Predicting the Presence of Breast Cancer using Biomarkers

Introduction

Breast cancer is the second leading cause of cancer-related deaths among women[1], emphasizing the need for further research and the development of new tests for earlier detection. The incidence of breast cancer has risen by 0.6% annually[1], with a sharper 1% increase among younger women[1]. However, from 1989 to 2021, the death rate declined by 42%, likely due to increased research and awareness[1]. Despite this progress, there are still significant issues with current testing methods.

Current testing methods often fail to detect early signs of breast cancer. A literature review on current breast cancer testing found that mammograms, a common initial diagnostic test often conducted after a woman self-reports symptoms, have a sensitivity of 82.5%[2]. This means mammograms miss 20% of positive breast cancer cases, potentially delaying treatment and reducing survival rates. Ultrasounds, another commonly used test, have a sensitivity comparable to mammograms but a low specificity of 62%[2]. This low specificity often leads to unnecessary anxiety and biopsies for women who do not have breast cancer. These results highlight the limitations of current testing methods and the need for improvement.

Early detection of breast cancer significantly increases the 5-year survival rate to 99%[1]. Bloodwork analysis, like the approach discussed in this paper, is an emerging but still an uncommon diagnostic method. By analyzing biomarkers in bloodwork and their correlation with positive and negative breast cancer cases, I aim to train machine learning models to identify the key features that distinguish breast cancer patients. This approach addresses the ongoing challenge of developing methods for earlier and more accessible cancer diagnoses. By improving

survival rates for women with breast cancer, it highlights the potential for blood-based biomarker analysis to improve early detection, not just for breast cancer but also for other types of cancer, paving the way for broader applications in oncology.

Methods

The paper I followed divided the analysis into two components: univariate and multivariate analysis[4]. For the univariate analysis, the prediction power of each variable was evaluated independently. These variables include age, BMI, glucose, insulin, HOMA (homeostasis model assessment), leptin, adiponectin, resistin, and MCP-1. The software packages I used were scikit-learn and scipy, which supported the univariate analysis by performing the Shapiro-Wilk test, Mann-Whitney U test, correlation matrix, ROC analysis, and Youden Index test. These tests helped identify key components that can independently distinguish breast cancer patients by evaluating the biomarker distribution, assessing correlations between variables and classifications, and determining the predictive power beyond random chance.

For the multivariate analysis, I focused on how the variables interact with each other, using feature importance from Random Forest. Using the same software packages, I conducted Logistic Regression, Random Forest, and SVM tests. Logistic Regression fits the data to a logistic function and is particularly effective for linear data, making it a suitable choice given that many biomarkers exhibit linearity. SVM testing maximizes the margins between support vectors and performs well with both linear and nonlinear data. Random Forest, an ensemble learning method, combines multiple decision trees to generate predictions and might be the most effective approach since some biomarkers exhibit nonlinear relationships. For cross-validation,

Monte Carlo cross-validation was performed to validate the sensitivity and specificity of the three tests conducted in the multivariate analysis.

To run the code and evaluate the performance of these tests, ensure that the CSV file, named “dataR2.csv,” is downloaded and placed in the same folder as the code. Additionally, the following software packages must be installed: pandas, matplotlib, math, seaborn, numpy, scipy.stats, and sklearn.

Results

Starting with the univariate analysis using the methods previously described, the results indicated that glucose was the most effective variable at distinguishing between the two groups, followed by insulin, HOMA, and resistin. Glucose showed a significant difference in distribution compared to the classification distribution, as demonstrated by the Mann-Whitney U test. It had an AUC (area under the curve) value of 0.76, indicating a clear distinction between the two classes, and a Youden Index value of .439, showing a reasonable balance between sensitivity and specificity. To further confirm these results, a correlation matrix test was performed, shown in Figure 1, which revealed that glucose has a high correlation with the classification, meaning that as glucose levels increase, the likelihood of having breast cancer also increases. In contrast, factors like age and BMI did not show distinguishable differences, as their AUC scores were below 0.50, and the Mann-Whitney U test indicated no significant difference between their distributions when compared to the classification. This aligns with the understanding that age and BMI alone are not sufficient to determine the presence of breast cancer.

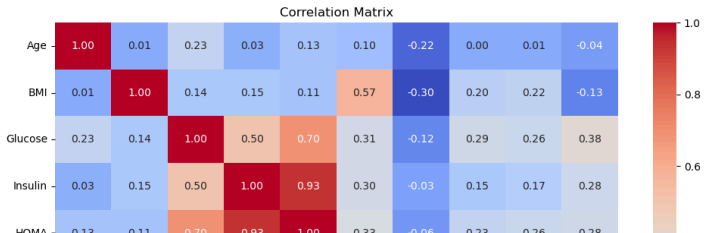

Figure 1. Correlation matrix comparing the independent biomarkers to the classification data.

Following these tests with the multivariate analysis, sensitivity and specificity results were obtained from logistic regression (LR), SVM, and random forest (RF) using Monte Carlo cross-validation. The features were ranked by importance using the feature importance method in random forest. Based on these rankings, the first two through six variables, as well as all nine variables, were tested. After running the iteration fifty times with cross validation, the specificity and sensitivity results are shown in Figure 2.

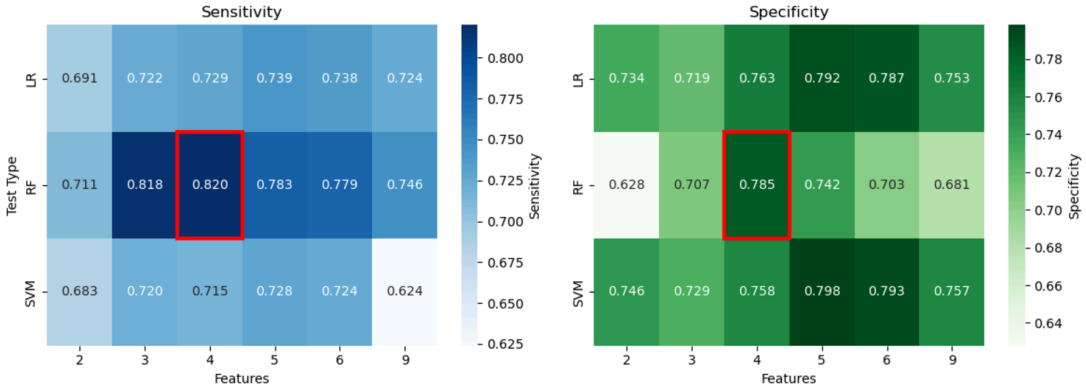

Figure 2. Results of the multivariate analysis, with the highest sensitivity and specificity highlighted in red.

The results indicated that random forest achieved the highest sensitivity and specificity with the top four features: glucose, resistin, age, and BMI. This contrasts with the findings of the paper, where logistic regression performed better for the same four variables. Several factors could explain this discrepancy. First, due to computational limitations, my test was run only 50 times, while the paper's test was conducted 500 times. Second, I did not adjust any hyperparameters for the three tests, which may have influenced the results. Third, because we used Monte Carlo cross-validation, which randomizes the data splits each time, the results are not reproducible, potentially leading to performance variability. Nevertheless, both analyses

identified glucose, resistin, age, and BMI as the most significant variables for distinguishing between the two groups.

The top four variables that produced the best predictive results also align well with biological reasoning. Glucose was the most significant variable in both univariate and multivariate analyses, which is biologically plausible as elevated glucose levels create an environment suitable for tumor growth[3]. Resistin, an inflammatory hormone that accumulates at sites of inflammation, promotes both cellular proliferation and tumor growth[3]. While age and BMI were not independently significant, their interactions with glucose and resistin enhanced the model's predictive power[3]. This finding aligns with biological insights, as increased age and BMI are linked to hormonal and metabolic changes that elevate glucose and resistin levels, reinforcing their role in the predictive model.

These results demonstrate that creating a powerful predictive model does not require incorporating all available variables—in fact, including too many can diminish the model's effectiveness. We found that using only half the variables resulted in a better predictive model than using all biomarkers. Additionally, not all biomarkers need to be purely biological to contribute meaningfully to a breast cancer prediction model, as seen with age and BMI. The findings confirm that biomarkers obtained from blood can reliably identify whether a patient has breast cancer, making them a valuable consideration in early diagnosis. Implementing these testing methods more widely could lead to earlier detection and ultimately improve survival outcomes for women with breast cancer.

Bibliography

1.American Cancer Society. "How Common Is Breast Cancer?" *Cancer.org*, https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html.

2. Bhushan, Arya et al. “Current State of Breast Cancer Diagnosis, Treatment, and Theranostics.” *Pharmaceutics* vol. 13,5 723. 14 May. 2021, doi:10.3390/pharmaceutics13050723

3. Kim, J., Das, R. N., & Lee, Y. (2019). Inter-relationship between diabetes and breast cancer biomarkers. *Oncology and Radiotherapy*

4. Patrício, Miguel et al. “Using Resistin, glucose, age and BMI to predict the presence of breast cancer.” *BMC Cancer* 18 (2018)