Scientific Journal of
Women's Health & Care

Research Article

A Statistical Study on Anti-Breast Cancer Drug Screening -

Xia Jiang[1] and Bin Zhao[2]*

[1]*Hospital, Hubei University of Technology, Wuhan, Hubei, China*

[2]*School of Science, Hubei University of Technology, Wuhan, Hubei, China*

***Address for Correspondence:** Bin Zhao, School of Science, Hubei University of Technology, Wuhan, Hubei, China, Tel: +86-130-285-175-72; E-mail: zhaobin835@nwsuaf.edu.cn

Submitted: 19 November 2021; **Approved:** 06 January 2022; **Published:** 07 January 2022

Citation this article: Jiang X, Zhao B. *A Statistical Study on Anti-Breast Cancer Drug Screening*. Sci J Womens Health Care. 2022 Jan 07;6(1): 001-006.

ABSTRACT

Breast cancer is one of the most lethal cancers, Estrogen Receptor α Subtype (ERα) is an important target. The compounds that able to fight ERα active may be candidates for treatment of breast cancer. The drug discovery process is a very large and complex process that often requires one selected from a large number of compounds. This paper considers the independence, coupling, and relevance of bioactivity descriptors, selects the 15 most potentially valuable bioactivity descriptors from 729 bioactivity descriptors. An optimized back propagation neural network is used for ERα, The pharmacokinetics and safety of 15 selected bioactivity descriptors were verified by gradient lifting algorithm. The results showed that these 15 biological activity descriptors could not only fit well with the nonlinear relationship of ERα activity can also accurately predict its pharmacokinetic characteristics and safety, with an average accuracy of 89.92~94.80%. Therefore, these biological activity descriptors have great medical research value.

Keywords: Breast cancer; Softmax function; Adam algorithm; Biological activity

INTRODUCTION

Breast cancer is one of the most common malignant tumors in women, and a malignant tumor occurring in ductal epithelium of the breast. It is the second leading cause of death by cancer among women in the United States. The total cost of illness for breast cancer has been estimated to be $3.8 billion, of which $1.8 billion represents medical care costs [1]. The difficulty in drug development is one of the huge treatment costs. How to reasonably choose the appropriate cancer drugs from tens of thousands of compounds, is one of the effective ways to reduce the cost of breast cancer treatment. According to Anthony Howell study, Estrogen may stimulate the growth and progression of infiltrating tumours [2]. Duan's study found that the estrogen is involved in the growth and differentiation of mammary epithelial cells in hormone dependent tumors. It plays an important role in the occurrence and development of breast cancer [3]. Zhang's study showed that estrogen mainly acts through the estrogen receptor expressed in the nucleus, that is, by binding with Estrogen Receptor (ER) to form a complex [4]. Through survival analysis, Chen found that the ERα positive rates of invasive carcinoma, invasive lobular carcinoma and ductal carcinoma in situ were 59.5%, 78.9% and 63.6%, respectively [5]. Accordingly, antihormone therapy is commonly used in breast cancer patients with ERα expression, which controls estrogen levels through regulating estrogen receptor activity. ERα mediates the E2 up regulation of PI3K/Akt signaling pathway and promotes cell proliferation [6]. So, compounds that can antagonize ERα activity may be candidates for treatment of breast cancer. For example, tamoxifen and renoxifene are the ERα antagonists for clinical treatment of breast cancer [7]. In order to screen potential active compounds, a potential compound model is usually established to collect compounds and bioactive data by targeting the specific estrogen receptor subtype targets associated with breast cancer. The Quantitative Structure-Activity Relationship (QSAR) model of compounds was constructed with the biological activity descriptor as the independent variable and the biological activity of compounds as the dependent variable. The model was used to predict the new compound molecules with good biological activity or guide the structural optimization of existing active compounds. A compound that wants to become a candidate drug, besides having good biological activity (here refers to anti breast cancer activity), also needs to have good pharmacokinetics and safety in human body. Pharmacokinetics is the study of the movement of a drug through the compartments of the body and the transformations (activation and metabolism) that affect it [8]. It is called ADMET property, including absorption, distribution, metabolism, excretion and toxicity. When determining the biological activity of a compound, it is also necessary to consider its ADMET properties as a comprehensive consideration. In this paper, the coupling degree between bioactivity descriptor and ERα activity is verified by BP neural network. After determining that the screened bioactivity descriptors can indeed affect ERα activity to a great extent, the ADMET property of bioactivity descriptors is further verified.

OVERVIEW OF BP NEURAL NETWORK

Artificial neural network is widely used in pattern recognition, function approximation and so on. BP neural network is a multilayer feed forward network simulating human brain. It has good adaptability and training ability, belongs to nonlinear dynamic system, and including two processes: forward propagation of information and back propagation of error. BP neural network consists of three parts: input layer, hidden layer and output layer. The input layer receives the input information, and then transmits the information to the hidden layer. The hidden layer analyzes and processes the data. Finally outputs acceptable information through the output layer. This information is continuously corrected through the reverse propagation of error, which can make full use of the coupling between data. BP neural network shows excellent accuracy in many fields. Therefore, this paper selects neural network as the main prediction method. Whether it is regression network or prediction network, the setting of the hidden layer and the number of hidden nodes of the network is very important. Too few hidden layers and hidden nodes will lead to less data information that the neural network can process, resulting in low prediction accuracy, and too many hidden layers will lead to over fitting of the model. There is no general calculation formula for the setting of the optimal number of hidden nodes. It is more based on the empirical formula or changing the number of hidden nodes to continuously train the model to find the number of hidden nodes with the smallest error [9-11]. Basic structure diagram of BP neural network is shown in figure 1.

The activation function of BP neural network usually uses softmax function to give corresponding weight to each node and transfer information between nodes in the network. In addition, there is an offset weight in the propagation of each layer of network, which is an additional constant of softmax function. We will use softmax function as an activation function of the BP neural network. The softmax function as follow:

$$S_i = \frac{e^i}{\sum_j e^j}$$

Its operating principle is shown in figure 2.

The softmax function in the figure maps the input values between 0.1 with the sum of 1, understanding these values as probabilities of

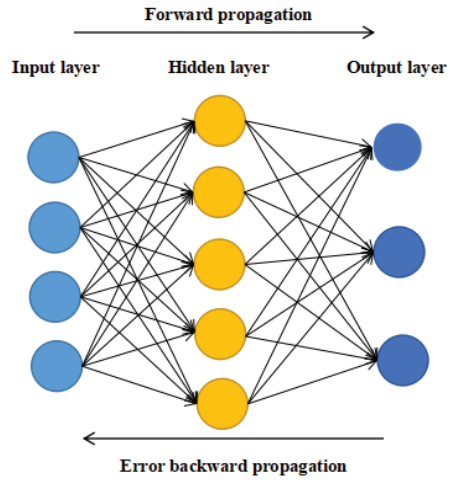


Figure 1: Basic structure diagram of BP neural network.

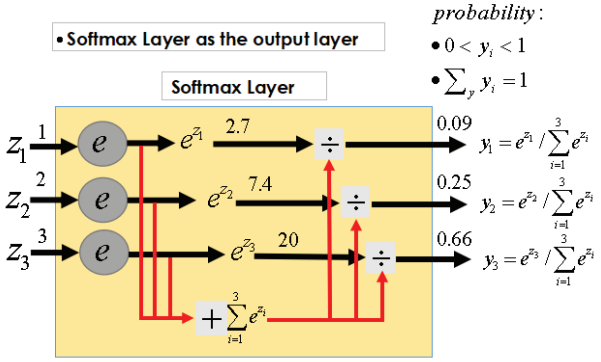


Figure 2: Principle of Softmax function.

probability, and the node with the highest probability can be selected as the prediction target. The BP neural network algorithm is as follows:

Enter layer to the hidden layer:

$$\alpha_h = \sum_{i=1}^{d} v_h * x_i$$

The softmax activation function passing through the hidden layer:

$$b_h = f(\alpha_h - \gamma_h)$$

Hideen layer to the output layer:

$$\beta_j = \sum_{h=1}^{q} w_j * b_h$$

Activation function passing through the output layer:

$$y_j^{k'} = f(\beta_j - \eta_j)$$

computing error:

$$E_k = 1/2 * \sum_{j=1}^{l} (y_j^{k'} - y_j^k)^2$$

Where x_i is the input, v_{ih} is the weight of the input layer to the hidden layer, γ_h is the hidden layer valve value, w_{hj} is the weights between hidden layer and output layer, η_j is the input layer valve value, $f(\bullet)$ is the softmax activation function.

In the model training, the gradient optimization algorithm (Adam algorithm) is used to optimize the model to obtain the best results [12].

Adam algorithm:

Initialize 1st, 2nd moment vector and time step:

$$m_0.v_0.t \leftarrow 0$$

do while:

$$t \leftarrow t+1$$

Computing the gradient:

$$g_t \leftarrow \nabla_\theta f(\theta_{t-1})$$

Update biased first moment estimate:

$$m_t \leftarrow \beta_1 * m_{t-1} + (1-\beta_1) * g_t$$

Update biased second moment estimate:

$$v_t \leftarrow \beta_2 * v_{t-1} + (1-\beta_2) * g_t^{\ 2}$$

Compute bias-corrected first moment estimate:

$$\hat{m}_t \leftarrow \frac{m_t}{(1-(\beta_1)^t)}$$

Compute bias-corrected second moment estimate:

$$\hat{v}_t \leftarrow \frac{v_t}{(1-(\beta_2)^t)}$$

Update parameters:

$$\theta_t \leftarrow \theta_{t-1} - \alpha * \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

Where α is the step length, $\beta_1; \beta_2$ is the momen estimation of exponential decay rate, and $f(\theta)$ is the random objective function of parameter θ. Adam algorithm will be used to optimize the parameters of BP neural network in order to accelerate convergence and improve accuracy. The model is:

Step 1: Initialize the network weight and bias, give each network connection weight a small random number, and each neuron with a bias will also be initialized to a random number.

Step 2: Forward propagation. Input a training sample, and then calculate the output of each neuron. The calculation method of each neuron is the same, which is obtained by the linear combination of its inputs.

Step 3: The gradient descent method is used to calculate the error and carry out back propagation. The weight gradient of each layer is equal to the input of the connection of the previous layer multiplied by the weight of the layer and the reverse output of the connection of the next layer.

Step 4: The weight gradient in the third step is used to adjust the network weight and neural network bias.

Step 5: Back propagation, Adam algorithm is used to accelerate the weight adjustment, initialize the moment vector and exponential weighted infinite norm to 0, update the parameters through vector operation, and iterate in t time from step size to 1. Sort errors and return.

Step 6: At the end of judgment, for each sample, judge if the error is less than the threshold set by us or has reached the number of iterations. We'll finish training, otherwise, return step 2.

DATA DESCRIPTION AND PREPROCESSING

In this paper, the bioactivity description data set is used to verify the ERα activity and ADMET properties respectively. The description dataset contains 729 biological activity descriptors of 1974 compounds. Because the data dimension is too large and contains a large number of repetitions and useless variables, this paper selects 15 most representative biological activity descriptors from the 729 biological activity descriptors of 1974 compounds. Firstly, low variance filtering is used to delete the biological activity descriptors with low information, then considering the correlation and independence between variables, Lasso regression is used to select these variables, and finally considering the coupling degree between variables and ERα activity. The final 15 most representative biological activity descriptors are obtained. The specific steps are as follows:

Step 1: Because the variance of variable can reflect the degree of dispersion, the variable with small variance contains little information, which cannot provide key and useful information for the construction of the model. Therefore, for 729 biological activity descriptors of 1974 compounds, the variance of 729 variables is calculated and arranged from large to small.

Step 2: After cleaning the biological activity descriptors with low information or no information, use the remaining molecular descriptors to further process the repeated information of the data, so as to make the data relatively independent. In this paper, Lasso [13] feature selection method is used to propose a variable from two variables with strong correlation to eliminate duplicate information. The essence of lasso feature selection method is to seek the sparse expression of the model and compress the coefficients of some features to 0, so as to achieve the purpose of feature selection. The parameter estimation of lasso feature selection method is as follows:

$$\hat{\gamma} = \arg\min_{\gamma}{}^{2} \mid y - \sum_{i=1}^{p} x_i \gamma_i \mid^{2} + \lambda \sum_{i=1}^{p} \mid \gamma_i \mid$$

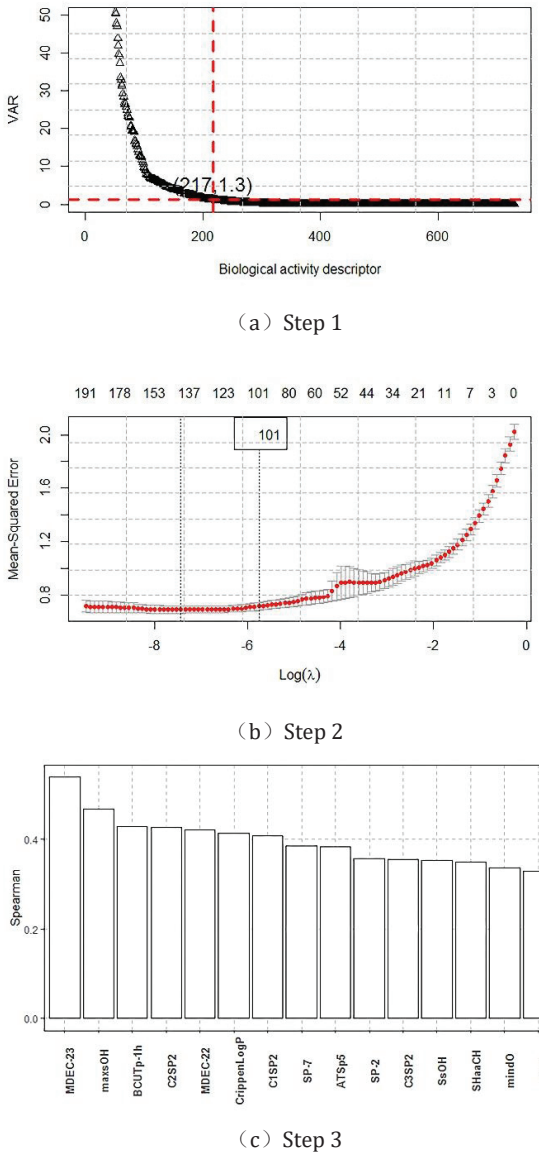
λ is a nonnegative regular parameter, which represents the complexity of the model. The greater its value, the greater the penalty of the linear model, The size of the regularization parameters was determined by cross-validation.

Step 3: Spearman rank correlation coefficient is a nonparametric index to measure the dependence of two variables, which can reflect the coupling degree between variables. This paper uses Spearman rank correlation coefficient to obtain the final 15 representative biological activity descriptors.

Three screening processes by figure 3 shows.

In step 1, 217 biological activity descriptors with variance greater than 1.3 were left.

In step 2, 101 bioactivity descriptors were retained by lasso feature selection.



(a) Step 1

(b) Step 2

(c) Step 3

Figure 3: Three screening processes.

In step 3, 101 biological activity descriptors are sorted according to Spearman rank correlation coefficient, leaving the most representative 15 biological activity descriptors. The final screening results are shown in table 1.

ADMET properties are composed of five aspects: absorption, distribution, metabolism, excretion and toxicity.

The corresponding values are provided in the form of two classifications, '1' represents good or yes, and '0' represents poor or no. Comparison table of ADMET properties are shown in table 2.

MODEL TRAINING AND PREDICTION

In order to avoid over fitting and improve the generalization ability of the model [14], we cut the remaining 15 bioactivity descriptors into 80% of the training set and 20% of the test set. Considering the coupling and the nonlinear relationship between the data, the neural network is used for training and prediction, the training set is used

to set the model parameters, and the test set is used to calculate the default accuracy and verify the rationality of the model. When training the model, we should also consider the convergence speed of the model. Neural network is a complex structure with large amount of calculation. When there are too many input variables in the input layer and the amount of data is too large, gradient optimization algorithm is usually used to accelerate the convergence speed of neural network. Adam algorithm is used for model optimization in this paper. The results are as follows:

As can be seen from figure 4 the red line is the logarithm of ERα, the blue line is the regression prediction result of neural network with one hidden layer, and the black line is the regression prediction result of neural network with two hidden layers. Among them, when the hidden layer is 1, the mean square error of prediction is 0.696, and when the hidden layer is 2, the mean square error of prediction is 0.759. Obviously, when the hidden layer is 1, the regression prediction result is more accurate, and the good prediction accuracy shows that the ERα activity can be controlled by controlling the 15 biological activity descriptors selected in this paper, so that we can inhibit the ERα activity.

In order to ensure that the selected bioactivity descriptors have good medical properties, the ADMET properties of these 15 bioactivity descriptors were verified. The commonly used machine learning methods are used for multiple prediction to eliminate contingency [15-17]. ROC curve shown in figure 5.

Table 1: Comparison table of biological activity descriptor.

Variable	Variable interpretation
MDEC-23	Molecular distance edge between all secondary and tertiary nitrogens
maxsOH	Maximum atom-type E-State: -OH.
BCUTp1h	nlow highest polarizability weighted BCUTS
C2SP2	Doubly bound carbon bound to two other carbons
MDEC22	Molecular distance edge between all secondary carbons
CrippenLogP	Crippen's LogP
C1SP2	Doubly hound carbon bound to one other carbon
SP7	Simple path, order 7
ATSp5	ATS autocorrelation descriptor, weighted by polarizability
SP2	Simple path, order 2
C3SP2	Doubly bound carbon bound to three other carbons
SsOH	Sum of atom-type E-State: -OH
SHaaCH	Sum of atom-type H E-State: CH
mindO	Minimum atom-type E-State: =O
SaaN	Sum of atom-type E-State: N

Table 2: Comparison table of ADMET properties.

Variable	Describe
Caco-2	Permeability of small intestinal epithelial cells
CYP3A4	Metabolize
hERG	Cardiotoxic
HOB	Oral bioavailability
MN	Genotoxic

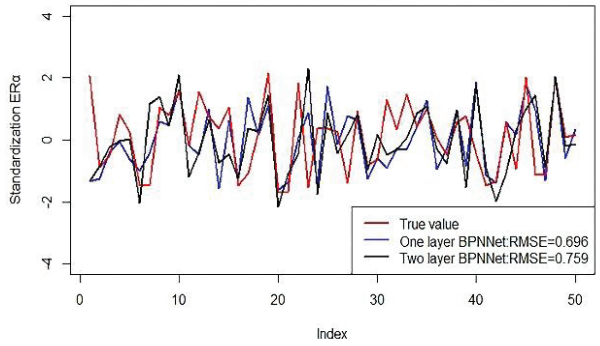
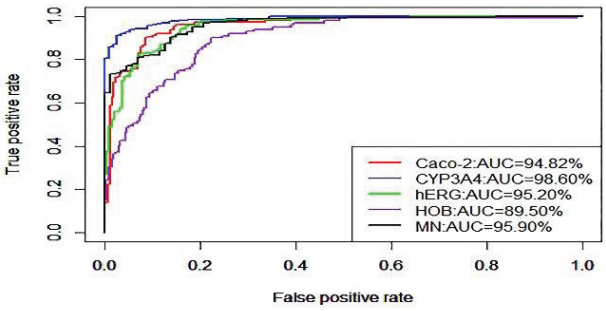
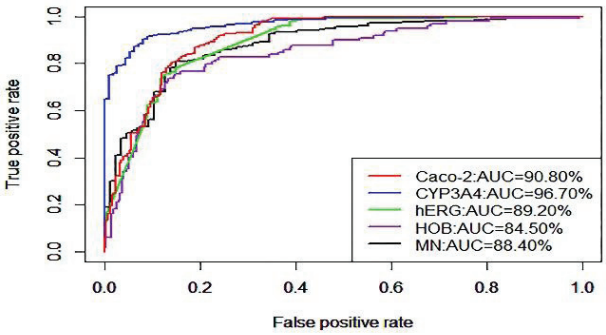


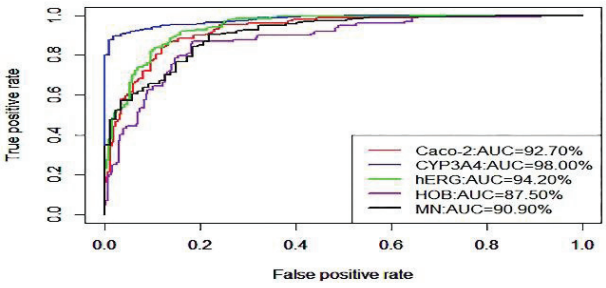
Figure 4: The predict of Adam-BPNNet.



(a) XGBoost-ROC.



(b) One layer BPNNet-ROC.



(c) GBM-ROC.

Figure 5: ROC curve of each classification method.

It can be seen from table 3 that the three models show very high prediction accuracy, as the table shows that the average accuracy of XGBoost is 0.948, the average accuracy of GBM is 0.927, the average accuracy of Adam-BPNNet is 0.899. Among the five metric variables that determine the pharmacokinetics, as well as the safety properties. The three models show that CYP3A4 is highly coupled with 15 biological activity descriptors, the coupling degrees are 0.986, 0.980 and 0.967 respectively. HOB is the lowest coupled with one biological activity descriptor, the coupling degrees are 0.895, 0.875 and 0.845 respectively. The pharmacokinetic and safety fit levels of the 15 biological activity descriptors selected in the article in the three models are between 0.8992 and 0.948, which are relatively stable. The average expression of drug permeability, absorption, metabolism, cardiotoxicity, genotoxicity and 15 biological activity descriptors reached more than 0.872. This shows that the 15 biological activity descriptors selected in this paper can not only reflect ERα activity to a great extent, it can also reflect good ADMET properties.

CONCLUSION

The results show that the 15 biological activity descriptors selected in this paper can predict ERα activity with a low mean square error of 0.676, which indicates that there is a high coupling between them. In addition, they can also reflect the properties of ADMET at an average level of 0.8992~0.948, so they have good medical value. The development of anti breast cancer drugs is a complex and long process. In this process, it is necessary to test the effects of drugs containing various biological components on target cells. If all the combined drugs are tested, it will be a long process. In order to improve the development cycle and cost of anti breast cancer drugs, we can consider using these bioactive descriptors to synthesize breast cancer resistant compounds.

Because the experimental data are limited, the influence of these 15 bioactive descriptors on the activity of other target cells is not considered. Therefore, the bioactive descriptors selected in this paper have limitations in the effect of breast cancer. Furthermore, lasso feature selection method is used to screen bioactivity descriptors, which may omit some important bioactivity descriptors.

When the synthetic breast cancer drugs are synthesized, the best value or range of bioactive descriptors can further reduce the development cost and development cycle of anti breast cancer drugs. Therefore, in this paper, we can further study the best values of various bioactive descriptors. At the same time, we also hope that the variable screening method and validation method can be applied to more biopharmaceutical processes.

Table 3: Comparison table of AUC values of different classification models.

Classifier	AUC
XGBoost	Caco-2: 0.948
	CYP3A4: 0.986
	hERG: 0.952
	HOB: 0.895
	MN: 0.959
GBM	Caco-2: 0.927
	CYP3A4: 0.980
	hERG: 0.942
	HOB: 0.875
	MN: 0.909
Adam-BPNNet	Caco-2: 0.908
	CYP3A4: 0.967
	hERG: 0.892
	HOB: 0.845
	MN: 0.884

ACKNOWLEDGEMENT

This work was supported by the Philosophical and Social Sciences Research Project of Hubei Education Department (19Y049), and the Staring Research Foundation for the Ph.D. of Hubei University of Technology (BSQD2019054), Hubei Province, China.

REFERENCES

1. Elixhauser A. Costs of breast cancer and the cost-effectiveness of breast cancer screening. Int J Technol Assess Health Care. 1991;7(4):604-15. doi: 10.1017/s0266462300007169. PMID: 1778705.
2. Howell A. Clinical evidence for the involvement of oestrogen in the development and progression of breast cancer. Proceedings of the Royal Society of Edinburgh. Section B. Biological Sciences. 1989;95:49-57.
3. Duan XN, Bai WP. China consensus on the safety management of endometrial endometrium in patients with breast cancer treated with selective estrogen receptor modulators (2021 Edition). Journal of Capital Medical University. 2021;42(4):672-677.
4. Zhang R, Du QL. Osteoglycin inhibits proliferation of Luminal breast cancer cells by up regulating estrogen receptor expression. Chinese Journal of Basic and Clinical Medicine. 2021;28(1):1-7. doi:10.7507/1007-9424.202106119.
5. Chen Q, Meng G, Huang W. The prognostic relationship between ER and PR expression in breast cancer and affecting the positive rate in association. Journal of Clinical and Experimental Pathology. 2016;32(01):13-18.
6. Zhu XX, Yang HY. Estrogen receptor subtypes α and β Expression and clinical significance in breast cancer. China Clinical Oncology and rehabilitation. 2014;21(6):766-768.
7. Liu R, Zhao J. Estrogen receptor subtype ERα; ERβ Expression in breast cancer. China Journal of Gerontology. 2012;32(16):3389-3390.
8. Smith T. Pharmacokinetics. Cambridge: Cambridge University Press; 2016. p. 562-577. doi: 10.1017/9781139626798.031.
9. Wang XM, Chen R, Qiao B. Application of BP Neural Network in Tea Disease Classification and Recognition. Guizhou Science. 2020;38(4):93-96.
10. Cao YF, Zhao YJ. Research on Computer Intelligent Image Recognition echnology based on GA-BP Neural Network. Applied Laser. 2017;37(1):139-143.
11. Qi Y. Experimental Study on NDVI Inversion Using GPS-R Remote Sensing Based on BP Neural Network. Xuzhou: China University of Mining and Technology; 2018.
12. Kingma D, Ba J. A Method for Stochastic Optimization. 3rd International Conference for Learning Representations. San Diego, 2015. https://tinyurl.com/4cmy3s3j
13. Li H. Statistical Learning Methods (in Chinese). Beijing: Tsinghua University Press; 2012. https://tinyurl.com/2p9d6z3j
14. Adcock B, Hansen A. The LASSO and its Cousins. Cambridge University Press; 2021. p. 129-141.
15. Ma B, Yan G, Chai B, Hou X. XGBLC: An Improved Survival Prediction Model Based on Xgboost. Bioinformatics. 2021 Sep 29:btab675. doi: 10.1093/bioinformatics/btab675. Epub ahead of print. PMID: 34586380.
16. Lombardi G, Bergo E, Caccese M, Padovan M, Bellu L, Brunello A, Zagonel V. Validation of the Comprehensive Geriatric Assessment as a Predictor of Mortality in Elderly Glioblastoma Patients. Cancers (Basel). 2019 Oct 9;11(10):1509. doi: 10.3390/cancers11101509. PMID: 31600898; PMCID: PMC6826848.
17. Korvink M, Martin J, Long M. Real-Time Identification of Patients Included in the CMS Bundled Payment Care Improvement (BPCI) Program. Infection Control & Hospital Epidemiology. 2020;41(S1):S367-S368. doi: 10.1017/ice.2020.993.