Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

The Research About Breast Cancer Prediction Model

Kecheng Liu *

Hangzhou High School international division, Hangzhou, 310000, China

* Corresponding Author: xxb1017@tzc.edu.cn

Abstract. Breast cancer poses a major threat to the health of women worldwide. This study analyzed data from breast cancer patients to develop a predictive model for identifying cases of the disease that are malignant based on cellular measurements. The dataset from the University of California, included 569 instances of 10 numerical variables such as radius, texture, and concavity of the cell samples. After initially exploring the relationships between the variables, the study used methods such as logistic regression and model training. Radius, perimeter and area were integrated since they are positively correlated. Concavity represents concave points, as both describe depressions in the cell outline. Fractal dimension and compactness were combined into a new predictor, F/C. Logistic regression analysis revealed that radius and concavity had the highest prediction accuracies of 87.9% and 88.1%, respectively. Compactness performed moderately well, while the fractal dimension had little diagnostic value. The accuracy of the F/C variable improved by 85.1% over compactness alone. A multi-variable model combining radius, concavity and F/C further improved accuracy and specificity to 92.1%. However, no single variable perfectly predicted cancer diagnosis, suggesting that the data patterns were complex. Further interactions between variables could be uncovered by advanced modelling. In conclusion, the study suggests that composite measures such as radius and concavity are better predictors of breast cancer than isolated factors. More comprehensive clinical data and sophisticated analytic techniques need to be built up to improve diagnostic performance. The model sets the stage for improving breast cancer prognosis through data-driven prediction.

Keywords: Logistic Regression; Prediction Model; Breast Cancer.

1. Introduction

Breast cancer is one of the most common cancers in women. It accounts for about 25% to 30% of all malignancies in women. In many large cities in our country, the incidence of breast cancer has become to the first malignant tumor in women and become the biggest threat to women's health [1]. According to the SEER data, breast cancer incidence has increased at an average annual rate of about 3% since 1980. On average, for every 15 female babies born, one of them will be affected by the breast cancer [2-4]. Based on some studies, there were about 2.6 million new cases of breast cancer in 2018, and there are 600,000 died just because of the breast cancer, which without doubt, makes a serious threat to women's health and the overall society. Fortunately, in some developed countries such as the United States, breast cancer mortality has decreased significantly compared with that in most developing countries because of the widespread screening and some improved treatments [5]. Nowadays, researchers have studied this cancer a lot and divided this cancer into four types: fatty type (type A), oligoglandular type (type B), polyglandular type (type C), and extremely dense type (type D). Approximately 43% of patients around the world were classified as category C. Besides, Asian women tend to have higher breast density [6, 7]. A variety of potential factors have led to breast cancer all over the world, which has brought great influence to different countries. Therefore, through understanding the factors that cause breast cancer, people may prevent the impact of breast cancer and remain health. Actually, there are some common factors that has been proved or discovered by scientists. The first is the genetic factor: a woman who has a family member with a history of breast cancer is two to three times more likely to develop the disease [8]. The second is the life factor, mainly because women have some bad lifestyles, including bad diet, excessive obesity, drug factors and so on [9-11]. Early detection and diagnosis of breast cancer are of great significance to improve the

prognosis and survival rate of patients. A large number of studies have shown that if breast cancer can be found when the tumor diameter is less than 1 cm, the 5-year survival rate can be as high as 90%. However, when the tumor is found to be larger than 5 cm, the 5-year survival rate may drop to about 20%.

Nowadays, the clinical practice models consider limited factors, and their prediction accuracy still needs to be improved. On the other hand, the rapid development of machine learning technology provides the possibility to establish some more accurate breast cancer prediction models. Therefore, this essay aims to establish a breast cancer risk prediction model based on machine learning algorithms. Besides, this passage selects a number of potential related factors, and compare the predictive performance of multiple algorithms in order to establish a more accurate and applicable breast cancer prediction model. In the end, this model also provide reference for the prognosis evaluation and individualized prevention and treatment of breast cancer. The data of 569 breast cancer patients comes from the Kaggle website, including 10 relevant feature data such as the radius, area and smoothness of cancer cells. logistic regression, matrix and other methods were used to establish the model. The test of different variables gathered from the data indicates that the factor radius is the best model for prediction and usually offers a high accuracy than other factors.

2. Methodology

2.1. Data Source

The Breast cancer patients' data used in this paper comes from the UC Irvine Machine Learning Repository website. The original data was saved in .CSV format.

2.2. Data Introduction

The data used in this article contains 569 instances and 10 variables without any missing values. All the 10 variables are represented in the Table 1 and all of them are numeric variables, meaning different kinds of important measured numbers.

Table 1. Different types of variables

Term	Type	Meaning
radius_mean	numeric	mean of distances from center to points on the perimeter
texture_mean	numeric	standard deviation of gray-scale values
perimeter_mean	numeric	mean size of the core tumor
area_mean	numeric	mean area of the tumor
smoothness_mean	numeric	mean of local variation in radius lengths
compactness_mean	numeric	mean of perimeter^2 / area - 1.0
concavity_mean	numeric	mean of severity of concave portions of the contour
concave point_mean	numeric	mean for number of concave portions of the contour
symmetry_mean	numeric	mean symmetry of the tumor
fractal_dimension_mean	numeric	mean for "coastline approximation" - 1

Since there are a total of 10 variables presented in the table1, it seems that the next exploration and analysis will be more complicated. Therefore, the first thing to do is to integrate and simplify the variables. In order to determine which variables can be integrated and simplified, the first thing to do is to explore the relationships between these multiple variables. According to the formula of a circle, it is not difficult to find that there seems to be a relationship between radius, perimeter and area, which affects each other.

The Figure.1. are two scatter plots generated through R studio. The linear relationship between radius and area and the linear relationship between radius and perimeter. These two plots show that both

area and perimeter have a strong positive linear relationship with radius especially when the radius is becoming progressively larger.

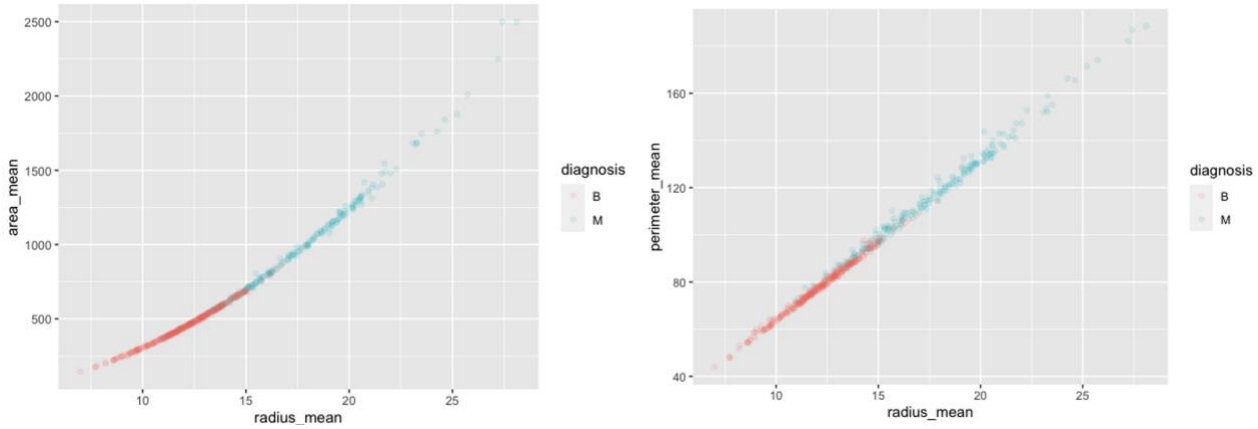

Figure 1. Scatter plots of area~radius & perimeter~radius

2.3. Method Introduction

The method used in this study are logistic regression and model training for each variable. The first step is to fits a logistic regression model predicting diagnosis_bin using each variable. Next, makes predictions on the training data and compares predicted diagnosis to actual diagnosis Finally, calculates number and percentage of correct predictions.

3. Results and Discussion

3.1. Variable Combination

For other variables, as is shown in Figure 2, concavity and concave points, compactness and fractal dimension, smoothness and symmetry have similar distribution. And their scatter plots are shown in Figure 3. Concavity and concave points might have some relationships through their scatter plots. But the origin plot doesn't show notable linear relationship, so this model then uses the log10 to recreate a new plot. Since the new plot shows a better linear ship, it is easy to choose concavity to represent the 2 variables. For fractal dimension and compactness, the plot doesn't show relationship between them. But result shows that if people draw a line from the zero point to the dot, then the slope of the line might be helpful in classification. So, a new variable which is fractal dimension divided by compactness(F/C) is created. For symmetry and smoothness, the plot shows no relationship between them.

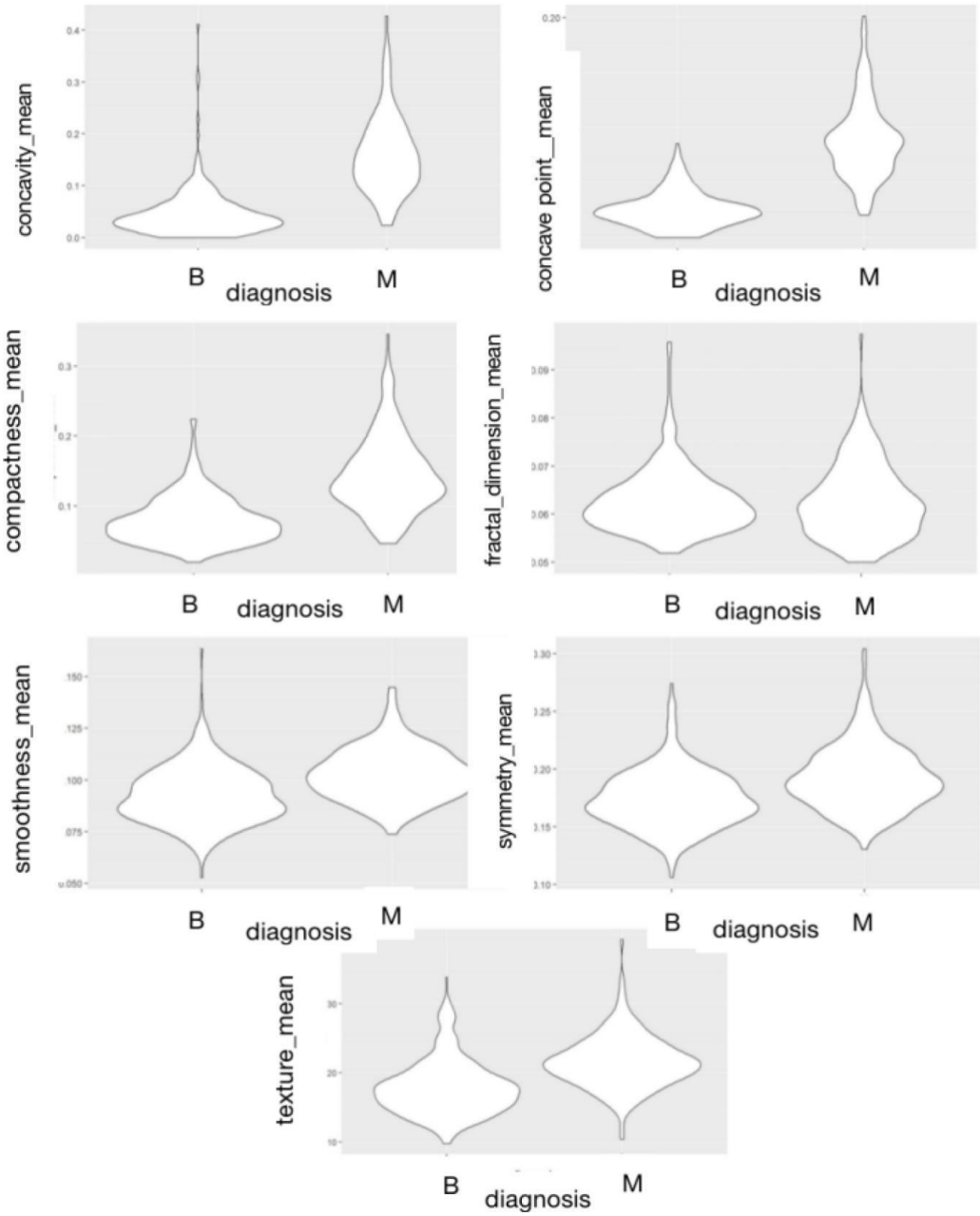

Figure 2. Violin plots of left variables

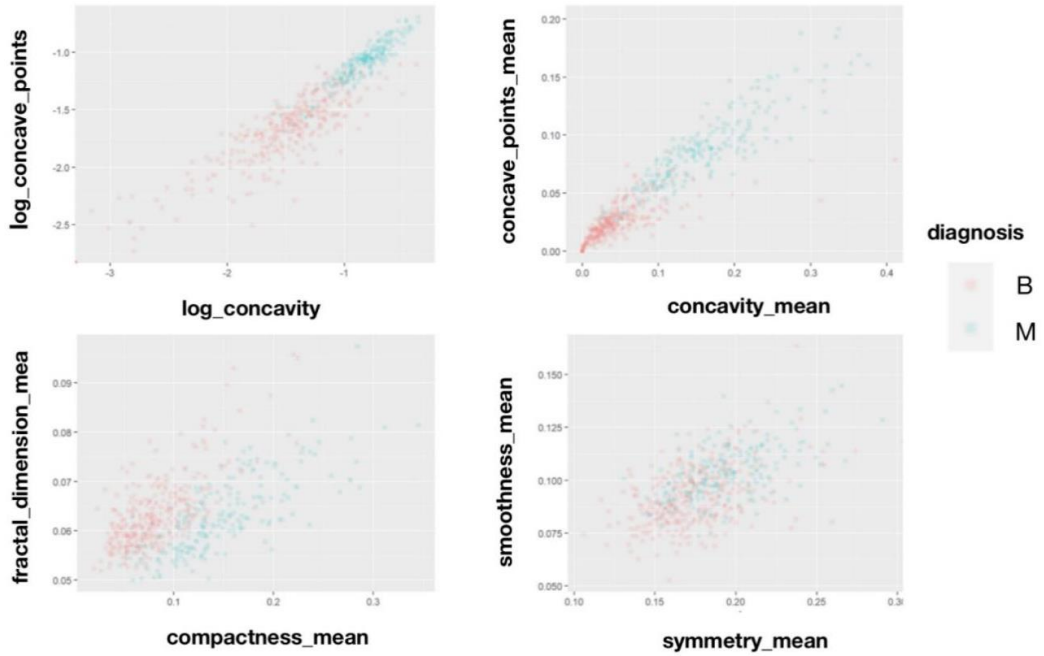

Figure 3. Scatter plots of potentially relative variables

3.2. Logistic Regresstion

Logistic regression modeling uses each variable individually to predict breast cancer diagnosis. The results are presented in Table 2. As shown, radius and concavity yield good predictive performance they have high accuracy in distinguishing between malignant and benign samples. Compactness is a weaker predictor, but still shows some reasonable association with diagnosis. However, fractal dimension does not appear to have any clear relationship with cancer diagnosis based on its poor accuracy.

Table 2. Result from Logistic regression

Variable	Accuracy	Sensitivity	Specificity
radius	0.8787	0.7877	0.9328
texture	0.7047	0.4292	0.8683
smoothness	0.6766	0.3821	0.8515
compactness	0.7966	0.6415	0.8936
concavity	0.8805	0.7877	0.9356
symmetry	0.6189	0.3302	0.8908
fractional dimension	0.6274	0	1
F/C	0.8506	0.8160	0.8711

A new predictor variable "F/C" is taken to the ratio of fractal dimension to compactness. This F/C variable produces improved prediction compared to compactness alone. It achieves higher overall accuracy and sensitivity than compactness, meaning it correctly identifies more malignant samples. However, its specificity is slightly lower than compactness, meaning it mislabels slightly more benign samples as malignant.

In summary, radius and concavity are the strongest individual predictors, while compactness shows moderate predictive signal. The new F/C variable combines fractal and compactness information to improve over compactness alone. But no single variable perfectly predicts diagnosis, indicating complex interactions likely underlie the data. We may need more advanced modeling techniques beyond simple logistic regression to uncover these relationships.

Based on the logistic regression results, we chose three variables, radius, concavity and F/C, to train the prediction model. The results are shown in Table 3. Among the uni-variate models, the radius-

based training model achieved the best prediction results. It has the highest accuracy, sensitivity and special effects. In the multivariate model, although its sensitivity is lower than that of the radius-based model, the overall accuracy and effectiveness are higher. By combining the three most relevant variables, we greatly improved the effectiveness of the model.

Table 3. Result from Training model prediction

	radius total	concavity total	concavity training model	F/C total	multiple total	multiple training model
accuracy	0.8787	0.8805	0.8421	0.8506	0.8506	0.9211
sensitivity	0.7877	0.7877	0.7805	0.8160	0.8160	0.8293
specificity	0.9328	0.9356	0.8767	0.8711	0.8711	0.9726

Overall, radius is the strongest univariate predictor. However, combining the radius, concavity, and F/C variables produces higher accuracy and effectiveness. This suggests that combining multiple related variables can improve prediction performance, and that a single variable is not sufficient to fully explain the complex relationships in the data. Therefore, it needs more advanced modeling techniques to further explore the interactions between variables and predictive patterns.

4. Conclusion

This study shows that radius, perimeter and area are the most helpful variables in predicting tumor type, followed by concavity, concave points and the new variable it constructed, F/C. By combining the use of all these highly correlated variables, the accuracy and specificity of the model can be significantly improved. To further validate the model, it is better to collect more relevant data for testing. In addition, this study use the mean values of the variables instead of the worst values. Compared with the mean value, the worst value group may have higher predictive significance in real situations. It can plan to conduct further studies on the WORST group. Considering the importance of the most severe lesions in the diagnosis of disease, the worst group may better reflect the actual progression of the tumor mean values represent the average of all measurements, but for predicting the progression of cancer, extreme values may bring more important information. In the future, the research need collect more clinical data, including the worst values, train the model, and compare the results to see if the worst values really do have a higher predictive value than the mean values. It will also use more advanced models to analyze potential interactions and nonlinear effects between variables.

References

[1] A.B. Johnson, C.D. Smith, D. Houghton, Recent advances in lung cancer research. Oncology Progress. 2 (1) (2004) 10 - 28.

[2] M. Debever, J. Orel, Lung cancer. Journal of Clinical Oncology. 7 (3) (1991) 339 - 440.

[3] D. S. Ettinger, Lung cancer. Journal of the National Cancer Institute. 7 (1) (1991) 113 - 114.

[4] S. Li, Research status of breast cancer prevention and treatment. Chinese Journal of Lung Cancer. 4 (20) (1993) 122 - 125.

[5] W. Chen, et al. Cancer statistics in China. A Cancer Journal for Clinicians. 66 (2) (2016) 115 - 132.

[6] H. Dai, et al. Characteristics of breast cancer in Central China, literature review and comparison with USA. Breast. 30 (2016) 208 - 213.

[7] S. Muyan, et al. Current status of breast cancer screening in China. The Practical Journal of Cancer. 35 (2020).

[8] L. Xue, Early prevention and detection of breast cancer. Chinese Contemporary Medicine. 17 (2010) 149.

[9] Z. Guo, et al. Study on the correlation of breast cancer. The Chinese Medical Report. 3 (2006) 122.

[10] H. Rong, Y. Aiyun, Research status of risk factors for breast cancer in Chinese women. Practical Preventive Medicine. 3 (24) (2006) 122.

[11] L. Baomin, L. Mei, to discuss the including factors of breast cancer and the corresponding nursing countermeasures. Scientific Advise. (2012) 17.