WEP
Warwick
Evans
Publishing

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

Classification of breast cancer data based on neural network algorithm

Mingyu Ouyang *

China Mobile Guizhou Company Limited Qiannan Branch, Guizhou, China, 558000

* Corresponding author: 18485368821@139.com

Abstract. In today's world, breast cancer has become the most common malignant tumor, so this paper mainly classifies the features of breast cancer through BP neural network to improve the accuracy of research and judgment. This paper mainly adopts 30 breast cancer feature data of 569 cases, and then obtains the feature data used in this BP neural network through correlation analysis, feature selection and principal component analysis. By training, the BP neural network that is most suitable for this data is constructed from aspects such as the number of hidden layers, loss function, and iteration times, ultimately improving the accuracy to 100%.

Keywords: BP neural network; correlation analysis; Feature Selection; principal component analysis.

1. Introduction

According to the World Health Organization International Agency for Research on Cancer(IARC)The latest global cancer data in 2020. As shown in Figure 1, there were 19.83 million new cases of malignant tumors worldwide, of which breast cancer surpassed lung cancer to become the largest cancer in the world with 2.26 million cases. Worryingly, however, low and middle-income countries account for nearly three-quarters of breast cancer deaths worldwide.

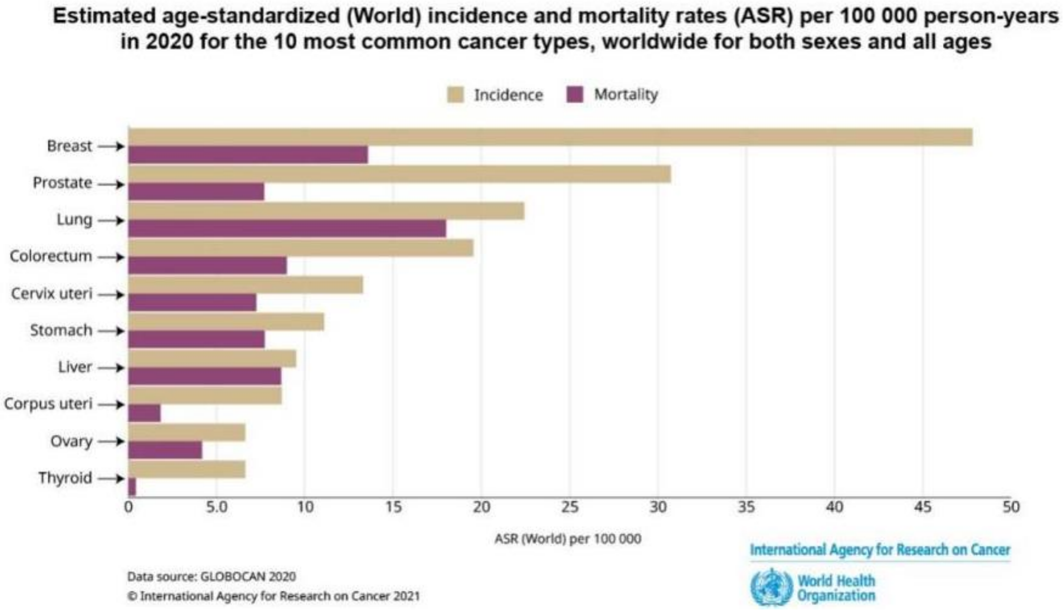

Figure 1. World Health Organization cancer incidence and mortality rankings (Source: IARC official website)

In China, there were approximately 4.57 million new patients with malignant tumors in 2020, and the incidence of breast cancer is still on the rise, especially in the eastern coastal areas and economically developed cities. Especially in rural areas, the decrease in breast cancer mortality is not significant. The main reasons include people's lack of awareness of the disease and the limitations of the medical environment. In this case, the diagnosis and treatment of breast cancer are delayed, resulting in reduced treatment effectiveness.

Because in the traditional manual classification of breast cancer, doctors' subjective judgment will affect the results. Different doctors may have different views and standards, leading to inconsistency in classification results. At the same time, limited by the doctor's personal knowledge and experience, it may not be possible to comprehensively and accurately determine the type of breast cancer, resulting in patients being unable to receive targeted treatment.

This project will use the BP neural network model to analyze medical data through deep learning, build a simple breast cancer classification model, provide intelligent diagnostic support for pathologists, and improve the accuracy of breast cancer pathological diagnosis[1]. It can improve patient diagnosis and treatment outcomes, increase survival rates, and promote the advancement of medical research.

2. Data preprocessing

2.1. Data correlation analysis

Therefore, the data in the above highly correlated feature groups will be deeply explored to select the most representative set of features from each group as necessary features for neural network training[2]. Correlation analysis will be used below to determine the correlation between various features and find the most appropriate features for training the model[3].

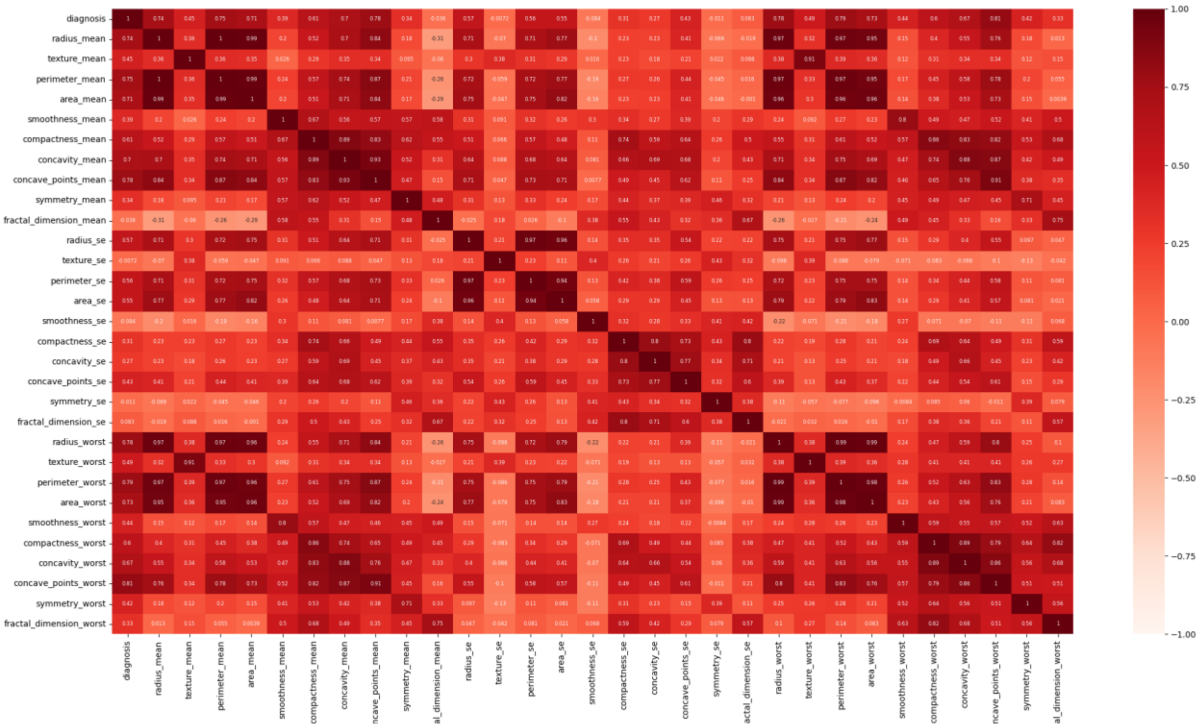

Figure 2. Correlation of each feature

As shown in Figure 2. above, the following conclusions can be drawn:

(1) radius_mean,perimeter_mean,area_mean,radius_se,perimeter_se,area_se,radius_worst,perimeter_worst,area_worst All are highly related. This is extremely meaningful,because perimeter_mean and area_mean is derived using radius as the only variable.

(2) texture_mean and texture_worst highly correlated.

(3) smoothness_mean and smoothness_worst highly correlated.

2.1.1. Deep exploration of data and feature selection

(1) As can be seen from the above summary radius_mean,perimeter_mean,area_mean,radius_se, perimeter_se,area_se,radius_worst, perimeter_worst, area_worst all are highly correlated, so below we select the necessary features for each group that can be used to train the model.

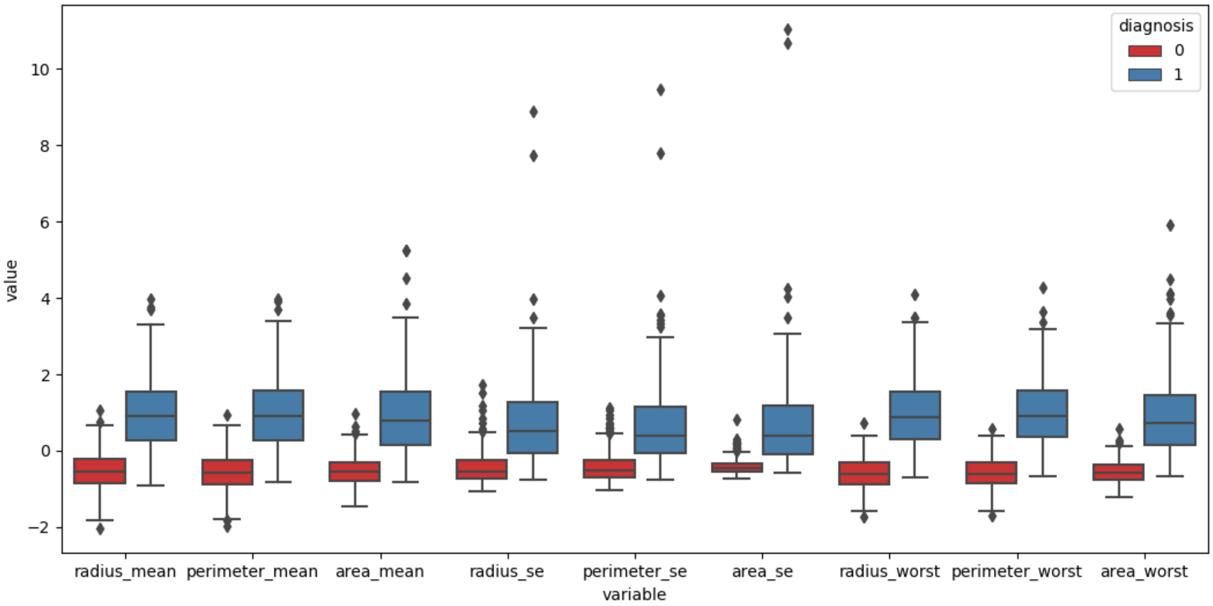

Figure 3. The first set of feature box plots

As shown in Figure 3. above. It was observed that almost all variables are well separated and there is no overlap in the data of the two labels. There are very few variables. but, like: radius_se, perimeter_se and area_se there are many outliers, so the component features will not be adopted.

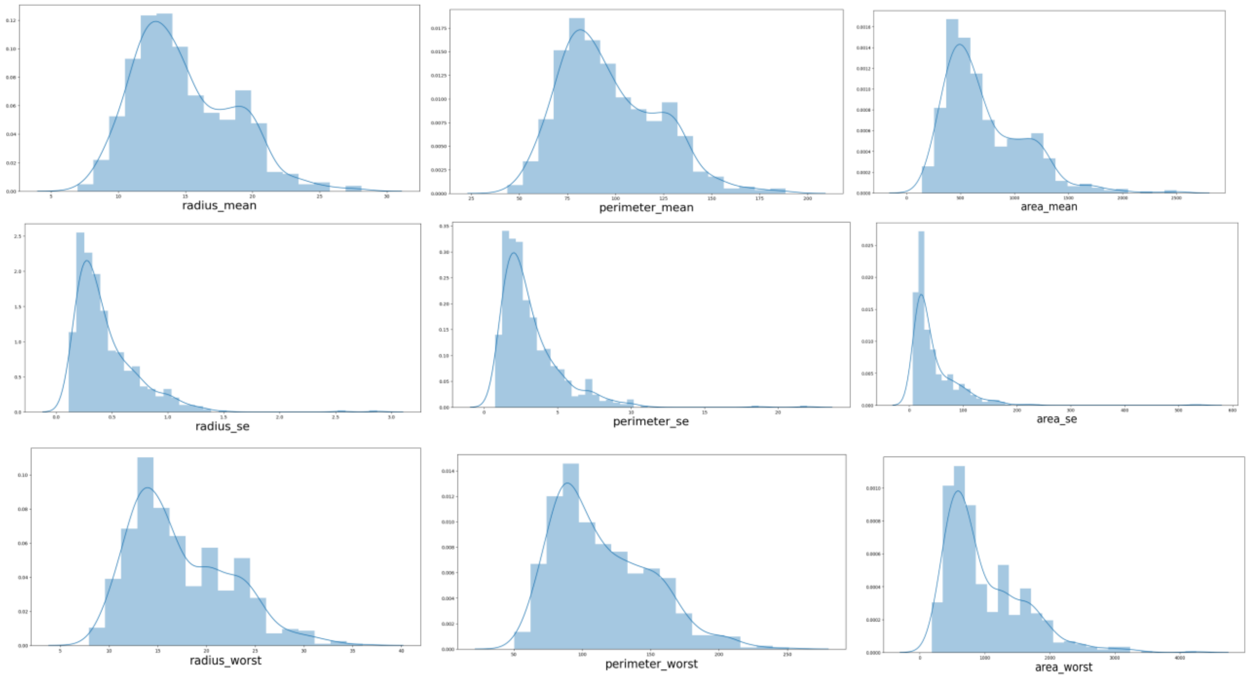

Figure 4. The first set of feature distribution diagrams

As shown in Figure 4. above, so radius_mean is probably a good predictor, almost linear, there are one or more points that appear to be outliers. Therefore, radius_mean is selected as a necessary feature.

(2) texture_mean and texture_worst;

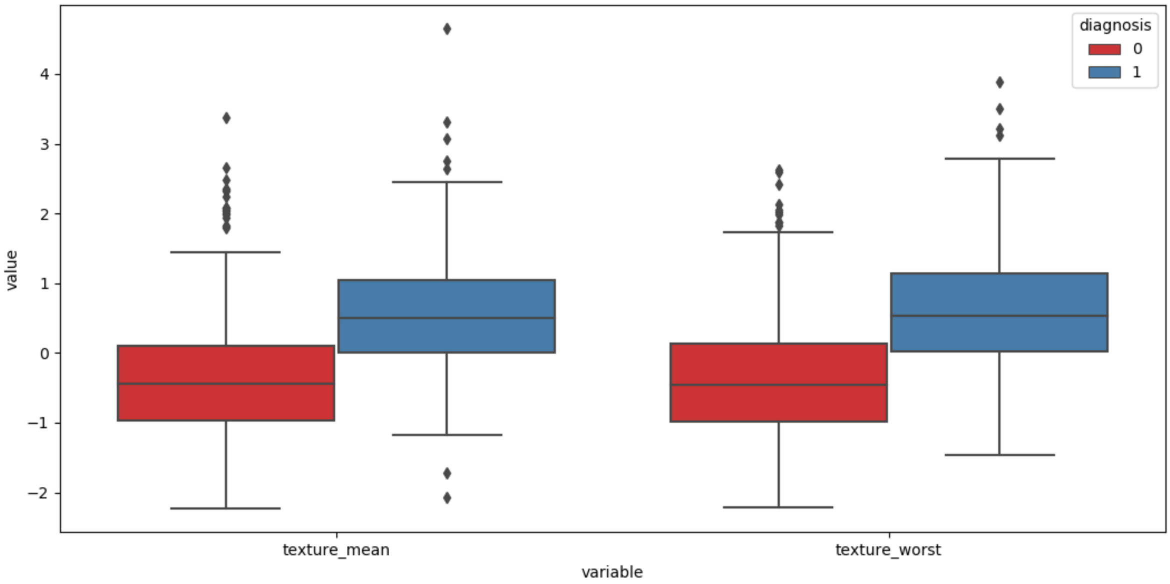

Figure 5. The second set of feature box plots

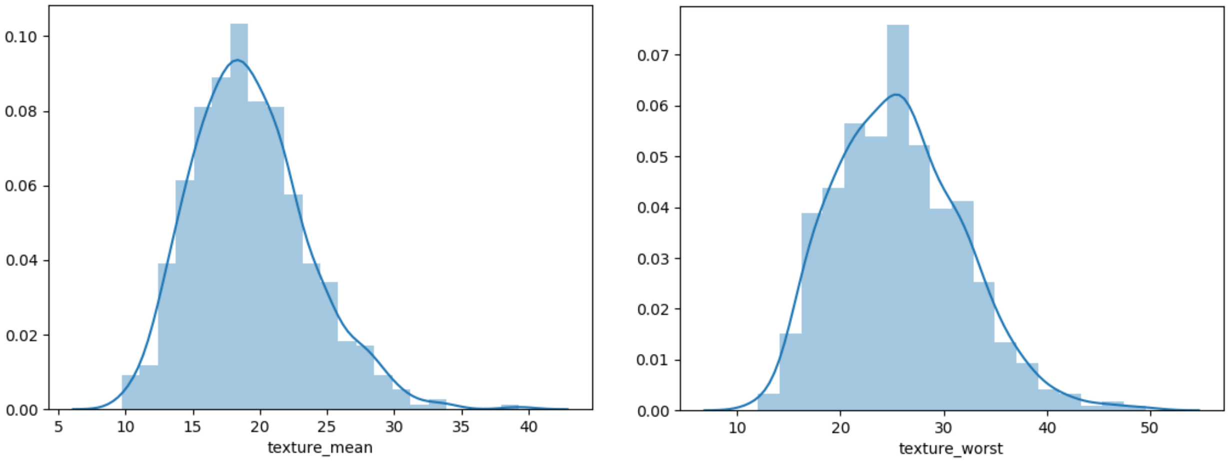

Figure 6. The second set of feature distribution diagrams

As shown in Figures 5. and 6. above, looking at outliers and variable overlap, texture_worst appears to be a better variable for the model. Therefore, texture_worst is selected as a necessary feature.

(3) smoothness_mean and smoothness_worst;

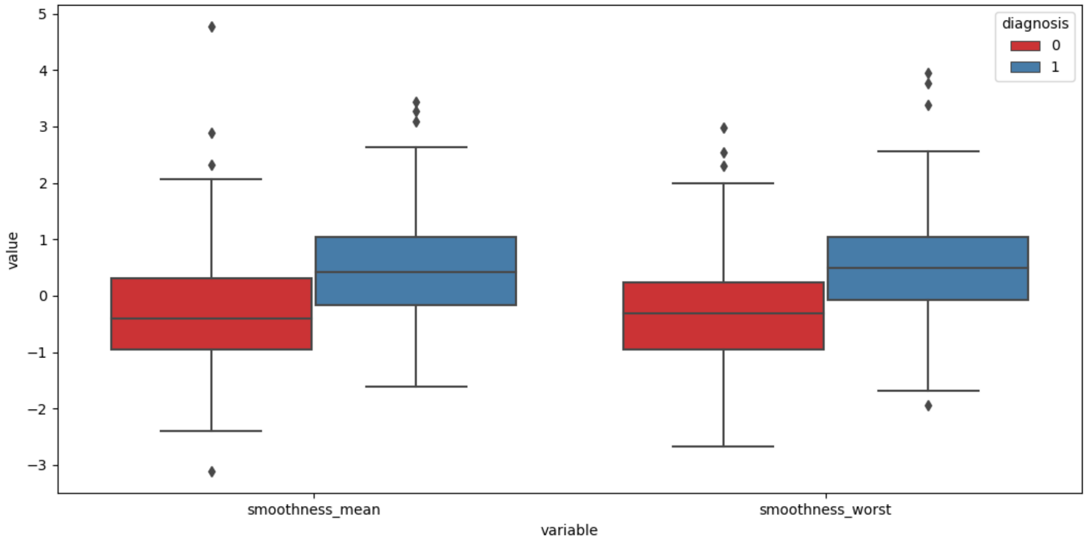

Figure 7. The third set of feature box plots

As shown in Figures 7. and 8. above, the overlap between the two variables cannot be used as a good predictor variable. However, considering the separation and outlier distribution, smoothness_worst is the better choice. Therefore, smoothness_worst is selected as a necessary feature.

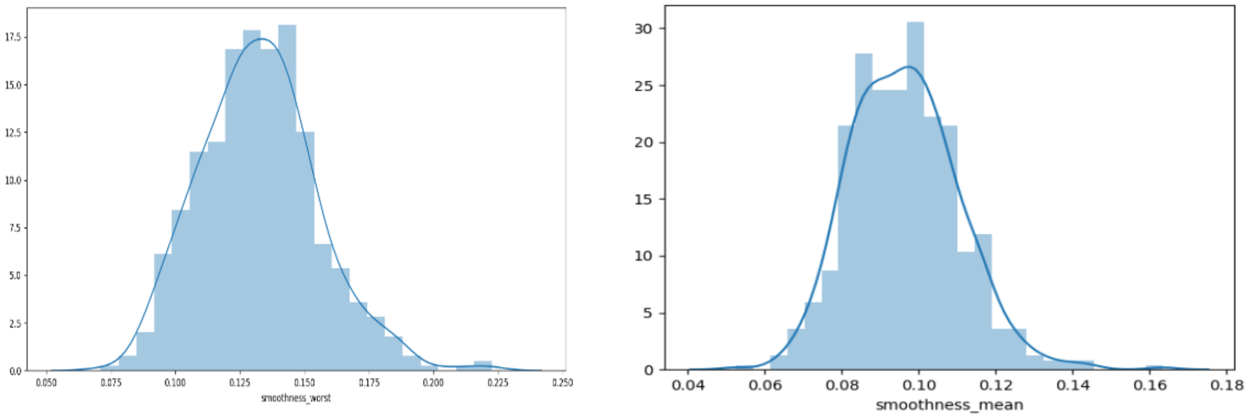

Figure 8. The third set of feature distribution diagrams

The remaining features do not have high correlation, so this article will use a better method to reduce the dimensionality of the data to achieve better classification results. Instead of completely giving up on a few of them, after the above analysis, this article will use PCA[4] to reduce the dimensionality of the remaining ten sets of unprocessed data.

2.2. Coping with high-dimensional data

2.2.1. Principal component analysis for feature extraction

This article will use principal component analysis (hereinafter referred to as PCA) to reduce the data dimension and improve the accuracy of classification, so as to achieve better results. The dimensionality reduction method of principal component analysis is used for data compression to eliminate redundancy and data noise elimination[5].

As shown in line chart 9, to explain the data contribution rate. All in all, almost all changes can be cumulatively attributed to the first three components, and the data contribution ratio reaches more than 0.98. In the end, this article uses the first three c omponents for explanation.

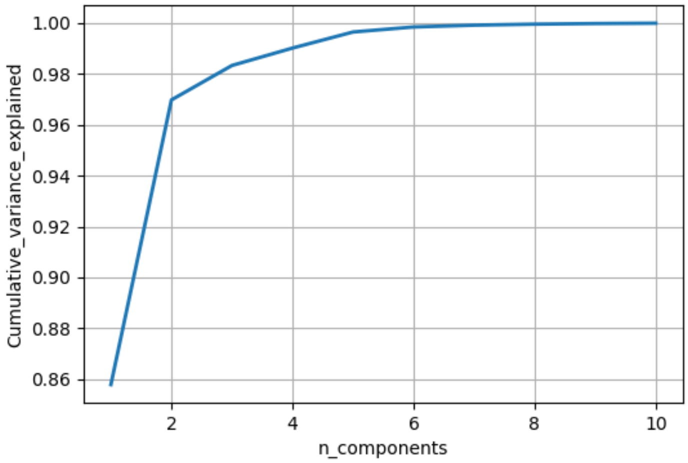

Figure 9. PCA dimensionality reduction contribution

3. Establishment of BP neural network

3.1. Research Process

This research will focus on obtaining the optimal classification. Principal component analysis was used to reduce the dimensionality of 10 feature data. Then put the necessary three feature data and the dimensionally reduced data obtained from the above summary into the constructed BP neural network for training and parameter adjustment.

3.2. BP neural network learning process

It is through such continuous forward transmission and reverse transmission operations that the BP network finally makes the actual output value of the network at the output layer tend to be the same as the expected trend.[6].

3.3. Construction of BP neural network

3.3.1. Selection of the number of hidden layer nodes

This article uses a hidden layer, with the number of nodes changing from 1 to 30, and analyzes how the accuracy changes with it[7].As shown in Figure 10 below, the accuracy rate of the test set is relatively stable at more than 98% between hidden layers 5 to 15, while the accuracy rate of the hidden layer is relatively unstable after 15 layers. Therefore, they are all within the hidden layer selection range. This article considers that the input layer has 6 layers and the output layer has 2 layers, so the number of hidden layer nodes is set to 10[8].

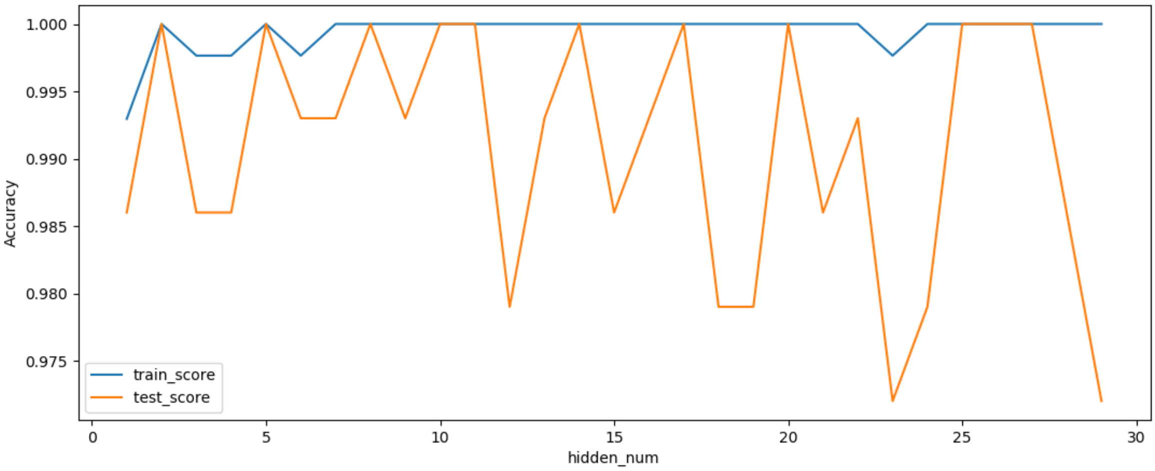

Figure 10. Changes in the number of hidden layers

3.3.2. loss function

This article uses cross entropy as the loss function. Its biggest advantage is that it can prevent the gradient from disappearing[9]. As shown in formula (1), given two probability distributions p and q, the cross entropy of p represented by q is:

$$H(p,q) = -\sum_{x} p(x)\log_2 q(x) \tag{1}$$

This article changes the number of iterations for a complete training of the model using all the data of the training set. The comparison of the loss value and accuracy is shown in Figure 11 below. As the number of iterations increases, the accuracy almost always remains at 100%, and the loss value gradually decreases, becoming unchanged when the number of iterations reaches about 3000. Therefore, considering that as the number of iterations gradually increases, the risk of overfitting becomes higher[10]. This article chooses the number of iterations to be 3000 iterations to train the model. As shown in Figure 11 below:

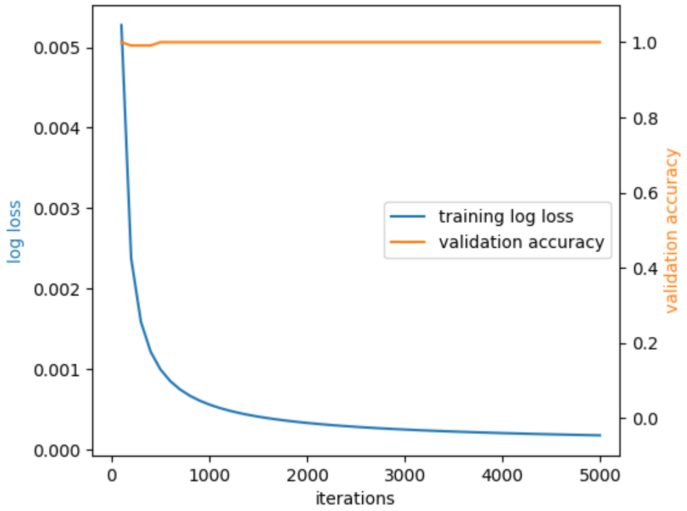

Figure 11. Changes in the number of iterations of loss and accuracy

3.4. Modeling results

As shown in Figure 12 below, after 3000 optimization iterations, the accuracy rate reached 100%. As shown in Table.1. below, the F1 score of each category is 1 point. The classification model has reached a relatively ideal expectation after a series of parameter adjustments. At the same time, this result is inseparable from the work in the data preprocessing stage.

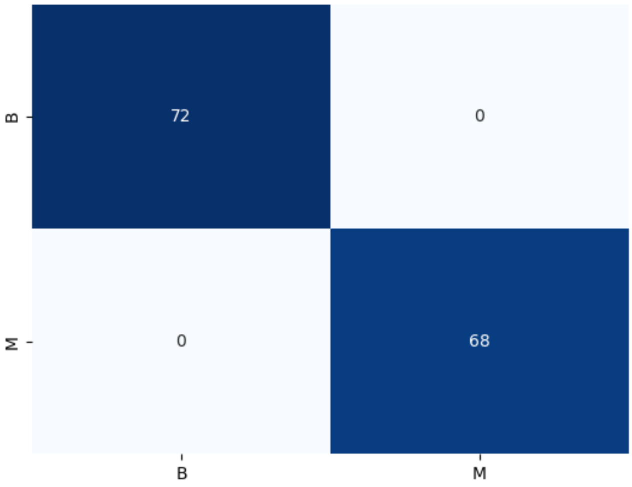

Figure 12. Confusion matrix result graph

Table 1. BP neural network modeling results

	precision	recall	f1-score	support
M	1.0	1.0	1.0	68
B	1.0	1.0	1.0	72
accuracy	1.0	1.0	1.0	140

4. Conclusion

The main contents of this paper are basically based on BP neural network to study the classification method of breast cancer feature data. This study mainly adopts 30 features of 569 cases. Eliminate unnecessary features through comprehensive sampling, correlation analysis, feature selection, principal component analysis, and other methods. After in-depth exploration of the data, six classification features were obtained and put into the constructed BP neural network to train the parameter adjustment training model. Finally, the classification accuracy of the model for breast cancer feature data reached 100%. Due to the small size of the text dataset, we can consider obtaining data from other dimensions in the future to classify more accurately, while increasing the amount of data to make the model more complete. Various optimization experiments were conducted on the BP neural network to demonstrate better performance on the basis of the original model.

References

[1] Shou Y, Meng T, Ai W, et al. Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism [J]. Computational Intelligence and Neuroscience, 2022, 2022.

[2] Shou Y, Meng T, Ai W, et al. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis [J]. Neurocomputing, 2022, 501: 629 - 639.

[3] Chen J, Liu Z, Yin Z, et al. Predict the effect of meteorological factors on haze using BP neural network [J]. Urban Climate, 2023, 51: 101630.

[4] Chen L, Jagota V, Kumar A. RETRACTED ARTICLE: Research on optimization of scientific research performance management based on BP neural network [J]. International Journal of System Assurance Engineering and Management, 2023, 14 (1): 489 - 489.

[5] Yang A, Zhuansun Y, Liu C, et al. Design of intrusion detection system for internet of things based on improved BP neural network[J]. Ieee Access, 2019, 7: 106043 - 106052.

[6] Shou Y, Cao X, Meng D, et al. A Low-rank Matching Attention based Cross-modal Feature Fusion Method for Conversational Emotion Recognition [J]. arXiv preprint arXiv: 2306. 17799, 2023.

[7] Wu Y, Gao R, Yang J. Prediction of coal and gas outburst: A method based on the BP neural network optimized by GASA [J]. Process Safety and Environmental Protection, 2020, 133: 64 - 72.

[8] Shou Y, Cao X, Meng D. Masked Contrastive Graph Representation Learning for Age Estimation [J]. arXiv preprint arXiv: 2306.17798, 2023.

[9] Wen L, Yuan X. Forecasting CO2 emissions in Chinas commercial department, through BP neural network based on random forest and PSO [J]. Science of The Total Environment, 2020, 718: 137194.

[10] Xu L, Wang H, Gulliver T A. Outage probability performance analysis and prediction for mobile IoV networks based on ICS-BP neural network [J]. IEEE Internet of Things Journal, 2020, 8 (5): 3524 - 3533.