Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230632

Prediction and feature analysis of breast cancer based on machine learning technology

Haotian Zhang

Aquinas International Academy, Garden Grove, 92845-1025, USA

23040527@hdu.edu.cn

Abstract. Breast cancer, whose incidence rate is increasing year by year, is one of the malignant tumours with the highest incidence rate in women. Every year, an increment of about 1300000 people suffers from breast cancer and 400000 people die from it globally. For the sake of those who may be at risk of breast cancer, it is of critical importance to establish a model that can make predictions of breast cancer. This study utilizes the Random Forest algorithm and Logistic Regression algorithm to construct an analysis model. The study is conducted on a breast cancer dataset that contains data derived from Wisconsin. Specifically, the research conducts feature selection and manages to work out the relationship between various features and tumour types and selects the 5 most significant features. Based on the data of those 5 features, the accuracy of the two models is compared and the Logistic Regression Model is further optimized to reach a higher prediction accuracy. This study is highly significant in the medical community since the model it created can help with breast cancer prediction, allowing for early intervention and a higher survival rate for possible breast cancer patients.

Keywords: Breast Cancer Prediction, Feature Selection, Random Forest Algorithm, Logistic Regression Algorithm.

1. Introduction

Breast cancer is a disease that results from the uncontrollable expansion of breast epithelial cells. It is a lethal illness that raises concern all around the world. According to estimates from the World Health Organization (WHO), 627,000 women will die from breast cancer in 2018. It ranks as the second most frequent cause of death in women [1]. Between 2012 and 2014, the cancer death rate climbed by about 6% [2]. Early attempts to forecast and diagnose tumours are desperately needed in order to reduce the fatality rate brought on by breast cancer and treat it at an early stage.

Many previous studies have been conducted, and various classification schemes as well as clustering algorithms have been employed to predict breast cancer. To deliver the continuous result of particular data, H. Tran employed logistics regression, a supervised learning approach that contains additional dependent variables [3]. Shen et al. chose the most crucial features and used the feature selection method Interactive Autonomy and Collaborative Technologies Laboratory (INTERACT) to create a model and found out that it outperformed the other model in terms of accuracy [4]. Following feature selection, Ahmed et al. found that Decision Tree and Naive Bayes perform better in the Receiver Operating Characteristic (ROC) curve [5]. Support Vector Machines (SVM) are superior because of their high accuracy, whereas expectation maximization has the lowest accuracy, according to Padmapriya et al.'s comparison of many methods [6]. Additionally, Jahanvi Joshi et al.'s study offered an assessment of classification simulations that can be employed for breast cancer diagnosis using the Waikato Environment for Knowledge Analysis (WEKA) tool [7]. Jaffar et al. and Khan et al. created a novel deep-learning-based model to identify as well as classify breast cancer by using mammographic images [8][9].

The major goal of this work is to develop an analysis model that can automatically detect breast cancer using machine learning technologies. Specifically, first, in the preprocessing stage, useful data is extracted from the data set, and conducts data mapping to build binary target values. Second, Random Forest and Logistic Regression are introduced as the analytical model. Logistic Regression has strong interpretability and high controllability. Random Forest does not need to perform feature selection and has a fast-training speed. Both the model in this study can effectively predict breast cancer and the performance and accuracy of Random Forest is better than that of Logistic Regression in terms of predicting breast cancer. The models conducted in this research could greatly help prevent breast cancer in the early stage and improve the survival rate of potential breast cancer patients.

2. Methodology

2.1. Dataset description and preprocessing

The Breast Cancer Wisconsin (Diagnostic) Data Set contains basic information related to the research of 569 people [10]. The dataset consists of two parts, 32 columns in total. The first part is comprised of personnel ID and diagnosis status. The second part includes data on 10 features, which are described by 30 variables. The features are listed as follows: 1. radius 2. texture 3. perimeter 4. area 5. smoothness 6. compactness 7. concavity 8. concave points 9. symmetry 10. fractal dimension. These features are extracted from the digital image of a breast mass, which is created by using fine needle aspirate (FNA). They depict visible characteristics of cells' nuclei. Dependent and independent variables are chosen and the dataset is divided into 2 parts, 20% test and 80% train. The 'id' and 'Unnamed: 32' variables are excluded from the dataset as they are not required for the model. Also, figure 1 is used to represent malignant, and 0 is used to represent benign in order to better utilize data to train models.

2.2. Proposed approach

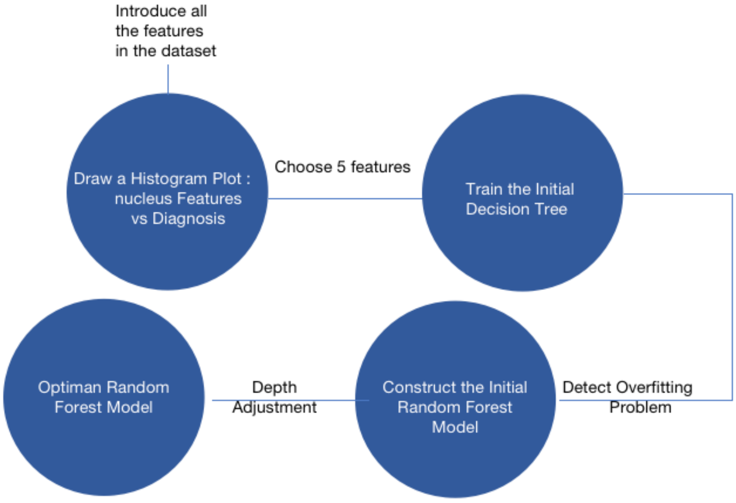
Introduce all
the features
in the dataset
Draw a Histogram Plot : nucleus Features vs Diagnosis
Choose 5 features
Train the Initial Decision Tree
Detect Overfitting Problem
Construct the Initial Random Forest Model
Depth Adjustment
Optiman Random Forest Model

Figure 1. Flow Chart Processing.

This study's primary goal is to create a concise, feasible, and reliable analysis model to automatically identify breast cancer. The research is conducted following the sequence in Fig.1. Firstly, after introducing the data, nucleus features are compared with the diagnosis, and a histogram plot is drawn.

From the histogram, 5 features with the highest correlation with cancer are selected. Second, when compared, Random Forest outperforms Logistic Regression in terms of accuracy and performance when predicting breast cancer. Thus, the Random Forest is chosen for further study. Then, the previously selected five features' data are used to train the decision tree model. Overfitting can be identified by graphically displaying the training and testing accuracy against the depth of the tree. The technique used to prevent overfitting in the study is to restrict the depth of the model. Enhancing generalization capacity requires limiting the depth. In the research, the depth is ultimately set as 7 and constructs the optimal Random Forest Model.

2.2.1. Random Forest Algorithm. Random Forest Algorithm, constructed on the basis of decision trees, is a supervised learning algorithm. Random forests have the ability to improve model performance by reducing overfitting brought on by overfitting data in decision trees. This is a key aspect of random forests.

In general, applying random forest algorithm need three steps. Step one is random sampling. Step two is random feature selection. step three is majority voting. Random sampling, which involves sampling the training dataset with dropout, is the first stage in a random forest. The purpose of random sampling is to reduce the risk of overfitting by averaging multiple random subsets, which can improve the model's generalization ability for new data. Random feature selection means that in each decision tree of a random forest, the feature selection of each node is random, and each node only considers a random subset of features without considering all feature attributes. The findings of all decision trees are combined to produce the ultimate prediction result of a random forest. For classification problems, each decision tree outputs a classification label, and the random forest performs a majority vote on the classification labels output by all decision trees to obtain the final classification result. Generally speaking, Random Forest is an integrated model that can increase model accuracy and decrease data overfitting. To lessen the decision tree's sensitivity to training data, it comprises of many decision trees and employs randomization approaches. The dataset for the study is split into two sets, one is the training and the other is testing sets. Then a model is created for the random forest classifier. Test set data are used to assess the model's accuracy after utilizing data derived from the training set to train the model.

2.2.2. Logistic Regression. Logistic regression is a frequently-used linear classification approach that, in an attempt to achieve classification, translates the results of linear regression to the probability space using a logarithmic probability function. The output of linear regression is mapped to the probability space using the logarithmic function as the activation function in the logistic regression model. And here is the logarithmic probability function's mathematical formula:

$$\mathrm{P}(\mathrm{y}=1|\mathrm{x})=\frac{1}{1+\mathrm{e}^{-(\omega\cdot\mathrm{x}+\mathrm{b})}} \tag{1}$$

where P (y=1 | x) denotes the likelihood that, given the input feature x, namely the weight vector, the data point belongs to the positive class. b stands for the bias term, and e is the base of natural logarithms. The cross-entropy loss, which serves as the loss function for the logistic regression model, is mathematically expressed as follows:

$$\mathrm{L}(\mathrm{x},\mathrm{y})=-\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}[\mathrm{x_i}\log(\mathrm{y_i})+(1-\mathrm{x_i})\log(1-\mathrm{y_i})] \tag{2}$$

where n represents the total number of data points, $\mathrm{x_i}$ is the i-th data point's real label, and $\mathrm{y_i}$ stands for its predicted probability. The Random Forest is configured as follows: there are 100 decision trees in the forest, a minimum of 25 samples are needed to split internal nodes, a tree can have a maximum depth of 7, and a maximum of 2 features are taken into account when determining the optimal segmentation. Next, a logistic regression classifier is created and trained by using a training set. Finally, the test set is used for the prediction and calculation of the accuracy of the model. The advantages of Logistic Regression models are simplicity, high computational efficiency, and strong interpretability.

2.3. Implemented details

The study uses Python 3.11.4 and the Scikit-learn library for implementing decision tree models. Data visualization is done using the Seaborn and Matplotlib libraries. The study is conducted on a macOS device with an Intel Core i5 SoC. The Random Forest is configured as follows: there are 100 decision trees in the forest, a minimum of 25 samples are needed to split internal nodes, a tree can have a maximum depth of 7, and a maximum of 2 features are taken into account when determining the optimal segmentation. With these parameters, the initial random forest model acts well to represent the patterns and connections in the data for further optimization and improvement.

3. Result and discussion

The random forest model is selected after investigation, optimization, and evaluation to guarantee effective classification by identifying key factors. The following processes make up the data analysis procedure to evaluate and improve the random forest model.

First, 10 features are estimated whether they tend to show a correlation with malignant, and histogram plots are made to show the correlation visually. Five characteristics—mean cell radius, mean perimeter, mean area, mean concavity, and mean concave points—show a substantial correlation with the development of malignancy, as can be seen in Fig. 2. So, the research focuses more on those 5 features.

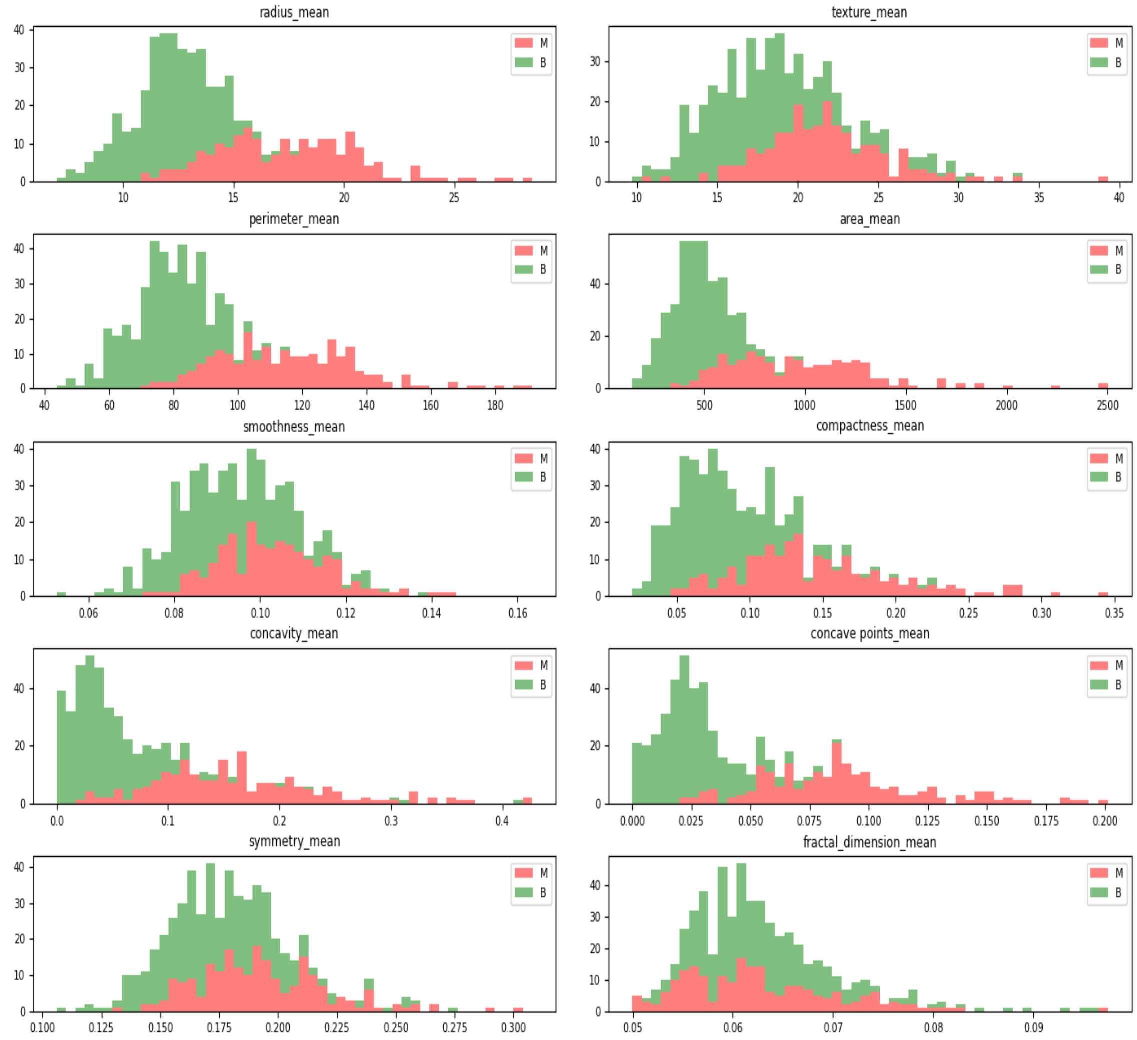

Figure 2. Histogram of the correlation between Feature Data and Malignant Tumor Type.

Second, both models are trained using data linked to the five features mentioned above. The training results reveal that the performance of the Random Forest Model surpasses that of theLogistic Regression Model in terms of accuracy, as can be observed in Tab. 1 after testing the accuracy of the two models and calculating the average accuracy.

Table 1. Accuracy Comparison of two models.

Model	Logistic Regression Model	Random Forest Model
Average Accuracy	86.181%	96.734%

Third, the Random Forest Model is chosen as the final model and creates a feature importance matrix in Table 2. According to Table 2, the data of the previously selected features are ultimately used to train and optimize the model.

Table 2. Feature Importance Matrix (am represents area mean, cpm represents concave points mean, pm represents perimeter mean, rm represents radius mean, con_m represents concavity mean, tm represents texture mean, com_m represents compactness mean, sm_m represents smoothness mean, fdm represents fractal dimension mean, sy_m represents symmetry mean).

Feature	Normalized Correlation Degree
am	0.214733
cpm	0.202946
pm	0.155607
rm	0.153618
con_m	0.116785
tm	0.061337
com_m	0.050892
sm_m	0.024448
fdm	0.012677
sy_m	0.006958

Then, the maximum depth of the Random Forest Model is altered. The maximum depth can range from 1 to 20, and a line chart in Fig. 3 is created. The maximum depth is shown by the horizontal axis, while accuracy is represented by the vertical axis. When the maximum depth is set to 7, according to an analysis of the relationship between the maximum depth and accuracy, the accuracy is at its highest. Through all the processes above, the Random Forest Model is optimized by setting the maximum depth as 7 and training with the data that have the highest correlation with malignancy.

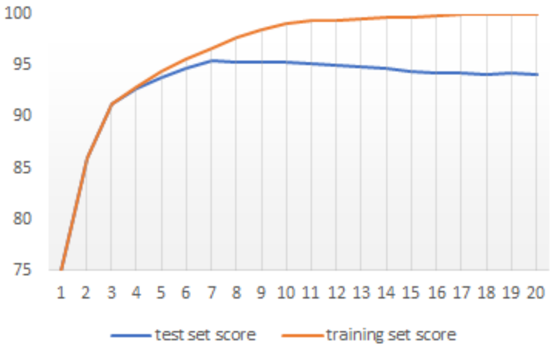

Figure 3. Line chart of accuracy and depth relationship.

4. Conclusion

The study employs the Logistic Regression Model as well as the Random Forest Model to create an efficient breast cancer prediction model. The maximum depth is set at 7 to further increase the model's accuracy in prediction and its ability to successfully reduce overfitting. The optimized model undergoes extensive testing to assess its performance, and the average forecasting accuracy was found to be 96.734%. This indicates that this research has successfully created an accurate model to predict breast cancer and has accomplished the primary objective. The researcher will continue to refine the Random Forest Model for further studies by modifying the maximum number of iterations and the minimal number of samples needed for internal node redistribution. Also, the researcher will try to apply other algorithms to realize the function of predicting breast cancer and try to improve the accuracy of predictions as much as possible.

References

[1] Sun Y S Zhao Z Yang Z N et al. 2017 Risk factors and preventions of breast cancer International journal of biological sciences 13(11): p 1387

[2] De Magalhães J P 2013 How ageing processes influence cancer Nature Reviews Cancer 13(5): pp 357-365

[3] Tran H 2019 A survey of machine learning and data mining techniques used in multimedia system Dept. Comput. Sci., Univ. Texas Dallas Richardson, Richardson

[4] Shen R Yang Y Shao F 2014 Intelligent breast cancer prediction model using data mining technique//2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics. Ieee 1: pp 384-387

[5] Pritom A I Munshi M A R Sabab S A et al. 2016 Predicting breast cancer recurrence using effective classification and feature selection technique//2016 19th International Conference on Computer and Information Technology (ICCIT) IEEE pp 310-314

[6] Padmapriya S Devika M Meena V et al. 2016 Survey on Breast Cancer Detection Using Weka Tool Imperial Journal of Interdisciplinary Research (IJIR) 2(4)

[7] Joshi J Doshi R Patel J 2014 Diagnosis of breast cancer using clustering data mining approach International Journal of Computer Applications 101(10): pp 13-17

[8] Jaffar M A 2017 Deep learning based computer aided diagnosis system for breast mammograms International Journal of Advanced Computer Science and Applications 8(7)

[9] Abdullah A A Hassan M M Mustafa Y T 2022 A review on bayesian deep learning in healthcare: Applications and challenge IEEE Acces 10: pp 36538-36562

[10] Buddhini 2016 reast cancer predictio https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data Kaggle