# Abstract

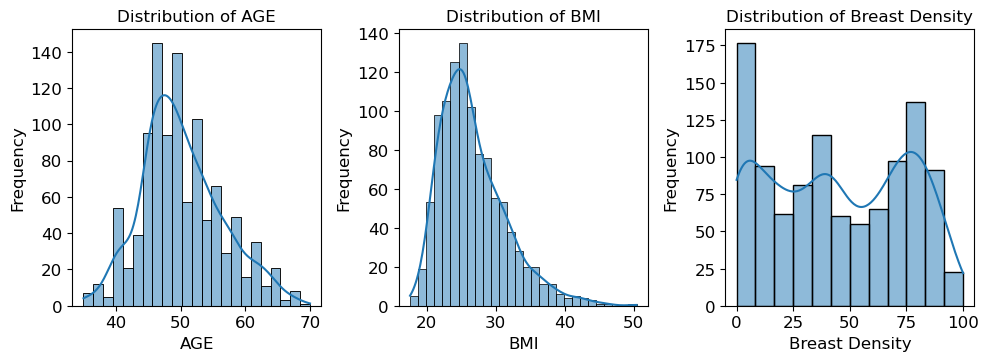
# This study examines the connection between mammographic density and the likelihood of developing breast cancer in a case-control analysis conducted during the IBIS-I trial. The objective is to establish whether breast density is a risk factor for breast cancer and to assess the relationship between breast density and two common demographic risk factors: age and body mass index. To improve the accuracy of breast cancer predictions, we employed machine learning models that used age, body mass index, and breast density as inputs. The study presents the performance of various machine learning models, compares them, and interprets the results. The findings of the study suggest that breast density is an independent risk factor for breast cancer in this case-control study. The study showed that there is a positive correlation between breast density and the risk of developing breast cancer, and this relationship was independent of age and body mass index. The study also discovered that the gradient boosting (GBDT) model performed well in predicting breast cancer, with an area under the curve (AUC) of the receiver operating characteristics of 0.78. In other words, the higher the AUC score, the more accurately the model predicts 0s as 0s and 1s as 1s. However, the study found that another machine learning model, such as random forest, was more likely to make better predictions using the same data and variables, with an AUC score of 0.88. This result implies that machine learning approaches may be more effective in predicting breast cancer risk than traditional statistical models. Such an approach may help identify high-risk individuals for targeted screening and prevention strategies. However, further research is required to validate these findings in other populations and examine the potential clinical usefulness of machine-learning approaches for breast cancer risk prediction.

# Methods:

# The following information was collected from the International Breast Cancer Intervention Study I (IBIS-I), a randomized controlled trial aimed at investigating the use of tamoxifen to prevent breast cancer in women with a higher risk of developing the disease. The data included mammograms, age, body mass index (BMI), and five-year treatment arm with either tamoxifen or a placebo. Mammograms were collected on a case-control basis for a subset of 1065 women, with cases being women who developed breast cancer after their first follow-up mammogram and controls being women who remained breast cancer-free at the time (Cuzick et al., 2011). We removed duplicates and handled missing values to prepare the data for analysis. We transformed the data types to numeric or categorical and scaled the numeric features using the MinMaxScaler method, which scales the data to a fixed range of 0 to 1. We also encoded the categorical features using OneHotEncoder, which encodes the categorical features into binary vectors. We balanced the classes using SMOTE to address the class imbalance in the target variable. We then split the data into training and testing sets and created a pipeline to preprocess the data and build the machine-learning models. We utilized GridSearchCV, a hyperparameter tuning technique, to tune the hyperparameters of the machine learning models. Further, we conducted univariate, bivariate, and multivariate analyses of the features in the dataset. We used statistical t-tests to test several hypotheses, including whether breast density is a risk factor for breast cancer, whether there is a significant relationship between breast density and age, whether there is a significant relationship between breast density and BMI, whether age has any effect on breast cancer, whether BMI has any effect on breast cancer, and whether breast density has an impact on breast cancer. We built several machine-learning models, including Decision Tree, AdaBoost, Random Forest, support vector machine (SVM), XGBoost, Naive Bayes, and Gradient Boosting Decision Tree (GBDT) (Zuo et al., 2023). We evaluated the models' performance using various metrics such as precision, recall, f1-score, sensitivity, and specificity. We also plotted the AUC-ROC curve for all the machine-learning models and used the best model to predict breast cancer. Finally, we interpreted the machine learning models by calculating the logistic regression model's odds ratio and confidence intervals.

# Results

Figure 1 displays female patients' age, BMI, and breast density distribution.The following information presents the distribution of three features - age, BMI, and density. The data on age indicates that the population's average age was around 50 years, with ages ranging from 35 to 70 years. The median age was 49 years, which is close to the average. There weren't many outliers in age, indicating a uniform age distribution. Moving on to BMI, the average BMI was 27, ranging from 17.6 to 50.4. The median (26) was also close to the mean, implying the absence of any significant outliers. Finally, the density distribution showed that the mean density was 44, with values between 0 and 100. The median (40) was lower than the mean, indicating that the density distribution was right-skewed.



**Figure 1 displays female patients' age, BMI, and breast density distribution.**

Figure 2 displays boxplots of three features, age, BMI, and density, arranged according to the target variable case. The left boxplot exhibits age against the case, the middle boxplot depicts BMI against the case, and the right boxplot represents breast density against the case. The boxplot of age indicates that the median age of cancer cases is higher than that of the controls, which is consistent with the known risk factor for breast cancer. Similarly, the boxplot of BMI shows that the median BMI of cancer cases is slightly higher than that of the controls, which is also a known risk factor for breast cancer. The boxplot of density reveals that the median breast density is higher for cancer cases than for non-cancer cases, which again aligns with the known risk factor for breast cancer. Notably, 88.4% of the observations correspond to non-cancer controls, while only 11.6% are cancer cases. This imbalanced the dataset since the number of observations in non-cancer controls is significantly higher than in cancer cases.

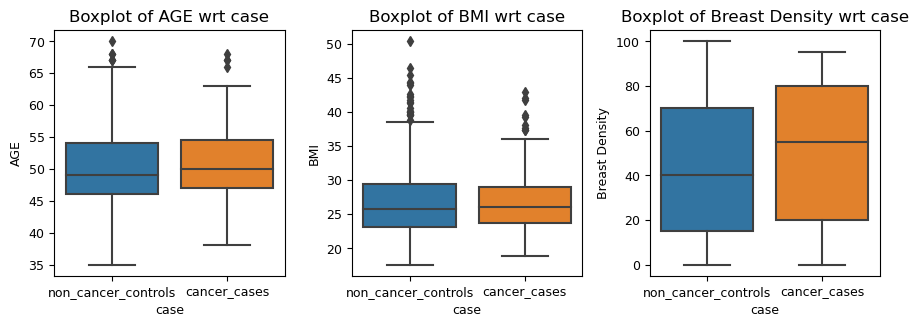


Figure 1 shows boxplots of age, BMI, and density features grouped by the target variable case. The left shows age vs case; the middle shows BMI vs case; and the right shows breast density vs case.

We conducted a study to test the association between breast density and breast cancer using the t-test, a statistical test used to determine if there is a significant difference between the means of the two groups. Our analysis showed that breast density is a significant risk factor for breast cancer (p-value = 0.017), with a positive correlation between age and breast density and a negative correlation between BMI and breast density. However, our t-test results showed that age and BMI had no significant effect on breast cancer.

Next, we used a logistic regression model to determine the odds ratio of developing breast cancer based on age, BMI, and breast density. Our analysis showed that for every one-unit increase in age, the odds of developing breast cancer increased by 2.4%. For every one-unit increase in BMI, the odds of developing breast cancer increased by 0.028682. For every one-unit increase in breast density, the odds of developing breast cancer increased by 1.4%.

We further analysed the significance of each coefficient and found that breast density had a significant effect on breast cancer (p-value = 0.001820). We also calculated the confidence intervals of the coefficients and found that we were 95% confident that the actual coefficient of breast density lay between 0.003070 and 0.018356.

We removed the columns with a p-value greater than 0.05 and refitted the model. Our analysis showed that the coefficient of breast density decreased to 0.010713, which meant that for every one-unit increase in breast density, the odds of developing breast cancer increased by 1.1%. The p-value of breast density was 6.007977e-03, indicating that breast density had a significant effect on breast cancer. Our study suggests that breast density is a significant risk factor for breast cancer and increases in breast density may lead to higher odds of developing breast cancer.

| **Model** | **Train Accuracy** | **Test Accuracy** | **Precision** | **Recall** | **F1 Score** | **Sensitivity** | **Specificity** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Decision Tree | 0.797023 | 0.720126 | 0.790780 | 0.973799 | 0.872798 | 0.973799 | 0.337079 |  |
| AdaBoost | 0.614853 | 0.517487 | 0.812057 | 0.962185 | 0.880769 | 0.962185 | 0.337500 |  |
| Random Forest | 0.997294 | 0.827044 | 0.904255 | 0.969582 | 0.935780 | 0.969582 | 0.509091 |  |
| SVM | 0.791610 | 0.776730 | 0.851064 | 0.971660 | 0.907372 | 0.971660 | 0.408451 |  |
| XGBoost | 0.693627 | 0.489049 | 0.840426 | 0.979339 | 0.904580 | 0.979339 | 0.407895 |  |
| Naive Bayes | 0.438085 | 0.397635 | 0.503546 | 0.904459 | 0.646925 | 0.904459 | 0.130435 |  |
| GBDT | 0.901576 | 0.539931 | 0.932624 | 0.981343 | 0.956364 | 0.981343 | 0.620000 |  |

Table 1: Performance of various machine learning models.

Table 1 shows the performance of various machine learning models, including Decision Tree, AdaBoost, Random Forest, SVM, XGBoost, Naive Bayes, and GBDT. The decision tree model had a training accuracy of 0.8 and a test accuracy of 0.72. Its precision was 0.79, the recall was 0.97, the f1 score was 0.87, the sensitivity was 0.97, and the specificity was 0.34. The adaboost model had a training accuracy of 0.61 and a test accuracy of 0.52. Its precision, recall, f1 score, sensitivity, and specificity were 0.81, 0.96, 0.88, 0.96, and 0.34, respectively. The random forest model had a training accuracy of 1.0 and a test accuracy of 0.83. Its precision, recall, f1 score, sensitivity, and specificity were 0.9, 0.97, 0.94, 0.97, and 0.51, respectively. The SVM model had a training accuracy of 0.79 and a test accuracy of 0.78. Its precision, recall, f1 score, sensitivity, and specificity were 0.85, 0.97, 0.91, 0.97, and 0.41, respectively. The xgboost model had a training accuracy of 0.69 and a test accuracy of 0.49. Its precision, recall, f1 score, sensitivity, and specificity were 0.84, 0.98, 0.9, 0.98, and 0.41, respectively. The Naive Bayes model had a training accuracy of 0.44 and a test accuracy of 0.4. Its precision, recall, f1 score, sensitivity, and specificity were 0.5, 0.9, 0.65, 0.9, and 0.13, respectively. The gradient boosting (GBDT) model had a training accuracy of 0.9 and a test accuracy of 0.54. Its precision, recall, f1 score, sensitivity, and specificity were 0.93, 0.98, 0.96, 0.98, and 0.62, respectively.

Based on the performance metrics, the Random Forest model has been found to be the best among all models. It has achieved the highest test accuracy, precision, recall, f1-score, sensitivity, and specificity. Here are the performance metrics of the Random Forest model: Test Accuracy: 0.83, Precision: 0.9, Recall: 0.97, F1 Score: 0.94, Sensitivity: 0.97, and Specificity: 0.51. These metrics indicate that the model has accurately predicted 83% of the test data, 90% of the positive cases, and 97% of the positive cases, respectively.

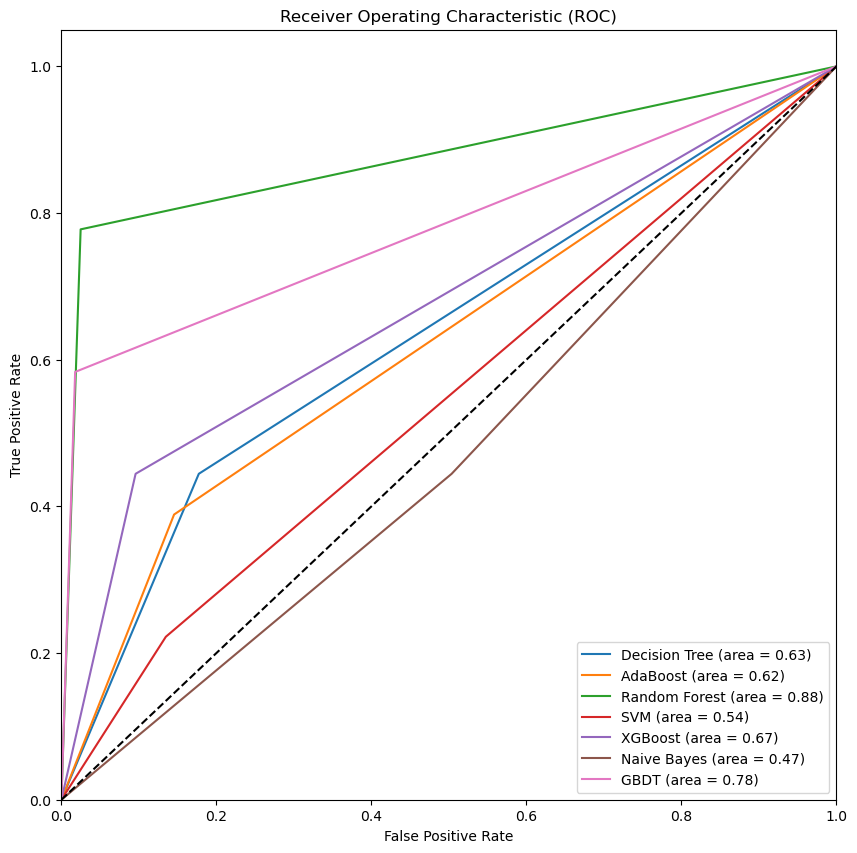


Figure 2: AUC-ROC curve for all the models.

# In Figure 3, you can see the AUC-ROC curve for all the models. AUC-ROC stands for the Area Under the Curve of Receiver Operating Characteristics. This curve is used to measure the performance of classification models at different threshold settings. The ROC curve is a probability curve, and the AUC tells us how well the model can distinguish between classes. The higher the AUC, the better the model predicts 0s as 0s and 1s as 1s. The AUC-ROC curve is plotted with the True Positive Rate (y-axis) against the False Positive Rate (x-axis). In this case, the Random Forest model has the highest area under the curve (0.88), meaning it performs best. Therefore, we can conclude that the Random Forest model is the best model for this dataset. We will use the Random Forest model to predict breast cancer and non-cancer cases in future work. The model will be evaluated using test data to check its performance. Further research is needed to validate these findings in other populations and to investigate the potential clinical utility of machine-learning approaches for breast cancer risk predictions.

# References

Cuzick, J., Warwick, J., Pinney, E., Duffy, S. W., Cawthorn, S., Howell, A., Forbes, J. F., & Warren, R. M. L. (2011). Tamoxifen-induced reduction in mammographic density and breast cancer risk reduction: A nested case-control study. *Journal of the National Cancer Institute*, *103*(9), 744–752. https://doi.org/10.1093/jnci/djr079

Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02377-z