The 2nd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/4/20220628

Research on Breast Cancer Classification and Prediction Model Based on Principal Component Analysis and Machine Learning

Qiushi Wang

School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

b20080613@njupt.edu.cn

Abstract. In recent years, breast cancer has become one of the biggest threats to women's health, accounting for the majority of cancer deaths among women. Because early treatment of breast cancer has a great effect on the recovery of breast cancer, the diagnosis of breast cancer is particularly important. Machine learning, as the most popular method, is also used for model construction in this field. This study is based on data from breast tumors, which contain 10 morphological features of breast tumor nucleus. In this study, homogenization, standardization and feature selection were used to process the data and KNN algorithm was used to construct the classification prediction model, with principal component analysis (PCA) used to optimize the model. Finally, the original 30 variables were reduced to 3 variables and the model parameters were adjusted in order to achieve the best model with the accuracy of 98.54%. The final model achieves the highest operating efficiency and accuracy. In this study, through the visualization of PCA and the comparison of different models, the classification effect of the final model can be the best. This model can be applied to the clinical diagnosis of breast cancer patients, which is helpful to the early treatment efficiency and greatly reduce the mortality of breast cancer patients.

Keywords: breast cancer, classification prediction model, machine learning, KNN, PCA.

1. Introduction

Over the past 30 years, cancer incidence and mortality rates have accelerated globally. As the biggest cause of death among female cancer patients, the incidence of breast cancer is increasing year by year. Breast tumor not only destroys the body of the patient, but also seriously affects female patients' psychology [1]. Nowadays, the incidence of breast cancer of female people is growing all around the world, especially in Asia, Africa, South America and other regions where the incidence of early breast cancer is low. As a result, the situation of global breast cancer is very alarming. Although the cure rate of early breast cancer is high, the death rate and recurrence rate of patients will greatly increase with the metastasis of cancer cells. Therefore, the diagnosis and treatment of early breast cancer is extremely important. Pathological diagnosis and imaging diagnosis are the main parts of the traditional diagnosis. The imaging diagnosis mainly consists of mammography, PET, MRI, CT and ultrasound. Breast ultrasound has been used routinely to detect breast cancer because of its good discriminative power. But

DOI: 10.54254/2753-8818/4/20220628

in traditional breast cancer diagnosis, doctors rely on visual information to analyze tissue images to determine how malignant a breast cancer tumor is. This process has the problems of long diagnosis time and easy misdiagnosis.

Therefore, improving the diagnostic efficiency and accuracy has become an important research content of breast cancer treatment. As a very popular data processing and mining method in recent years, machine learning has gradually played an excellent role in disease diagnosis. According to the development of the disease, clinical data were collected, including the patient's age, the size, shape and other data of tumor cell, changes in physical condition of the patient and so on. Therefore, many researchers have used machine learning to construct classification model for breast cancer diagnosis. Hamilton et al. used the method of rule derivation to discriminate breast cancer sample data, and the prediction accuracy reached 96% [2]. Setiono proposed an algorithm to extract classification rules from trained neural networks and applied it to the prediction model of breast cancer [3]. Polat proposed a classification algorithm using least square support vector machine (LS-SVM) to construct the model, which obtained a high accuracy result, but the results would be different when the model is evaluated in different ways [4]. Alickovic et al. used normalized multilayer perceptron neural network to constructed a classification model, dividing the training set and test set, and got 99.27% accuracy [5]. However, in the existing research on breast cancer diagnosis models, many pursue to improve the accuracy but ignore the training efficiency of the model. Too many variables used in the training process will reduce the efficiency of the model, thus affecting the diagnosis of breast cancer.

This study used data processing and feature selection based on breast cancer data, combined with KNN algorithm and principal component analysis, to establish a diagnostic classification and prediction model for breast cancer. In this study, many parameters such as the number of principal components of the PCA and the K value of KNN algorithm were adjusted to reduce the variables with low contribution in the breast cancer classification model, and the dimensionality of other variables was reduced to improve the operation efficiency and accuracy of the model. This study also evaluated the model by comparing the data accuracy and variables of different treatments. This model helps solve the problem of low efficiency of traditional models for breast cancer clinical data training. It has high reliability and practicability, which can be used as a software for clinical equipment to predict patients' condition according to clinical data. This model can also improve the efficiency of early diagnosis and treatment of breast cancer.

2. Dataset

2.1 Data source

The data set used was the clinical data contributed by Dr. W. H. Wolberg of the University of Wisconsin Clinical Science Center, USA [6], which contains the attributes calculated from the digital images of the breast masses. This dataset contains a total of 569 instance samples and 32 attributes, including 2 information attributes and 30 feature attributes. In the original sample data, the classification label used 'M' to represent malignant and 'B' to represent benign. In this experiment, the classification was redefined, with '0' representing benign and '1' representing malignant, accounting for 63% and 37% of the total samples, respectively.

2.2 Variable declaration

The mean, standard deviation and maximum of 10 feature values are calculated for each sample image in the data set, a total of 30 feature values. The mean is the average of the variables of the nucleus in the sample image. The standard deviation reflects the fluctuation of variables of different nuclei. The maximum value is not the maximum value of the whole sample, but the average value of the top three variables, which can weaken the influence of extreme values on variables. As shown in Table 1, the means of the 10 variables are interpreted as an example.

DOI: 10.54254/2753-8818/4/20220628

Table 1. Interpretation of the mean variables

Variable	Interpretation
radius_mean	the distance of the nucleus from the center to the peripheral points(mean)
texture_mean	standard deviation of gray value(mean)
perimeter_mean	circumference of nucleus(mean)
area_mean	area of nucleus(mean)
smoothness_mean	local variation in radius length(mean)
compactness_mean	compactness(mean)
concavity_mean	the severity of the contour concave(mean)
concave points_mean	the number of contour recesses(mean)
symmetry_mean	symmetry(mean)
fractal_dimension_mean	fractal_dimension(mean)

The above variables describe the characteristics of a nucleus from different perspectives. By establishing the model, the tumor status of patients can be diagnosed and the efficiency of diagnosis can be improved. It is possible to improve survival and cure rates.

3. Data preprocessing

3.1 Sample equalization

Before training the model, it is necessary to preprocess the data. Through TSNE visualization algorithm, the original feature attributes of 30 dimensions are reduced to the two-dimensional plane. Observing the visual images as shown in figure 1, it can be seen that the data has good separability, which is suitable for classification training and testing using the model. However, there is a certain imbalance in the data.

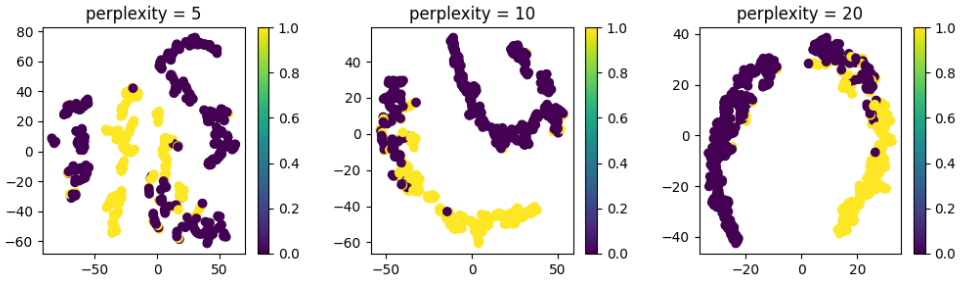

Figure 1. Tsne plots of the original data

In the classification learning algorithm, category imbalance represents the large difference in the proportion of different categories' samples, which will cause interference to the model training. In this data, the number of benign breast masses is significantly less than the number of malignant breast masses, which will affect the division of the training set and test set, thus leading to low accuracy of final model construction. Therefore, SMOTE oversampling strategy was used in this paper for sample balance [7].

SMOTE oversampling strategy is that for each minority sample A, in order to increase the new minority sample, the algorithm will randomly select a sample B from the nearest one of A, and the new minority sample will be created randomly on the line between A and B. As an improved algorithm, it can avoid overfitting and obtain better results. With the implementation of Python code, the original

DOI: 10.54254/2753-8818/4/20220628

dataset was equalized. In the end, the adjusted sample contained 311 benign and 308 malignant samples. TSNE was again used to visualize the dataset.

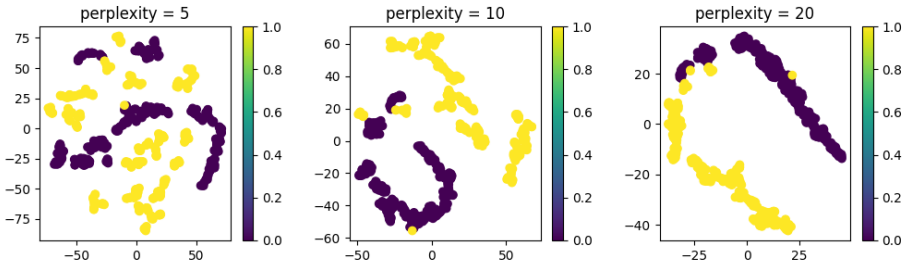

Figure 2. Tsne plots of the adjusted data

Figure 2 shows that the samples of the two categories are in an equilibrium state, and the separability of the samples has also been improved to some extent. The adjusted data are more suitable for model training and testing, so as to get more accurate results.

3.2 Standardization

The sample data consists of different orders of magnitude. The results obtained through the analysis of the original data are of no practical significance. Generally speaking, indicators with higher values will play a higher role in the comprehensive analysis than indicators with lower values. So the results are more likely to be influenced by indicators with higher values [8]. Thus, standardization of the data is need. Because of the large number of extreme values in the original data, normalization is not appropriate. Therefore, the data was regularized so that the mean of each sample after regularization is 0 and the variance is 1.

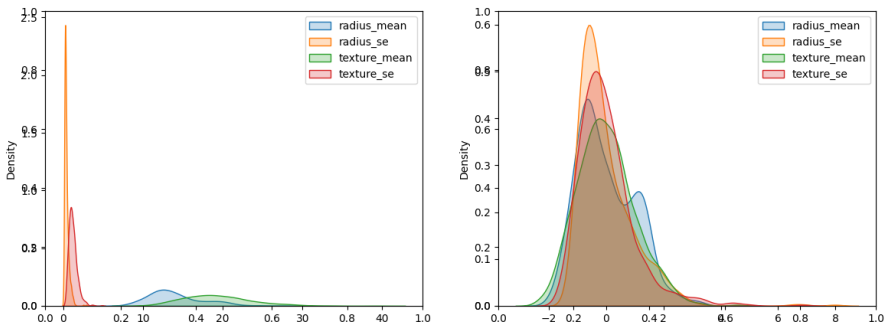

Figure 3. Kernel density maps before and after standardization

Four variables were selected to generate the kernel density map as shown in the left of figure 3. The distribution of different variables is quite different, for example, texture_se is very concentrated and has a small value, while radius_mean is widely distributed and has a large value. After the feature data is standardized, the kernel density map is shown in the right of figure 3, which eliminates the dimensional difference between features, so that the function of features in the model is not affected by their dimensions, and ensures the reliability of classification results.

3.3 Feature selection

For model training and testing, some variables in the original data set will play a very small role. Feature selection was conducted by studying the correlation between feature and target. The low correlation between the feature and the target indicates that the change of the feature will not bring much influence on the change of the target [9], so the few variables can be eliminated with the lowest correlation. SelectKBest in Python was used to sort the correlations and select the last 10 variables.

DOI: 10.54254/2753-8818/4/20220628

There are a lot of variables of the standard error in the last 10 variables. It can be inferred that for breast cancer tumors, there is no significant difference in the dispersion of size, shape, and other indicators between malignant and benign tumors while the differenceis more pronounced in the mean and worst values.

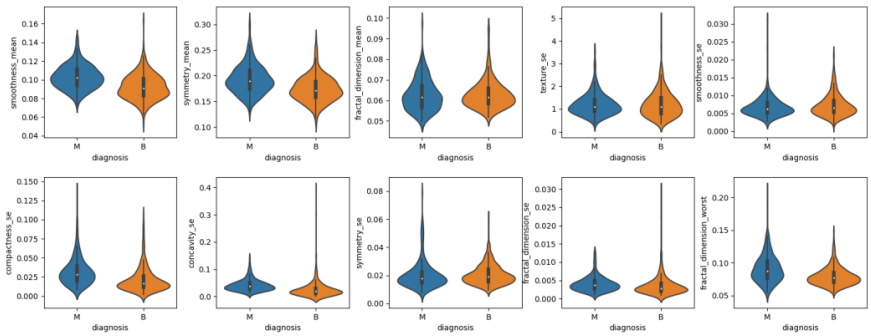

Figure 4. Violin plots of the last 10 variables, where blue is malignant and orange is benign.

Violin plots of the 10 variables with the lowest correlation were created. As can be seen from Figure 4, the malignant and benign violin plots under each variable were very similar, which means that the distribution of samples of different categories was almost the same. Therefore, in these variables, it is difficult to distinguish malignant and benign samples, which cannot help the model construction of breast cancer diagnosis. The rationality of relevant screening was further verified. Therefore, The 10 variables were removed with the lowest correlation and a new dataset with 20 variables was generated.

4. Classification model construction

4.1 KNN's principle

KNN classification algorithm was chosen to use to train the model. The principle of classification prediction process of KNN is: when predicting a new value x, we only need to find the most frequent category in the set through the set composed of k points closest to it. Then the new value x can be considered as a member of this category [10].

Based on the above ideas, the following algorithm execution steps can be given:

1. Select k sample points with the nearest distance in the vicinity of x in the training set (Euclidean distance method is generally used for distance measurement), and create a set N_k(X) containing these k sample points. The formula of Euclidean distance is as follows:

$$d_{xy} = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{1}$$

Determine instance X belongs to category Y according to the principle of majority voting as shown below:

$$y = argmax \sum_{x_i \in N_{k(x)}} I(y_i, c_j), i = 1,2,\dots,N; j = 1,2,\dots,K, \tag{2}$$

where, I is the indicator function:

$$I(x,y) = \begin{cases} 1, if\ x = y \\ 0, if\ x \neq y \end{cases} \tag{3}$$

DOI: 10.54254/2753-8818/4/20220628

4.2 Selection of training parameters

The selection of k value in KNN algorithm is important in the prediction result. If the k value is relatively small, the approximate error of KNN will be relatively small, and the prediction result is easily affected by the nearest neighbor, which leads to overfitting. At the same time, the value of k also has a range limitation on the other side, because when the value of k is larger, the nearest neighbor error of the prediction result will be larger, which will also lead to a large deviation between the prediction result and the actual result [11]. In this case, the model is prone to underfitting.

Therefore, k value was selected by means of cross validation. Cross-validation's principle is to process data repeatedly, divide the given data, combine the segmented data into training set and test set, and conduct training test and model selection repeatedly on this basis. The cross_val_score implementation in Python was used to select the k value.

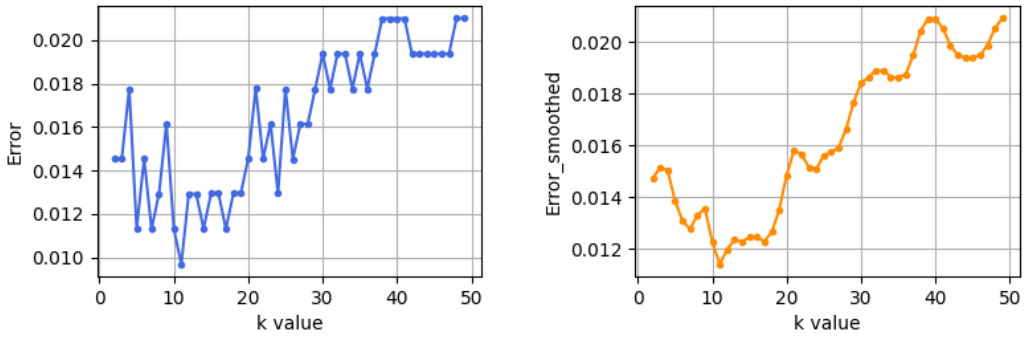

Figure 5. Errors of different value of k in KNN

The error and smoothed error of different K values are shown in the figure 5. It can be seen that when k=11, the loss value is the smallest, so k=11 was chosen for training. After the model training, the accuracy of the results can be obtained as 98.38%.

4.3 PCA's principle

In practical applications, the time cost of the model is highly required if the results are expected to be obtained quickly. Although the feature selection was used to delete 10 variables that had little influence on the model construction early, 20 variables were still needed for the experiment. Therefore, PCA method was chosen to reduce the dimension of variables to optimize the model.

PCA is a dimensionality reduction method. When there are a large number of indicators and multiple indicators have strong correlation, PCA can exclude overlapping information, reduce the dimension of the original large number of indicators, and transform them into a small number of comprehensive indicators with low correlation [12]. PCA is suitable to be used in a large number of index data sets, which can effectively reduce the number of variables, so as to simplify the process of machine learning and improve the efficiency of the model.

4.4 Dimension reduction based on PCA

By using PCA, the explanation variance ratio (EVR) of different principal components was obtained, as shown in the figure 6 below.

DOI: 10.54254/2753-8818/4/20220628

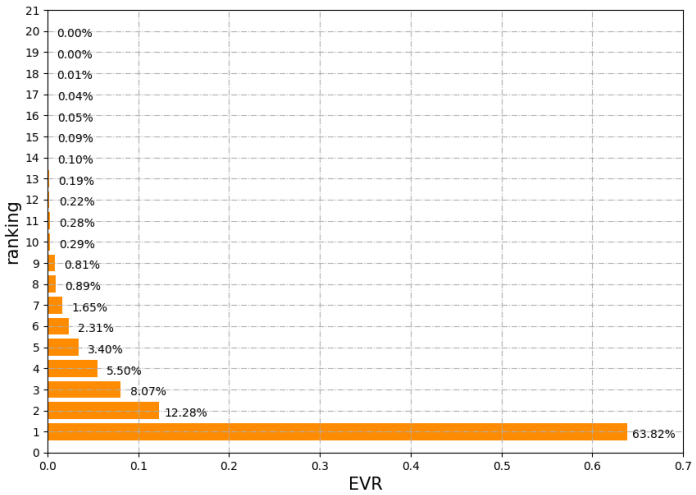

Figure 6. EVRs of different components

As can be seen from the figure 6, the EVR of the first principal component is 63.82%, that is, the explanation rate of the original variable reaches 63.82%, and the contribution rate is the largest. When the principal component increases, the corresponding EVR decreases sharply. When the principal component reached the eighth place, its EVR was only 0.89%, indicating that its explanatory rate to the original variable was very small and its contribution was very low. As the principal component continues to increase, its corresponding EVR is almost close to 0, representing that the contribution degree is close to 0.

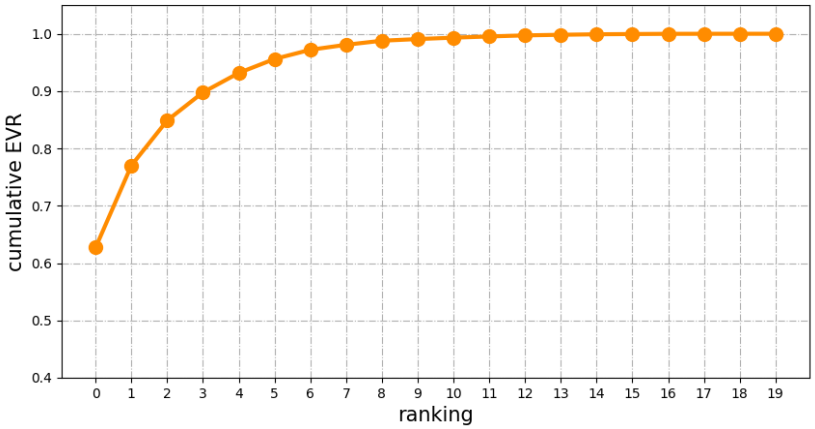

Figure 7. Cumulative EVR plot

A cumulative EVR plot was made as shown in figure 7. A threshold of 95% was set and the number of principal components obtained is 6. That is, when the dimension of the original variable is reduced to 6 variables, the cumulative explanation rate of the original variable has reached 95%, indicating a good effect of dimension reduction.

4.5 Model optimization based on PCA

When PCA was used to reduce the dimension to 6 variables, it is believed that 95% explanation rate could replace the original model. Therefore, KNN was re-used, substituted the variables after dimension reduction, and the best k value of 12 by cross-validation method was obtained. Then, when k=12 is inserted into the model for training, the training result accuracy is 98.38%.

DOI: 10.54254/2753-8818/4/20220628

By constantly adjusting the number of components, the number of principal components is reduced as much as possible, and the accuracy of KNN classification is not reduced. Finally, the least number of principal components is 3, the best K value is 12, and the accuracy of the model is improved to 98.54%.

5. Model evaluation

In this experiment, sample balance and data standardization are carried out on the original data to make the data more suitable for machine learning model training. In addition, the feature selection method, PCA dimensionality reduction, and the adjustment of K value in KNN algorithm are used to optimize the model. Through comparative experiments, it can be shown in table 2 that the prediction accuracy of KNN model under different data processing was obtained.

Table 2. Comparison of the accuracy of different methods

	Original data	After feature selection	After PCA	After PCA and adjusted K
Accuracy	97.84%	98.39%	98.39%	98.54%
Variables	30	20	6	3

As can be seen from the table 2, after feature selection, the original data set deleted the 10 variables with less impact, the accuracy was improved, and the feature variables were reduced, and the model efficiency was improved. After PCA dimensionality reduction, the number of feature variables is further reduced to 6, which can have a great effect on reducing the operation time. By adjusting the k value of KNN algorithm and further adjusting the number of principal components, finally, the accuracy is improved to the highest, reaching 98.54%, and the number of variables is reduced to 3. The efficiency and accuracy of the model are further improved.

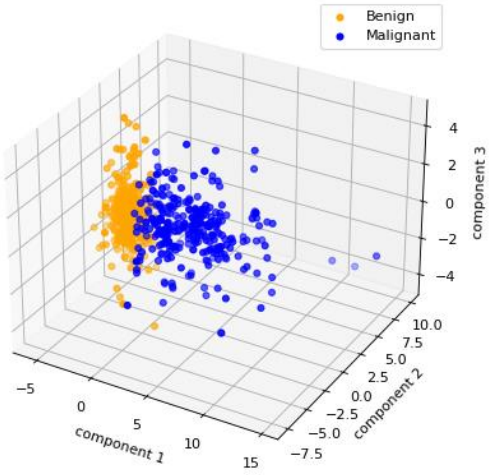

Figure 8. 3D visualization of PCA

Finally, the model's accuracy reaches the highest when the dimension of the original feature variables is reduced to three feature variables. Therefore, the classification results of PCA were visualized in 3D, as shown in the figure 8. It can be seen that the malignant and benign samples of breast masses in 3D space are clearly demarcated and can be separated by a definite two-dimensional plane. This also verifies the high accuracy of the model.

6. Discussion

Through the whole research, the maximum values of each feature in the dataset played a significant role in the classification of breast tumors. The dimensionality of these features can be reduced into three by

PCA so that it can achieve high precision classification of breast cancer. At the same time, the model still has many limitations. Firstly, the small dataset may result in a low malleability of the model. Secondly, although KNN model has many advantages, it still has the problem of slow running speed caused by complex operation, which may affect the work efficiency when applied to medical equipment. Thirdly, although the model improvement based on PCA has achieved good results, it may not be the optimal dimensionality reduction method due to the principle. PCA may lead to low model robustness in larger datasets.

Therefore, in future research, this study will look for larger data sets, conduct more statistical analysis on the eigenvalues in the data sets, and select different algorithms based on the characteristics of the data, in order to build a breast cancer diagnosis prediction model with better scalability and higher robustness. In future work, this study will also build a model for the prognosis of breast cancer patients. A prediction model will be built based on the physiological data of different breast cancer patients after diagnosis, and strive to build a personalized treatment plan for patients diagnosed with breast cancer.

7. Conclusion

This research uses homogenization, standardization and feature selection of breast cancer data, combined with KNN algorithm to train breast cancer samples, and it constructs a high-precision breast cancer diagnostic classification model. Meanwhile, principal component analysis(PCA) is added to optimize the model. Finally, the dimension of the 30 feature values in the original dataset is reduced to three attribute variables, and maintained the classification accuracy of the model above 98%.

In this study, the classification effect of PCA and KNN algorithm is good, but there is still room for improvement in accuracy. In future work, the effects of different dimensionality reduction algorithms and different classifiers will be compared in order to build a better classification and prediction model for breast cancer diagnosis. This study can be used as a software to carry on the medical instrument. By analyzing the section data of the breast mass of the patient, the tumor condition of the patient can be predicted. It can achieve the purpose of early diagnosis, help patients to take treatment in advance, and reduce the mortality of breast cancer.

References

[1] L. Gao, J. Liu, X. Zhou, Y. Su and P. Wang, Supporting her as the situation changes: A qualitative study of spousal support strategies for patients with breast cancer in China. Eur J Cancer Care **29**, e13176(2020).

[2] H. J. Hamilton, N. Shan and N. Cercone, RIAC: A Rule Induction Algorithm Based on Approximate Classification. Tech. Rep. CS **96**, 06 (1996).

[3] R. Setiono, Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in Medicine **18(3)**, 205-219(2000).

[4] K. Polat and S. Gunes, Breast cancer diagnosis using least square support vector machine. Diqital Signal Prolessing **17(4)**, 694-701(2006).

[5] Alickovic E and Subasi A, Normalized neural networks for breast cancer classification. International Conference on Medical and Biological Engineering **73**, 519-524(2019).

[6] W. H. Wolberg and O. L. Managasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceeding of the national a cadeny of sciences **87(23)**, 9193-9196(1990).

[7] D. Li, J. Liu and J. Liu, NNI-SMOTE-XGBoost: A Novel Small Sample Analysis Method for Properties Prediction of Polymer Materials. Macromol. Theory Simul **30**, 2100010(2021).

[8] M. S. Mohamad, S. Omatu, S. Deris and M. Yoshioka, Three-Stage Method for Selecting Informative Genes for Cancer Classification. IEEJ Trans Elec Electron Eng **4**, 725-730(2009).

[9] B. Li, Q. Wang and J. Hu, Feature subset selection: a correlation-based SVM filter approach. IEEJ Trans Elec Electron Eng **6**, 173-179(2011).

[10] P. Sadhukhan and S. Palit, Multi-label learning on principles of reverse k-nearest neighbourhood. Expert Systems **38**, e12615(2021).

DOI: 10.54254/2753-8818/4/20220628

[11] M. Rachdi, A. Laksaci, Z. Kaid, A. Benchiha and F. A. Awadhi, k-Nearest neighbors local linear regression for functional and missing data at random. Statistica Neerlandica **75**, 42– 65(2021).
[12] M. Stefanucci, L. M. Sangalli and P. Brutti, PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. Statistica Neerlandica **72**, 246– 264(2018).