Numerical	0.11	2.87	0.41	0.23	0.32	0.48	0.03
12	Texture se	Numerical	0.36	4.89	1.22	0.83	1.11	1.47	0.05
13	Perimeter se	Numerical	0.76	21.98	2.87	1.61	2.29	3.36	2.02
14	Area se	Numerical	6.80	542.20	40.34	17.85	24.53	45.19	45.49
15	Smoothness se	Numerical	0.00	0.31	0.01	0.01	0.01	0.01	0.00
16	Compactness se	Numerical	0.00	0.14	0.03	0.01	0.02	0.03	0.001
17	Concavity se	Numerical	0.00	0.40	0.03	0.02	0.03	0.04	0.003
18	Concave points se	Numerical	0.00	0.05	0.01	0.01	0.01	0.01	0.00
19	Symmetry se	Numerical	0.01	0.08	0.02	0.02	0.02	0.02	0.008
20	Fractal dimension se	Numerical	0.00	0.03	0.00	0.00	0.00	0.00	0.002
21	Radius worst	Numerical	7.93	36.04	16.27	13.01	14.97	18.79	4.83
22	Texture worst	Numerical	12.02	49.54	25.68	21.08	25.41	29.72	6.15
23	Perimeter worst	Numerical	50.41	251.20	107.26	84.11	97.66	125.40	33.60
24	Area worst	Numerical	185.2	4254.0	880.6	515.3	686.5	1084.0	569.36
25	Smoothness worst	Numerical	0.07	0.22	0.13	0.12	0.13	0.15	0.002
26	Compactness worst	Numerical	0.03	1.06	0.25	0.15	0.21	0.34	0.016
27	Concavity worst	Numerical	0.00	1.25	0.27	0.12	0.23	0.38	0.02
28	Concave points worst	Numerical	0.00	0.29	0.11	0.06	0.10	0.16	0.065
29	Symmetry worst	Numerical	0.16	0.66	0.29	0.25	0.28	0.32	0.062
30	Fractal dimension worst	Numerical	0.06	0.21	0.08	0.07	0.08	0.09	0.018

4. Results and Discussions

This section deals with various results and findings obtained out of the prediction and classification of breast cancer. The results are related to decision tree analysis and discriminant analysis.

4.1 Decision Tree Analysis

Decision tree learners are powerful classifiers, which utilize a tree structure to model the relationships among the features and the potential outcomes. In the case a final decision can be made, the tree is terminated by leaf nodes (also known as terminal nodes) that denote the action to be taken as the result of the series of decisions. In the case of a predictive model, the leaf nodes provide the expected result given the series of events in the tree. A great benefit of decision tree algorithms is that the flowchart-like tree structure is not necessarily exclusively for the learner's internal use. After the model is created, many decision tree algorithms output the resulting structure in a human-readable format. This provides tremendous insight into how and why the model works or doesn't work well for a particular task. Diagnosis of medical conditions based on laboratory measurements, symptoms, or the rate of disease progression. Although the previous applications illustrate the value of trees in informing decision processes, this is not to suggest that their utility ends here. In fact, decision trees are perhaps the single most widely used machine learning technique, and can be applied to model almost any type of data—often with excellent out-of-the-box applications.

The motivation behind the choice of decision tree as a potential model to find the significant input variables out of 30 input variables for the diagnosis of breast cancer estimates of the patients is the simplicity, easy interpretability, and high accuracy of the DT algorithm. We apply an optimal DT model to the dataset consisting of 569 different patients and try to find out potential casual variables from the set of available variables that are related to the diagnosis of breast cancer of the patients. DT is implemented using 'rpart' package in R with "minsplit" equals to 10% of the data as a control parameter. We have used RMSE, MAPE, coefficient of multiple determination (R^2), and adjusted R^2 (AdjR^2) to evaluate the predictive performance of the tree model used in this study. An optimal regression tree is built with 7 variables with 'minsplit' = 5 with equal costs for each variable. A variable importance list from the DT is given in Figure 1 and the fitted tree is provided in Figure 2. From the variable importance plot based on the complexity parameter of the DT model (also see Figure 3), seven causal variables are obtained out of 30 potential input variables having higher importance. Our results are consistent with previous results. But interestingly, we obtained seven essential causal variables like radius worst, concave points worst, texture mean, texture worst, concavity mean, area worst and concave points mean can be managed to diagnose against this deadly disease. Once these variables are taken care of, the respective country may diagnose properly of the breast cancer at a significant rate.

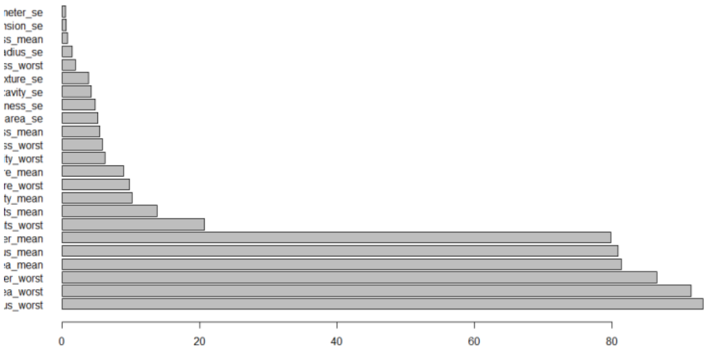

Figure 1 *Feature Importance Percentages Affecting of Breast Cancer Diagnosis Based on a Complexity Parameter Applying Decision Tree (DT)*

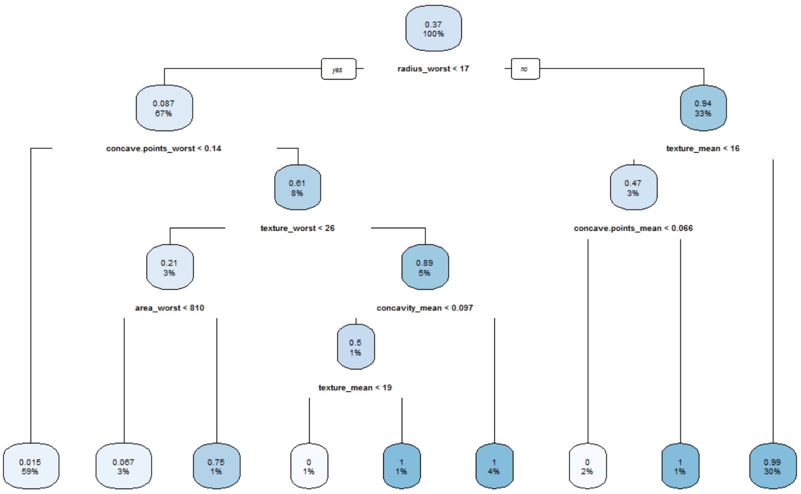

Figure 2 *Optimal Decision Tree Representing the Relationships between the Causal Variables and Breast Cancer Diagnosis*

By applying rpart the range of cost complexity can be evaluated. To compare the error for each cost complexity value rpart performs a 10-fold cross validation so that the error associated with a given cost complexity is computed on the hold-out validation data. Figure 2 shows the optimal tree having 8 internal nodes resulting in 9 terminal nodes. Basically, this tree is partitioning on 7 variables to produce its model. A tree with 9 terminal nodes we can force to generate a full tree by using cp=0 (See Figure 3). In Figure 3 y-axis is cross validation error, lower x-axis is cost complexity value, upper x-axis is the number of terminal nodes. After 9 terminal nodes, we see diminishing returns in Error reduction as the tree grows deeper. To predict the diagnosis of breast cancer of a patient, these seven risk factors out of 30 variables are identified like radius worst, concave points worst, texture mean, texture worst, concavity mean, area worst and concave points mean. All 569 number of patients go through the DT (see Figure 2), are assessed at a particular node, and proceed to the left if the answer is “yes” or proceed to the right if the answer is “no”. So, first all 569 patients whose radius worst is less than 17 go to the left branch, all other patients proceed to the right branch. All the patients whose radius worst is greater than 17 have texture mean is greater than 16 have diagnose of breast cancer is 0.99%. All the patients whose radius worst is greater than 17 have texture mean is less than 16 and have concave points mean is greater than 0.066 have diagnose of breast cancer is 1%. All the patients whose radius worst is greater than 17 have texture mean is less than 16 and have concave points mean is less than 0.066 have diagnose of breast cancer is 0%. All the patients whose radius worst is less than 17 have concave points worst is greater than 0.14 have texture worst is greater than 26 have concavity mean is greater than 0.097 have diagnose of breast cancer is 1%. All the patients whose radius worst is less than 17 have concave points worst is greater than 0.14 have texture worst is greater than 26 have concavity mean is less than 0.097 and have texture mean is greater than 19 have diagnose of breast cancer is 1%. All the patients whose radius worst is less than 17 have concave points worst is greater than 0.14 have texture worst is greater than 26 have concavity mean is less than 0.097 and have texture mean is less than 19 have diagnose of breast cancer is 0%. All the patients whose radius worst is less than 17 have concave points worst is greater than 0.14 have texture worst is less than 26 have area worst is greater than 810 have diagnose of breast cancer is 0.75%. All the patients whose radius worst is less than 17 have concave points worst is greater than 0.14 have texture worst is less than 26 have area worst is less than 810 have diagnose of breast cancer is 0.067%. All the patients whose radius worst is less than 17 have concave points worst is less than 0.14 have diagnose of breast cancer is 0.015%. The splitting process continues, visiting all variables each time a split is made until all the diagnosis of breast cancer patients data are divided into nine partitions with predicted diagnosis of breast cancer patients (0.015, 0.007, 0.75, 0, 1, 1, 0, 1, 0.99) based on only seven input variables namely radius worst, concave points worst, texture mean, texture worst, concavity mean, area worst and concave points mean.

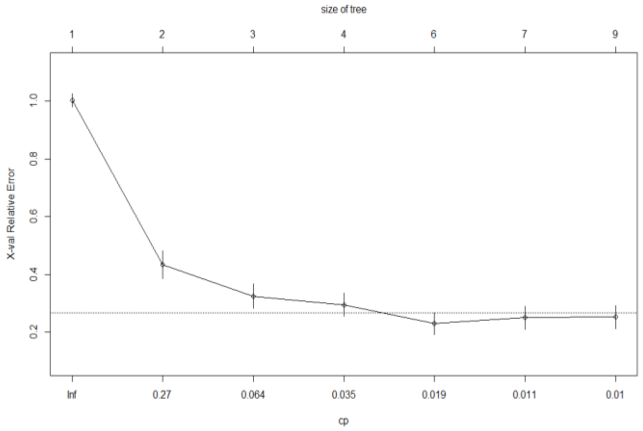

Figure 3 *Complexity Parameter and Size of the Tree verses Cross Validation Error*

4.2 Linear Discriminant Analysis

Discriminant analysis is used to distinguish distinct sets of observations and allocate new observations to previously defined groups. On the collected samples in order to separate the samples from which they were excavated, discriminant function can be used as a function. The function can then be applied to classify and to predict. In R, we can fit a model to data with the lda function. The first part of the output displays the formula that was fitted. The second part is the prior probabilities of the groups, which reflects the proportion of each group within the dataset. In other words, if it had no measurements and the number of measured samples represented the actual relative abundances of the groups, the prior probabilities would describe the probability that any unknown sample would belong to each of the groups. The third part shows the group means, which is a table of the average value of each of the variables for each of the groups. Scanning this table can help you to see if the groups are distinctive in terms of one or more of the variables. The fourth part reports the coefficients of the discriminant function. Because there are two groups, there are 2-1 linear discriminants (if it had only two groups, it would need only 1 [2-1] linear discriminants). For each linear discriminant LD1, there is one coefficient corresponding, in order, to each of the variables. The predict() function, also part of the MASS package, uses the lda() results to assign the samples to the groups. In other words, since lda () derived a linear function that should classify the groups, predict() allows you to apply this function to the same data to see how successful the classification function is. Following the statistical convention that x-hat is the prediction of x, (hat is added to the object name to make it clear that these are the predictions). The output starts with the assigned classifications of each of our samples. Next, it lists the posterior probabilities of each sample to each group, with the probabilities in each row (that is, for each sample) summing to 1.0. These posterior probabilities measure the strength of each classification. If one of these probabilities for a sample is much greater than all the others, that sample is assigned to one group with a high degree of certainty. If two or more of the probabilities are nearly equal, the assignment is much less certain. If there are many groups, the following command is a quick way to find the maximum probability for each sample: Since most of the probabilities in the dataset are large (>0.9), this indicates that most of the samples in the set have been assigned to one group. If most of these probabilities are large, the overall classification is successful. The last part of the predict() output lists the scores of each sample for each discriminant function axis. These scores can be plotted to show graphically in Figure 5 how the groups are distributed in the discriminant function. The two groups occupy distinctly different and non-overlapping regions. There is just one case of group 1 being close to group 2, so one can clearly state that the discrimination has been successful. The plot is as shown in the following Figure 5. Again, note the good separation of the groups along discriminant function 1, and particularly so for group 1. The effectiveness of discriminant function in classifying the groups must be evaluated, and this is done by comparing the assignments made by predict() to the actual group assignments. The table() function is most useful for this. By convention, it is called with the actual assignments as the first argument and the fitted assignments as the second argument. The rows in the output correspond to the groups specified in the original data and the columns correspond to the classification made by the discriminant function. In a perfect classification, large values would lie along the diagonal, with zeroes off the diagonal, which would indicate that all samples that belong to group 1 were discriminated by the discriminant function as belonging to group 1, and so on. The form of this table can give you considerable insight into which groups are reliably discriminated. It can also show which groups are likely to be confused and which types of misclassification are more common than others. Then we calculate the overall predictive accuracy, that is, the proportion of cases that lie along the diagonal. The result is 0.9693878. Here the predictive accuracy is almost 97%, quite a success. This approach measures what is called the resubstitution error, how well the samples are classified when all the samples are used to develop the discriminant function. The classification of breast cancer by the LDA classifier was remarkably successful, with an Area Under the Curve (AUC) score of 0.979, see Figure 4.

Linear Discriminant Model is:

diagnosis ~ β_1 ×concave.points_mean + β_2 ×concave.points_worst + β_3 ×concavity_mean + β_4 ×texture_mean + β_5 ×texture_worst + β_6 ×area_worst

diagnosis ~ 17.86 ×concave.points_mean + 14.22 ×concave.points_worst -3.97 ×concavity_mean -0.02 ×texture_mean + 0.08×texture_worst + 0.001×area_worst

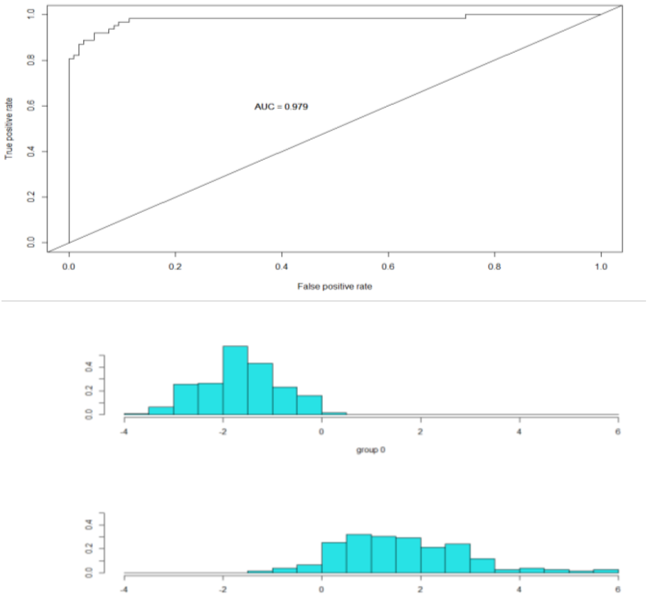

Figure 4 *AUC Results*

5. Conclusions

The motivation behind the choice of decision tree as a potential model to find the significant input variables out of 30 input variables for the diagnosis of breast cancer estimates of the patients is the simplicity, easy interpretability, and high accuracy of the DT algorithm. We apply an optimal DT model to the dataset consisting of 569 different patients and try to find out potential casual variables from the set of available variables that are related to the diagnosis of breast cancer of the patients. An optimal regression tree is built with 7 variables. From the variable importance plot based on the complexity parameter of the DT model, seven causal variables are obtained out of 30 potential input variables having higher importance. Our results are consistent with previous results. But interestingly, we obtained seven essential causal variables like radius worst, concave points worst, texture mean, texture worst, concavity mean, area worst and concave points mean can be managed to diagnose against this deadly disease. Once these variables are taken care of, the respective country may diagnose properly of the breast cancer at a significant rate. With the accuracy rate of 97.9% we anticipate that our findings will not only contribute to the evolving field of medical research but also pave the way for innovative approaches in breast cancer diagnosis, ultimately bolstering our collective efforts in combating this complex and challenging health concern.

6. References

1. Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D., & Olugbara, O. O. (2022). A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Applied Sciences (Switzerland), 12(22). https://doi.org/10.3390/app122211455
2. Ahmad, R., Ahmed, B., & Ahmed, B. (2022). Effectiveness of MRI in screening women for breast cancer: A systematic review. Breast Cancer Management, 11(2). https://doi.org/10.2217/bmt-2021-0016
3. Birchha, V., & Nigam, B. (2022). Performance Analysis of Averaged Perceptron Machine Learning Classifier for Breast Cancer Detection. Procedia Computer Science, 218(2022), 2181–2190. https://doi.org/10.1016/j.procs.2023.01.194
4. El Massari, H., Gherabi, N., Mhammedi, S., Sabouri, Z., Ghandi, H., &Qanouni, F. (2023). Effectiveness of applying Machine Learning techniques and Ontologies in Breast Cancer detection. Procedia Computer Science, 218(2022), 2392–2400. https://doi.org/10.1016/j.procs.2023.01.214
5. Hassan, M. M., Hassan, M. M., Yasmin, F., Khan, M. A. R., Zaman, S., Galibuzzaman, Islam, K. K., & Bairagi, A. K. (2023). A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction. Decision Analytics Journal, 7(April), 100245. https://doi.org/10.1016/j.dajour.2023.100245
6. Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., &Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1, 1-14.
7. Li, J., Chen, Y., Ye, W., Zhang, M., Zhu, J., Zhi, W., & Cheng, Q. (2023). Molecular breast cancer subtype identification using photoacoustic spectral analysis and machine learning at the bio macromolecular level. Photo acoustics, 30, 100483. https://doi.org/10.1016/j.pacs.2023.100483
8. Liu, Y., Fu, Y., Peng, Y., & Ming, J. (2024). Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. Heliyon, 10(2), e24876. https://doi.org/10.1016/j.heliyon.2024.e24876
9. Moore, G. E. (1988). Breast Cancer Diagnosis. Archives of Surgery, 123(8), 1024. https://doi.org/10.1001/archsurg.1988.01400320110026
10. Munshi, R. M., Cascone, L., Alturki, N., Saidani, O., Alshardan, A., &Umer, M. (2024). A novel approach for breast cancer detection using optimized ensemble learning framework and XAI. Image and Vision Computing, 104910.
11. Nemade, V., & Fegade, V. (2022). Machine Learning Techniques for Breast Cancer Prediction. Procedia Computer Science, 218(2022), 1314–1320. https://doi.org/10.1016/j.procs.2023.01.110
12. Omotehinwa, T. O., Oyewola, D. O., & Dada, E. G. (2023). A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. Healthcare Analytics, 4(April), 100218. https://doi.org/10.1016/j.health.2023.100218
13. Rawal, R. (2020). Breast cancer prediction using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR), 13(24), 7.
14. Sahu, A., Das, P. K., & Meher, S. (2023). Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms. Physica Medica, 114(March), 103138. https://doi.org/10.1016/j.ejmp.2023.103138
15. Sharma, A., Goyal, D., & Mohana, R. (2024). An ensemble learning-based framework for breast cancer prediction. Decision Analytics Journal, 10(August 2023), 100372. https://doi.org/10.1016/j.dajour.2023.100372
16. Sharma, D., Kumar, R., & Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. Measurement: Sensors, 24(November), 100560. https://doi.org/10.1016/j.measen.2022.100560
17. Shayea, I., Saoud, B., & Hadri, M. (2024). Machine learning , IoT and 5G technologies for breast cancer studies : A review. 89(December 2023), 210–223. https://doi.org/10.1016/j.aej.2024.01.043
18. Singh, G. (2020). Breast cancer prediction using machine learning. Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol., 8(4), 278-284.
19. Uddin, K. M. M., Biswas, N., Rikta, S. T., & Dey, S. K. (2023). Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. Computer Methods and Programs in Biomedicine Update, 3(February), 100098. https://doi.org/10.1016/j.cmpbup.2023.100098
20. World Health Organization: WHO & World Health Organization: WHO. (2023, July 12). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=Scope%20of%20the%20problem,the%20world's%20most%20prevalent%20cancer.
21. Yadav, R. K., Singh, P., & Kashtriya, P. (2022). Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey. Procedia Computer Science, 218, 1434–1443. https://doi.org/10.1016/j.procs.2023.01.122
22. Yarabarla, M. S., Ravi, L. K., & Sivasangari, A. (2019, April). Breast cancer prediction via machine learning. In 2019 3rd international conference on trends in electronics and informatics (ICOEI), 121-124. IEEE.