Scientific Research Publishing

Journal of Biosciences and Medicines, 2023, 11, 252-259
https://www.scirp.org/journal/jbm
ISSN Online: 2327-509X
ISSN Print: 2327-5081

A Novel Method for Diagnosis of Breast Cancer Tumors Based on Random Forest

Mengying Cai

School of Information Engineering, Yancheng Teachers University, Yancheng, China
Email: caimengying2022@163.com

How to cite this paper: Cai, M.Y. (2023) A Novel Method for Diagnosis of Breast Cancer Tumors Based on Random Forest. *Journal of Biosciences and Medicines*, **11**, 252-259.
https://doi.org/10.4236/jbm.2023.114018

Received: March 13, 2023
Accepted: April 21, 2023
Published: April 24, 2023

Abstract

GLOBOCAN 2020 cancer data shows that female breast cancer has become the most common cancer over lung cancer for the first time. As a disease threatening the life safety of women all over the world, how to improve the accuracy of breast cancer diagnosis and help patients get treatment as early as possible is of great importance. This paper introduces a new random forest-based breast cancer diagnosis method (NRFM), using the average radius, average texture, average circumference and other 30 indicators in the nucleus of breast mass as characteristics, to diagnose the benign and malignant breast cancer. NRFM proposed to randomly miss a certain percentage of breast cancer data, using random forest regression to fill in the experiment proved that using the method proposed in this paper, when the proportion of missing data reached 50%, the accuracy of breast cancer diagnosis will be as high as 96.85%. Experiments show that NRFM is easy to understand, convenient to operate, and has practical application value, which can assist doctors to improve the accuracy of breast cancer diagnosis.

Keywords

Random Forest, Breast Cancer Diagnosis, Machine Learning

1. Introduction

GLOBOCAN 2020 released its latest database in December 2020, estimating incidence and death from 36 types of cancer in 185 countries, analyzing differences in cancer by region and gender, and predicting cancer in 2040. According to the GLOBOCAN 2020 cancer data, there were 19,292,789 new cases of cancer worldwide in 2020, of which 9,958,133 cancer patients died. Notably, for the first time, breast cancer in women has surpassed lung cancer as the most common cancer, and breast cancer and lung cancer are the leading causes of death in patients (**Table 1**, **Table 2**), as in [1].

Table 1. Top 10 cancer incidence in the world.

Pathogenetic sequence	Total male and female				
	Cancer species	Number of cases	Proportion (%)	Coarse ratio (1/100,000)	World standard rate (1/100,000)
1	breast Cancer	2,261,419	11.7	58.5	47.8
2	lung cancer	2,206,771	11.4	28.3	22.4
3	colorectal cancer	1,931,590	10.0	24.8	19.5
4	prostate cancer	1,414,259	7.3	36.0	30.7
5	gastric cancer	1,089,103	5.6	14.0	11.1
6	liver cancer	905,677	4.7	11.6	9.5
7	cervical cancer	604,127	3.1	15.6	13.3
8	esophageal carcinoma	604,100	3.1	7.8	6.3
9	thyroid carcinoma	586,202	3.0	7.5	6.6
10	bladder cancer	573,278	3.0	7.4	5.6

Table 2. Top 10 cancer deaths in the world.

deaths sequence	Total male and female				
	Cancer species	Number of deaths	Proportion (%)	Coarse ratio (1/100,000)	World standard rate (1/100,000)
1	lung cancer	1,796,144	18.0	23.0	18.0
2	colorectal cancer	935,173	9.4	12.0	9.0
3	liver cancer	830,180	8.3	10.7	8.7
4	gastric cancer	768,793	7.7	9.9	7.7
5	breast Cancer	684,996	6.9	17.7	13.6
6	esophageal carcinoma	544,076	5.5	7.0	5.6
7	Pancreatic cancer	466,003	4.7	6.0	4.5
8	prostate cancer	375,304	3.8	9.5	7.7
9	cervical cancer	341,831	3.4	8.8	7.3
10	leukemia	311,594	3.1	4.0	3.3

From the perspective of women, breast cancer ranks the first place in the spectrum of morbidity and cause of death in most countries in the world. In 2020, there were 2,261,419 new cases of female breast cancer worldwide, and 684,996 deaths, accounting for 24.5% and 15.5% of the total number of new cases and deaths of female cancer respectively, as in [2], and its incidence ranked the first among female malignant tumors. Early breast cancer has a cure rate of 90%, mid-stage breast cancer has a cure rate of 60% - 80%, and advanced breast cancer only has a cure rate of 10% - 20%, as in [3]. It can be seen that as long as breast cancer is diagnosed at an early stage, there is a great probability of successful cure. Therefore, early diagnosis of breast cancer is malignant or benign is very important.

With the progress of The Times, artificial intelligence has made major breakthroughs in the fields of disease diagnosis and imaging medicine. As a disease threatening the life safety of women all over the world, artificial intelligence has also made great progress in the pathological diagnosis of breast cancer, as in [4]. These AI-based methods can assist doctors in breast cancer diagnosis, improve the accuracy of breast cancer diagnosis, and help patients get treatment as early as possible.

2. Research Status

KHAMENEH *et al.* proposed a machine learning model to segment and classify immunohistochemical (IHC) breast cancer images of breast patients, as in [5]. The feature learning classifier of support vector machine (SVM) was first used for classification, then the segmentation method based on convolutional neural network (CNN) was used for segmentation, and finally the total score of each piece was evaluated. Thus reduce the workload of pathologists, and greatly reduce the medical costs for the purpose.

SHAMI *et al.* developed a morphological based molecar profiling (MBMP) machine learning method, as in [6]. Two groups of queues were used to train the proposed convolutional neural network (CNN) model. To predict ER expression in HE-stained histological images. The results showed that the positive predictive values were 97% and 98%, the negative predictive values were 68% and 76%, and the accuracy were 91% and 92%, respectively, which were not inferior to the transmitted immunohistochemistry (PPV: 91% - 98%, NPV: 51% - 78%, accuracy: 81% - 90%), which showed that tissue morphology was significantly correlated with biomarker molecular expression.

NAIK *et al.* proposed a neural network model based on multi-instance learning to measure the estrogen receptor status (ERS) in stained WSI with a sensitivity and specificity area under the curve (AUC) of 0.92, as in [7]. Because of this study, doctors can use the absorbed technology to avoid some technical problems of antigen repair, and also avoid false negative results caused by insufficient, as in [8]. The method is applied to pathological images, without personal interpretation of doctors, and improves the shortcomings of existing clinicopathological analysis methods, which are long in time and high in price.

JABER *et al.* proposed an ISM classifier and heterogeneous detection system based on WSI by using an image classifier based on neural network architecture, as in [9]. This classifier classifies the subtypes of WSI through the mechanism of voting. The results showed that the WSI classification of the breast biopsy tissue slices of HE chromosome was 65.92% consistent with the molecule-based classification method.

HAMEED *et al.* proposed an integrated method based on deep learning, which classified the pathological patterns of breast cancer tissues of cancer patients and non-cancer patients on 544 complete WSI, and the accuracy rate of breast cancer diagnosis with this method was as high as 95.29%, as in [10]. MAHMOOD *et al.* proposed a multi-stage mitosis detection method based on deep convolutional neural networks, which was applied to two breast cancer histophiologic image datasets, as in [11].

Nowadays, many deep learning algorithms have been applied in breast cancer research and made great progress. However, there are still some obstacles in practical application. One of the main problems is that decisions made by artificial intelligence models are difficult to be understood and explained. Therefore, this paper proposes a new diagnosis method for breast cancer based on random forest, which is easy to understand, convenient to operate, and has more practical application value, and can assist doctors to improve the accuracy of breast cancer diagnosis.

3. NRFM

3.1. Random Forest Profile

Random Forest is a theory that utilizes the bootsrap resampling method. The principle of random forest classification is as follows: firstly, t samples are selected from the original samples, and the sample size of each sample is the same as that of the original training set. Secondly, t samples are trained using the decision tree model to obtain t results. Finally, the results are determined by voting, as shown in **Figure 1** below. By constructing different training sets, random forest can make significant differences among classification models, thus improving the prediction ability of combinatorial classification model, as in [4]. Through T-round training, a sequence of classification models {h is obtained1(x), h2(x)… ht(x)}, using them to form a multi-classification model system, and finally using the voting method to determine the result, classification decision:

$$H\left(X\right)=\arg\max\sum\nolimits_{i=1}^{t}I\left(hi\left(x\right)=Y\right) \tag{1}$$

where, H(x) represents the combinatorial classification model, hi is the single decision tree classification model, and Y represents the output variable. Formula (1) illustrates that the random forest uses the voting method to determine the final result.

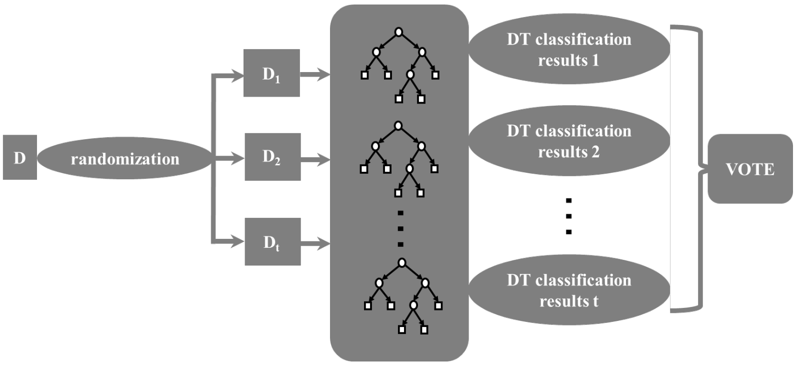
D
randomization
D1
D2
Dt
DT classification results 1
DT classification results 2
DT classification results t
VOTE

Figure 1. RF schematic diagram.

3.2. NRFM

This paper proposes a new method for the diagnosis of breast cancer tumor based on random forest. The design idea of this method is: firstly, through some medical means to obtain the average radius, average texture, average circumference and other 30 indicators data in the nucleus of breast mass of patients. Of course, we do not need to obtain all 30 index data in the nucleus of breast mass of every patient. If some data is missing, it will not affect the use of NRFM at all, because data missing is the more general situation. As we all know, not all index data can be obtained in real cases like experiments, and NRFM also takes advantage of this feature. So it has more practical application value.

When all 30 indicators in the nucleus of the patient's breast mass are obtained, it is called the raw data. First, the raw data is randomly lost in a certain ratio; Then, random forest regression is used to fill in the missing original data, and finally, the filled data is substituted into the model for training.

When the partial data of 30 indicators in the nucleus of the patient's breast mass is obtained, it is called the original data. Random forest regression was directly used to fill in the missing original data, and then the data after filling was used to substitute into the model for training.

In a word, we first identify the original data. If we are missing data, we directly use random forest regression to fill in the missing original data, and then use the data after filling into the model for training. If it is complete data, first of all, the original data will be randomly lost according to a certain proportion; Then, random forest regression is used to fill in the missing original data, and finally, the filled data is used to substitute into the model for training. The NRFM method proposed in this paper is easy to understand, convenient to operate, and has practical application value, which can assist doctors to improve the accuracy of breast cancer diagnosis.

4. Experiments

The data set used in this experiment is the breast cancer data set. The dataset built into sklearn. datasets, included data from 569 patients in the State of Wisconsin on the malignant/benign (1/0) category of breast cancer (training goals), as well as physiological indicators in 30 dimensions. The dataset has 30 features and 569 samples. Firstly, we processed the breast cancer data set, randomly lost a certain percentage of data, and got a missing breast cancer data set. Then, random forest regression, mean, and 0 were used to fill in the missing data. After filling in the data, the random forest algorithm was used for training, and the accuracy rate was selected as the measurement index to compare the effect. The accuracy rate refers to the proportion of the number of different predicted values and actual values of the model. In order to obtain reliable and stable results, K-fold cross-validation was used.

In **Figure 2**, a 50% data loss was performed, and then three methods were used for data filling. It can be seen from the results that random forest regression has the highest accuracy of filling in the missing value, the accuracy of the original data is slightly lower than that of random forest regression, followed by the accuracy of filling in the missing value with 0 and the accuracy of filling in the missing value with the mean value. It can be seen that the method proposed in this paper has a higher prediction accuracy than the original data.

Figure 3 shows the diagnostic accuracy of four data filling methods in the case of different proportions of missing data. As can be seen from Figure 2, no matter in which ratio, the accuracy of using mean and 0 to fill in missing data is not as high as that of the original data. However, the effect of using random forest to fill in the missing data is related to the proportion of missing data. When the missing data is about 32%, the accuracy of using random forest to fill in the missing data is consistent with that of the original data. When the proportion of missing data is lower than 42 values, the accuracy of random forest to fill in the missing data is worse than that of the original data. When the proportion of missing data is higher than 32 values, the diagnostic accuracy of using random forest to fill in the missing data is higher than that of the original data. When the proportion of missing data is 50%, the accuracy of using random forest to fill in

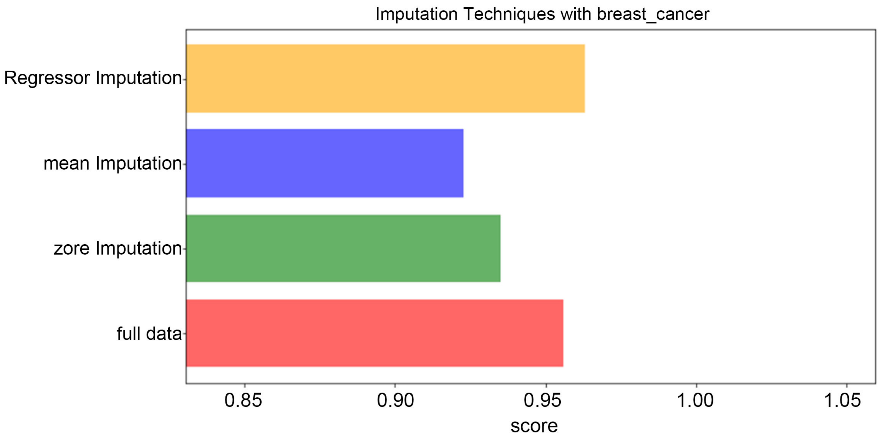

Figure 2. Comparison of the accuracy of different methods in filling in missing values.

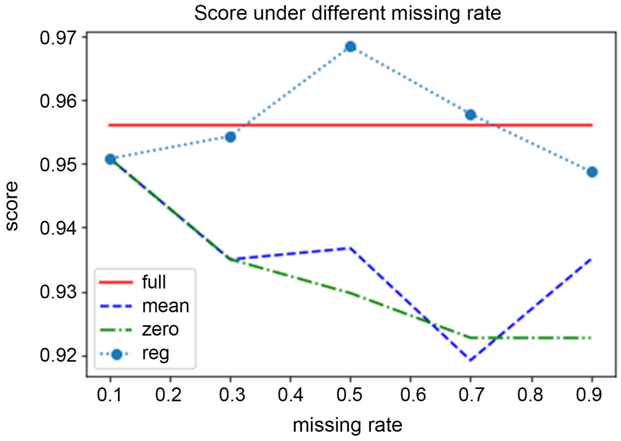

Figure 3. Comparison of diagnostic accuracy of different methods.

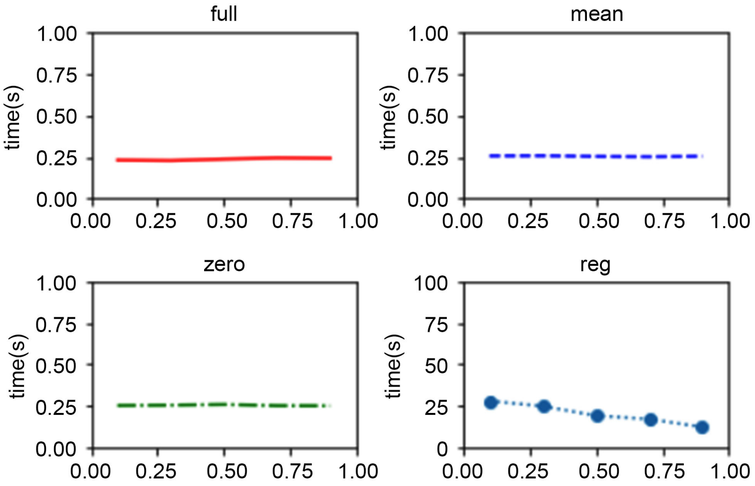

Figure 4. Comparison of time of different methods.

the missing data reaches the highest value. When the proportion of missing data is lower than 72%, the accuracy of using random forest to fill in the missing data is lower than that of the original data. This is because there are too many missing data to make accurate prediction with random forest.

Figure 4 shows the time taken by the four data filling methods in the case of different proportions of missing data. According to the results, the time taken to diagnose breast cancer with full data, mean filling or 0 filling in the missing data was almost the same regardless of the proportions of missing data. The time of filling in missing data with random forest was significantly higher than that of the other three methods, because it took a certain amount of time to fill in missing data with random forest regression.

5. Conclusion

This paper proposes a new method of breast cancer diagnosis based on random forest, which is easy to understand, convenient to operate, and has practical application value, and can assist doctors to improve the accuracy of breast cancer diagnosis. However, there are also shortcomings. When the proportion of missing data is large, random forest regression is used to fill in the missing value, and then the data after filling is used for model training. This operation will produce the phenomenon of overfitting. What is the range of missing data ratio, which can not only ensure the generalization ability of the model, but also improve the accuracy of diagnosis, is the problem we will study in the next step.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Chao, M.M. and Chen, W.Q. (2021) Interpretation of Global Cancer Statistics in

GLOBOCAN 2020. *Chinese Journal of the Frontiers of Medical Science*, **13**, 63-69.

[2] Liu, Z.C., Li, Z.X., *et al*. (2021) Interpretation on the Report of Global Cancer Statistics 2020. *Journal of Multidisciplinary Cancer Management* (*Electronic Version*), **7**, 1-13.

[3] Wang, Z., Wu, Q., Wang, H.L., Zhang L.N., Ning, N. and Zhang L.Z. (2021) Progress in the Application of Artificial Intelligence in the Pathological Diagnosis of Breast Cancer. *Journal of Modern Oncology*, No. 1, 174-177.

[4] Fang, K.N., Wu, J.B., Zhu, J.P. and Xie, B.C. (2011) A Review of Technologies on Random Forests. *Statistics and Information Forum*, **26**, 32-37.

[5] Khameneh, F.D., Razavi, S. and Kamasak, M. (2019) Automated Segmentation of Cell Membrenes to Evaluate HER2 Status in Whole Slide Images Using a Modified Deep Learning Network. *Computers in Biology and Medicine*, **110**, 164-174. https://doi.org/10.1016/j.compbiomed.2019.05.020

[6] Shamai, G., Binenbaum, Y., Slossberg, R., *et al*. (2019) Artificial Intelligence Algorithms to Assess Hormonal Status from Tissue Microarrays in Patients with Breast Cancer. *JAMA Network Open*, **2**, e197700. https://doi.org/10.1001/jamanetworkopen.2019.7700

[7] Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B. and Socher, R. (2020) Deep Learning-Enabled Breast Cancer Hormonal Receptor Status Determination from Base-Level H&E Stains. *Nature Communications*, **11**, Article No. 5727. https://doi.org/10.1038/s41467-020-19334-3

[8] Ehteshami, B.B, Mullooly, M., Pfeifferf, R.M., *et al*. (2018) Using Deep Convolutional Neural Networks to Identify and Classify Tumor Associated Stroma in Diagnostic Breast Biopsies. *Modern Pathology*, **31**, 1502-1512. https://doi.org/10.1038/s41379-018-0073-z

[9] Jaber, M.I., Song, B., Taylor, C., *et al*. (2020) A Deep Learning Image-Based Intrinsic Molecular Subtype Classifier of Breast Tumors Reveals Tumor Heterogeneity that May Affect Survival. *Breast Cancer Research*, **22**, Article No. 12. https://doi.org/10.1186/s13058-020-1248-3

[10] Hameed, Z., Zahia, S., Garcia-Zapirain, B., Aguirre, J.J. and Vanegas, A.M. (2020) Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models. *Sensors*, **20**, 4373-4390. https://doi.org/10.3390/s20164373

[11] Mahmood, T., Arsalan, M., Owais, M., Lee, M.B. and Park, K.R. (2020) Artificial Intelligence-Based Mitosis Detection in Breast Cancer Histopathology Images Using Faster R-CNN and Deep CNNs. *Journal of Clinical Medicine*, **9**, 749-773. https://doi.org/10.3390/jcm9030749