Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 2, No. 3, 2021

An Optimized Modeling Algorithm for Breast Cancer Drug Candidates Based on NSGAII

Chenyao Fan, Huawei Mei

North China Electric Power University, Baoding, Hebei, China

Abstract: Breast cancer is one of the most common malignant tumors in women. It seriously threatens the safety of women worldwide. It is an important and urgent task to research and develop anti-breast cancer drugs and improve the therapeutic effect of breast cancer. Taking the actual sample data as the main starting point, firstly, the prediction model of pIC50 is established by ResNet residual network and neural network (NN) to judge the biological activity. Then the classification model of ADMET property is established by ResNet residual network and LightGBM, and the model fusion is realized by Choquet fuzzy integral. Finally, the NSGAII multi-objective optimization algorithm is used to determine the range of values that each molecular descriptor obtains in the range of good biological activity, and ultimately to optimize the modeling of anti-breast cancer drug candidates. The experimental results show that the algorithm improves the prediction accuracy of biological activity, realizes the efficient and accurate classification of ADMET properties, and accurately describes the impact of molecular descriptors on biological activity.

Keywords: ResNet residual network, NSGAII multi-objective optimization algorithm, Choquet fuzzy integral; bioactivity prediction, Neural network.

1. Introduction

Breast cancer is a common cancer nowadays, also known as pink killer, which has a high incidence rate of over 20% in women worldwide, ranking the first place in women's cancer. Related studies [1] have found that estrogen receptor α Subtype (ERα) is expressed in 50% - 80% of breast tumor cells, and the experiment of ERα gene deletion mice also confirmed that ERα has a great impact on the development of breast. Therefore, ERα is considered to be an important target for breast cancer treatment, and it can antagonize Erα. α Active compounds are considered candidates for possible treatment of breast cancer. Screening for potentially active compounds economically and efficiently has also become a key link in the development and treatment of breast cancer drugs.

At present, in order to save time and cost in drug research and development, the method of establishing compound activity prediction model is usually used to screen potential active compounds. The specific method is: for a target (ERα) related to disease, collect a series of compounds acting on the target and their bioactivity data, and then take a series of molecular structure descriptors as independent variables and the bioactivity value of the compound as dependent variables to construct the quantitative structure-activity relationship (QSAR) model of the compound, Then the model is used to predict new compound molecules with better biological activity, or to guide the structural optimization of existing active compounds.

In the process of drug research and development, a compound activity prediction model is usually established to screen suitable active compounds. The steps are as follows:

(1) Find out the effect on ERα The bioactivity data of the compounds were analyzed;

(2) Taking the molecular structure descriptor as the independent variable and the bioactivity value of the compound as the dependent variable, the relationship model between them was established;

(3) The established model was used to predict and screen the compounds with high biological activity for in-depth research and analysis.

Qualified breast cancer drugs need to meet the properties of ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity). Popularly speaking, it is necessary to have good biological activity, pharmacokinetic properties, and safety, which can effectively prevent cancer at a moderate metabolic rate without toxic side effects. In order to facilitate modeling, only five ADMET properties such as intestinal epithelial cell permeability (Caco-2), cytochrome P450 (CYP), 3A4 subtype (CYP3A4), human ether-a-go-go related gene (hERG), human oral bioavailability (HOB), and micronucleus test are considered. Finally, when looking for the value range of some molecular descriptors, the compounds can have a better biological activity to inhibit ERα and better ADMET properties.

2. Method

2.1. Data Preprocessing

The flow chart of data preprocessing is shown in Figure 1. In this paper, the Spearman method, grey correlation analysis method, and principal component analysis method are used for correlation analysis to preliminarily observe the correlation between each molecular descriptor and compound activity, so as to provide a reference for future work; In order to make efficient use of the limited data, this paper uses the cross-validation method to construct the training data and test data, uses the LightGBM model training data integrating the tree model and L1 and L2 penalty values, and sorts the importance of the molecular descriptors. Five importance ranking documents are obtained through five-fold cross-validation, and the final correlation degree is obtained by weighted summation of five correlation degrees, so as to obtain the results of statistical analysis.

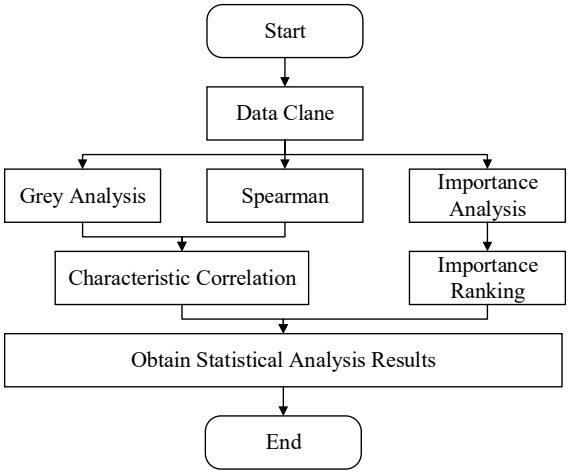

Figure 1. Flow chart of data preprocessing

2.1.1. LightGBM Model Integrating Tree Model with L1 and L2 Penalty Values

The main idea of LightGBM[4] is to use a weak classifier (i.e. decision tree) for iterative training to obtain the optimal model. LightGBM supports efficient parallel training and has faster training speed, lower memory consumption, better accuracy, and can quickly process massive data. Its principle is as follows:

LightGBM takes the CART regression tree as the decision tree and introduces L1 and L2 penalty coefficients. In addition to screening the features, it also reduces the dimension. The specific work steps are as follows:

As shown in equation (1), first initialize the weak learner:

$$f_0(x) = -[argmin_c \sum_{i=1}^{N} L(y_i, c)] \quad (1)$$

For $m = 1,2,\ldots,M$, Calculate the negative gradient of each sample $i = 1,2,\ldots,N$ according to equation (2), i.e. residual:

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad (2)$$

Take the residual obtained in the previous step as the new real value of the sample, and take the data (xi， rim), $i = 1,2,\ldots,N$ as the training data of the next tree to obtain a new regression tree fm(x), whose corresponding leaf node area is Rjm, $j = 1,2,\ldots,J$, where J is the number of leaf nodes of regression tree t.

Calculate the best fitting value for the leaf area $j = 1,2,\ldots,J$, as shown in formula (3):

$$r_{jm} = argmin \sum_{xi \in R_{jm}} L(y_i, f_{m-1}(x_i) + r) \quad (3)$$

Then update the strong learner, as shown in equation (4):

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} r_{jm} I(x \in R_{jm}) \quad (4)$$

Finally, the final learner is obtained, as shown in equation (5):

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} r_{jm} I(x \in R_{jm}) \quad (5)$$

2.1.2. Grey Correlation Analysis

Grey correlation analysis is a method to measure the degree of correlation between factors according to the degree of similarity or difference in the development trend between factors, that is, "grey correlation degree". The specific steps are as follows:

Determine the reference sequence reflecting the characteristics of system behavior and the comparison sequence affecting system behavior. The data sequence reflecting the behavior characteristics of the system is called the reference sequence. The data sequence composed of factors affecting system behavior is called comparison sequence.

The reference sequence and comparison sequence are dimensionless. Due to the different physical meanings of various factors in the system, the dimensions of the data are not necessarily the same, which is not convenient for comparison, or it is difficult to get a correct conclusion during comparison. Therefore, dimensionless data processing is generally necessary for grey correlation analysis.

Find the grey correlation coefficient between reference sequence and comparison sequence ζ_{X_i}. The so-called correlation degree is essentially the difference degree of geometry between curves. Therefore, the difference between curves can be used as a measure of correlation degree. For a reference sequence X0, there are several comparison sequences $X_1, X_2, ..., X_n$, and the correlation coefficient between each comparison sequence and the reference sequence at each time (i.e. each point in the curve) ζ_{X_i} can be calculated by the following formula: ρ The resolution coefficient is generally between 0 and 1, usually 0.5. $\Delta(min)$ is the second level minimum difference, $\Delta(max)$ is the maximum difference between two levels. The absolute difference between each point on the curve of the comparison sequence X_i and each point on the curve of the reference sequence X_0 is recorded as $\Delta_{oi}(k)$。 So correlation coefficient ζ_{X_i} can also be simplified as equation (6):

$$\zeta_{X_i} = \frac{\Delta(min) + \rho\Delta(max)}{\Delta_{oi}(k) + \rho\Delta(max)} \quad (6)$$

Calculate the correlation degree. Because the correlation coefficient is the correlation degree value of the comparison series and the reference series at each time (i.e. each point in the curve), it has more than one number, and the information is too scattered to facilitate the overall comparison. Therefore, it is necessary to concentrate the correlation coefficient of each time (i.e. each point in the curve) into one value, that is, calculate its average value as the quantitative expression of the correlation degree between the comparison series and the reference series. The correlation degree formula is as follows (7):

$$r_i = \frac{1}{N} \sum_{k=1}^{N} \zeta_i(k) \quad (7)$$

r_i is the grey correlation degree of the comparison sequence xi to the reference sequence X_0, or called sequence correlation degree, average correlation degree and line correlation degree. The closer the r_i value is to 1, the better the correlation.

Relevance ranking. The correlation degree between factors is mainly described by the order of correlation degree, not only the size of correlation degree. The correlation degree of M subsequences to the same parent sequence is arranged in order of magnitude to form the correlation order, which is

recorded as $\{X\}$, which reflects the "good and bad" relationship of each subsequence to the parent sequence. If $r_{0i} > r_{0j}$, $\{X_i\}$ is said to be better than $\{X_j\}$ for the same parent sequence $\{X_0\}$, and is recorded as $\{X_i\} > \{X_j\}$; R0i represents the eigenvalue of the i-th subsequence to the parent sequence.

2.2. The Bioactivity Prediction Model of pIC50 was Established

There is not a simple linear relationship between molecular descriptors and compound molecular activity, so cluster analysis and decision tree are not suitable. A neural network is needed to solve this nonlinear problem. Firstly, based on the above work, this paper selects the molecular descriptor variables with high correlation with compound activity and then carries out the feature construction, feature dimension increase, and PCA dimension reduction of the variables takes them as the feature parameters and selects the pIC50 parameter between pIC50 and IC50_nM as the measurement standard of biological activity. In the part of the neural network model, this paper intends to use LightGBM, traditional neural network, and ResNet residual network to establish models respectively for model comparison, and finally do model fusion. The overall solution flow is shown in Figure 2.

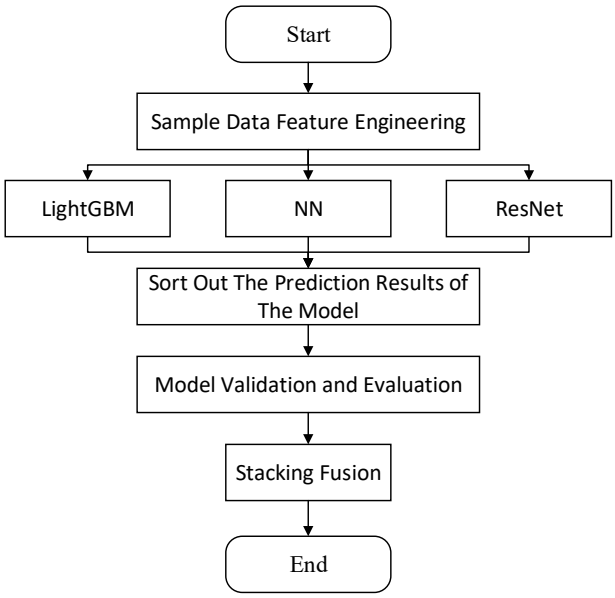

Figure 2. Flow chart of building prediction model

2.2.1. Selection of Measurement Criteria

Two groups of values are used in the compound molecular activity table to quantify the molecular activity, namely pIC50 and IC50_nM. IC50_nM is the semi inhibitory concentration (or semi inhibitory rate), that is, the concentration of drugs or inhibitors required to inhibit half of the specified biological process (or a component in the process, such as enzyme, receptor, cell, etc.), It is used in pharmacy to characterize the Antagonistic Ability of antagonists in vitro. There is a positive correlation between plc50 and IC50_nM, and both have the function of characterizing the molecular activity of compounds. Observe the data distribution of pIC50 and IC50_nM respectively. As shown in Figure 3, the distribution of IC50_nM is a heavy-tailed distribution, which is quite different from the normal distribution, which is not conducive to model training. Moreover, according to the general data processing method, the long tail on the right needs to be truncated, which is not applicable when the amount of data given in this question is small. It can be seen that the distribution of pIC50 is relatively uniform and close to the normal distribution, which is more suitable as the result set than IC50_nM. Therefore, pIC50 is selected as the measurement standard of biological activity in this paper.

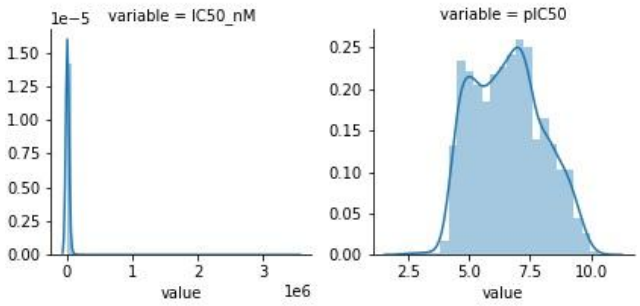

Figure 3. Data distribution

2.2.2. ResNet

Compared with ordinary deep neural networks, ResNet [3] has many "bypasses", i.e. shortcut paths. The layers circled at the beginning and end form a residual unit. Each residual unit is stacked by convolution Conv layer, batch normalized BN layer, and nonlinear activation function Relu layer.

It has two layers, as shown in the formula, in which σ Represents the nonlinear function ReLu; Then, the output y is obtained through a shortcut and the second ReLu. When the input and output dimensions need to be changed (such as changing the number of channels), a linear transformation Ws can be made for x during the shortcut. The experiment shows that there is no need for another dimension transformation. Unless the demand is the output of a specific dimension, the residual block often needs more than two layers, A single layer of residual blocks does not improve. This residual element has two layers. The expression of F is as shown in equation (8), and then the output y is obtained through a shortcut and the second ReLu, as shown in equation (9).

$$F = W_2\sigma(W_1x) \tag{8}$$

$$y = F(x,\{W_i\}) + x$$
$$y = F(x,\{W_i\}) + W_sx \tag{9}$$

In ResNet, all residual blocks do not have a pooling layer, and down sampling is realized through the stride of Conv; Get the final feature through average pooling, not through the full connection layer; Each convolution layer is followed by a Batch Norm layer. ResNet structure is very easy to modify and expand. By adjusting the number of channels in the block and the number of stacked blocks, you can easily adjust the width and depth of the network to obtain networks with different expression abilities without worrying too much about the "degradation" of the network. As long as the training data is sufficient and the network is gradually deepened, you can obtain better performance.

2.3. Establish ADMET Property Classification Model

According to the properties of ADMET, they are divided

into two categories. First, we need to process the molecular descriptor data, eliminate the columns with unique values, and normalize all the data. LightGBM, Logistic, and ResNet were used to establish models and compare them. The training and prediction are carried out through 50% cross-validation of the LightGBM model and logistic model. The model fusion of the three models is realized by Choquet fuzzy integral [5], and the final prediction results are obtained. The idea flow chart of question 3 is shown in Figure 4.

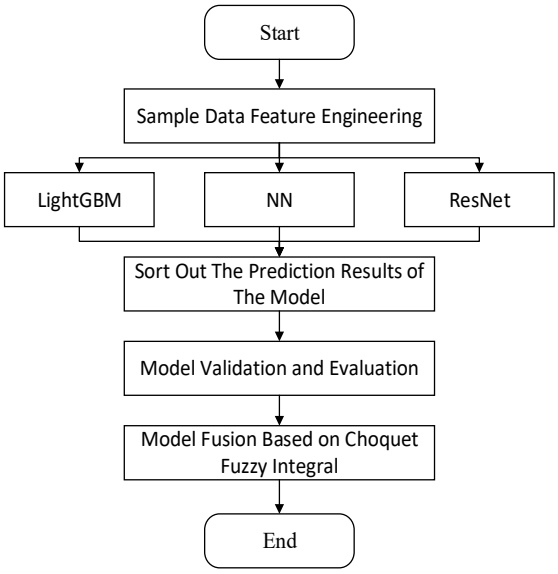

Figure 4. Flow chart of establishing classification model

2.3.1. Choquet Fuzzy Integral

Choquet integral is an integral representation theory based on convex set and convex cone theory. For any $x \in K$ in compact convex set K in locally convex space, there is a probability measure on the endpoint set $extK$ of K, so that for any $x^* \in X^*$, there is equation (10):

$$\langle x^*, x\rangle = \int_{extK} \langle x^*, y\rangle d\mu_x(y) \quad (10)$$

2.3.2. Logistic Regression Model

Logistic regression is a common binary classification algorithm model. Although it has the word regression, it is a classification model. It is mainly used in epidemiology. It is often used to explore the risk factors of the disease and predict the probability of a disease according to the risk factors, which is similar to the background of this question. The basic form of the logistic model is shown in equation (11). However, in the practical application of the model, P is generally not directly regressed, but the monotonic continuous probability function π is defined first, as shown in equation (12).

$$P(Y = 1|x_1, x_2 \ldots x_k) = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)} \quad (11)$$

$$\pi = P(Y = 1|x_1, x_2 \ldots x_k) 0 < \pi < 1 \quad (12)$$

Therefore, the logistic model is transformed into equation (13):

$$P(Y = 1|x_1, x_2 \ldots x_k) = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)} \quad (13)$$

According to the above principles, this paper establishes a logistic binary classification model, taking BCE as the loss function, and its calculation formula is shown in equation (14). Where y is the label 0 or 1, a is the output of the network through the sigmoid function, the range is (0,1), and N is the number of samples. The network outputs a value, and BCE gives a loss according to the value and the corresponding label.

$$loss = -\frac{1}{n}\sum_x [y\, ln\, a + (1-y)\, ln(1-a)] \quad (14)$$

All models adopt 50% cross validation, and then realize data fusion through Choquet fuzzy integral, so as to improve the generalization ability of the model.

2.4. NSGA II Algorithm

NSGAII algorithm [2] is a genetic algorithm that sorts the non dominated solutions mixed by parents and children, selects the winning individuals according to the crowding degree and crowding operator in the same level order, then retains the elite solutions, and continues to cross, mutate and select until the final convergence. It proposes a fast non dominated sorting. The core idea of fast non dominated sorting is to obtain the dominated degree Np of each P in the population through calculation and comparison, and obtain a multi-layer non dominated surface through the size of the dominant degree, which can well reduce the complexity of the algorithm. In addition, the NSGAII algorithm also proposes a crowded comparison operator to avoid the uncertainty of human input parameters.

(1) Non-dominated solution: assuming that any two solutions S1 and S2 are better than S2 for all targets, S1 dominates S2. If the solution of S1 is not dominated by other solutions, S1 is called non dominated solution (non dominated solution), also known as Pareto solution.

(2) Congestion comparison operator. Firstly, the crowding distance is defined to quantitatively calculate the density of each individual. The calculation of crowding distance needs to first sort each individual in the same non dominated surface in ascending order on each target, in which the crowding distance of the individuals with the maximum and minimum values on this target is set to infinity. The crowding distance of individual i is equal to the sum of the absolute values of the difference between the function values of individual i-1 and i+1 on each target. Through the crowding distance, the crowding comparison operator can be obtained.

(3) According to the above model principle, firstly, set the model objective function f(x), f1(x) as the value of pIC50 predicted by molecular descriptor data to analyze the biological activity. The greater the value, the higher the biological activity; f2(x) is to predict the ADMET property through the molecular descriptor data and calculate its score. Through the property analysis, it is proposed to adopt the case that the Coca-2 value is 1, CYP3A4 value is 1, hERG value is 0, HOB value is 1 and MN value is 0, indicating that the five properties are better. Then set the inequality constraint condition of the model, g1(x) is the predicted pIC50 value, which shall not be less than a certain value; g2(x) is the predicted five ADMET properties, and at least three properties are good. Finally, the objective function and constraints are obtained, as shown in equation (15):

$$\begin{aligned} f1(X) &= predict2(X) \\ f2(X) &= predict3(X) \\ g1(X) &= pIC50 - predict2(X) \\ g2(X) &= 3 - predict3(X) \end{aligned} \quad (15)$$

3. Experimental Results

3.1. Fitting Effect of Prediction Model

Comparing the predicted value of the LightGBM model with the real value, the error is obvious, and the average absolute error (MAE) is about 0.42; Compared with the real value, the predicted value of the NN model and ResNet has a good fitting effect. The predicted value accurately describes the changing trend of the real value, and the average errors are about 0.07 and 0.12 respectively. Figure 5 shows the comparison between the predicted value and the real value of ResNet. Table 1 shows the comparison of evaluation indexes of the three models.

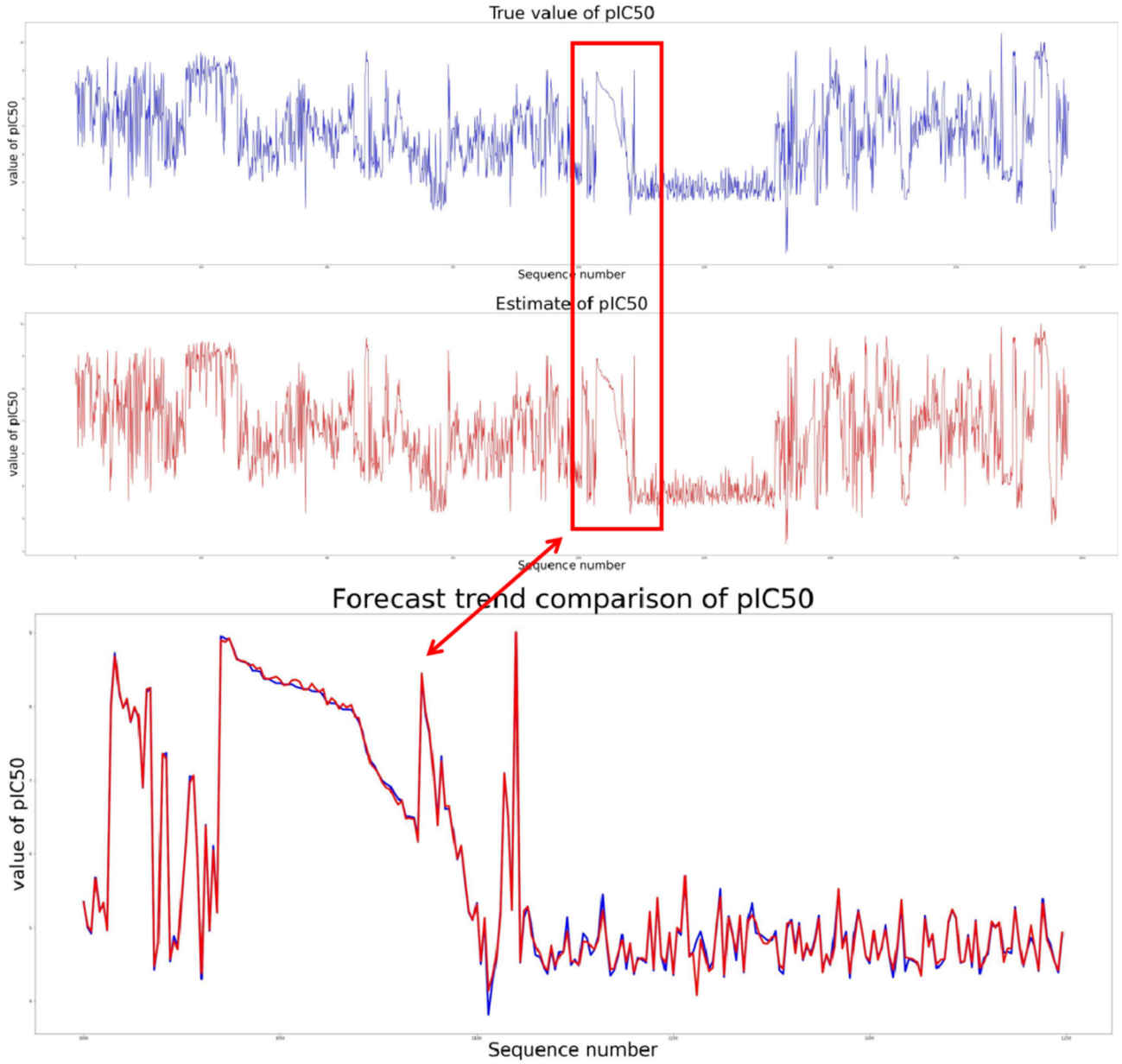

Figure 5. Comparison between predicted value and real value of ResNet

Table 1. Comparison of evaluation indexes of each model

Model Type	MAE
LightGBM	0.422
NN	0.140
ResNet	0.069

Table 1 shows the average absolute errors of the above three models are 0.422, 0.140, and 0.069. From the analysis of the above table, it can be seen that the prediction data of NN and ResNet models have better estimation advantages. The average absolute error is as low as 0.06.

3.2. Fitting Effect of Classification Model

The sample data of main variables are divided into a training set and verification set, in which the verification set accounts for 20% of the total data set, and the test results are shown in the following figure:

Figure 6 shows the prediction results of LightGBM, Logistic, ResNet, and Choquet fusion models for the training set, "1" means the prediction is correct, and "0" means the

prediction fails. It can be seen that the overall prediction effect of the four models is good, in which the prediction effect error of the logistic model is obvious, and the prediction effect of the Choquet fusion model is more ideal.

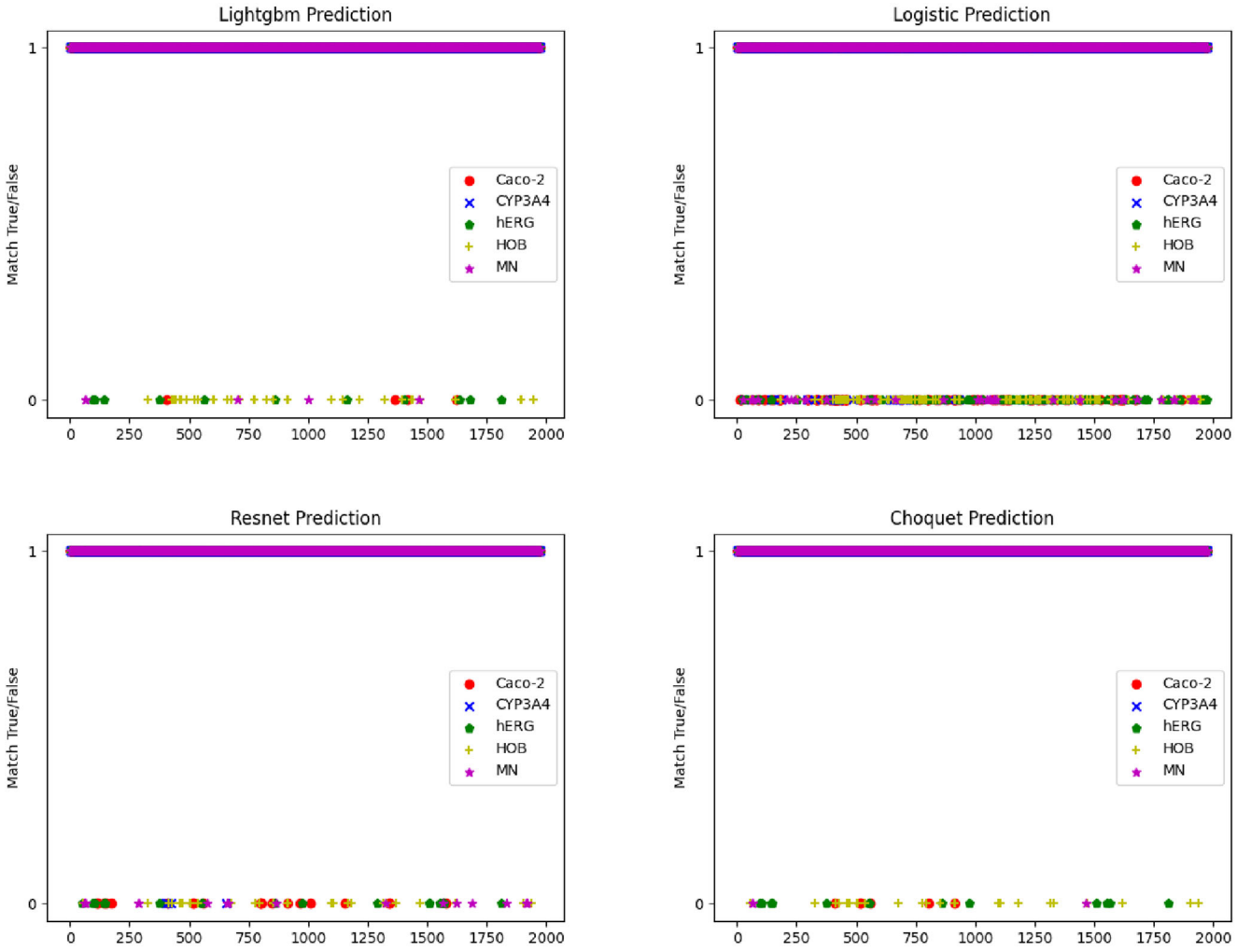

Figure 6. Prediction results of four models

Firstly, this paper analyzes that there is no extreme imbalance in the distribution of data categories in ADMET, and uses accuracy and F1-score as the evaluation index of the model. Figure 7 shows the accuracy evaluation results of LightGBM, Logistic, ResNet and Choquet fusion models.

Table 2. Comparison of model evaluation indexes

Model Type	F1-Score	Accuracy
LightGBM	0.87	0.94
Logistics	0.84	0.87
ResNet	0.93	0.97
Choquet fusion model	0.93	0.98

Table 2 shows that the results of Choquet fusion after 50% cross-validation improve the generalization ability of the model and improve the accuracy of model classification to varying degrees. Among them, the effect of the Logistics model is not ideal, the effect of the Choquet fusion model is the best, and each model has a good classification effect on the properties of CYP3A4 in ADMET.

3.3. Optimizing Modeling Effect

The numerical variation trend of molecular descriptors is shown in Figure 8. More than 70% of molecular descriptors can basically outline the value range under the condition of good biological activity. The value range results of some representative molecular descriptors are shown in Table 3.

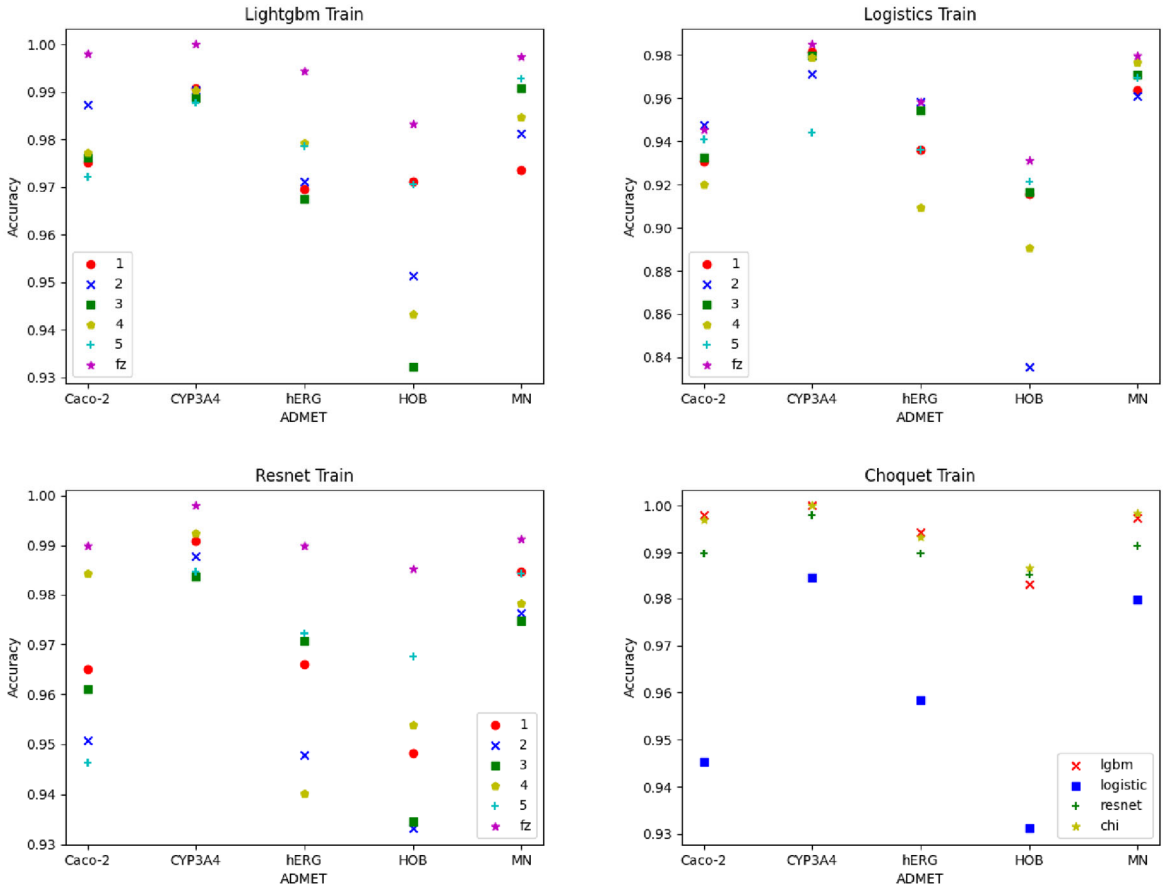

Figure 7. Accuracy evaluation results of four models

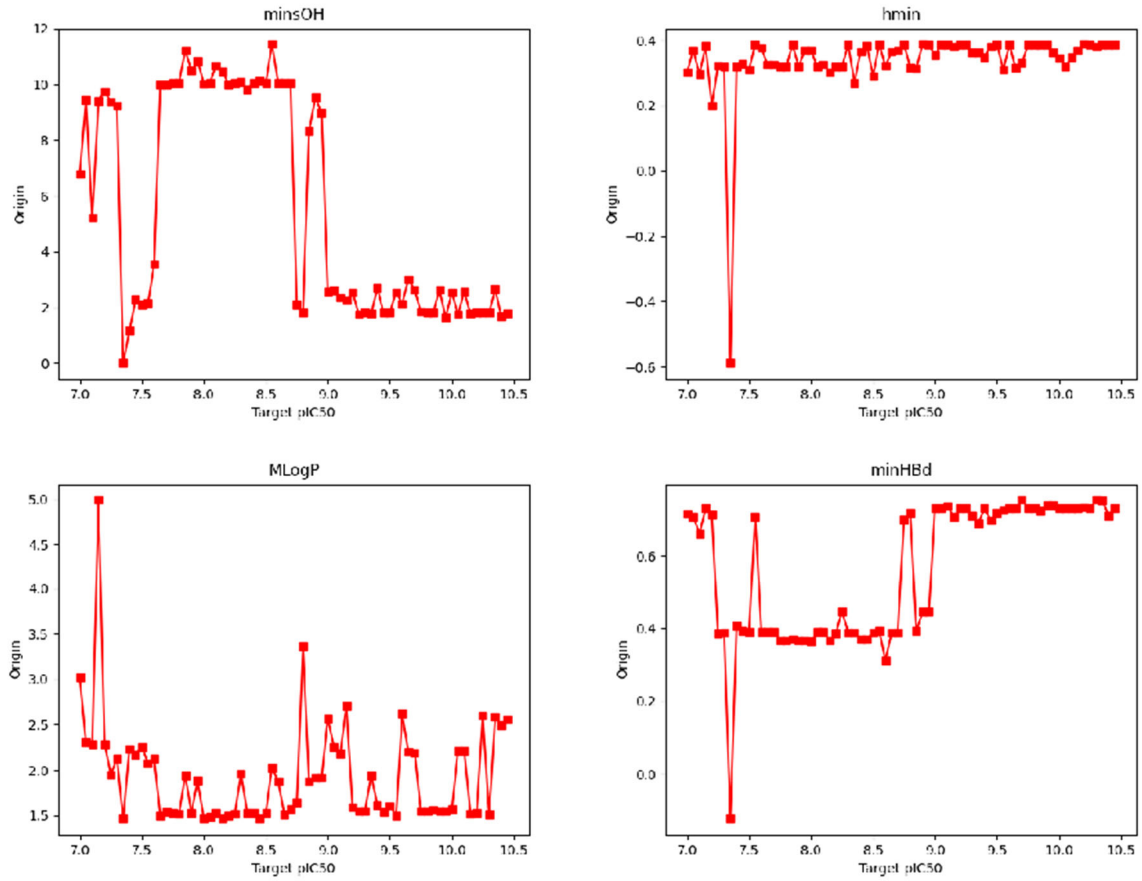

Figure 8. Change trend of some molecular descriptors

Table 3. Value range of partial molecular descriptors

Molecular Descriptor	Value Range
ATSc5	[0.4, 0.6]
ATSc4	[-0.9, -0.7]
ATSc3	[0.4, 0.45]
ALogp2	[140, 150]
XLogP	[11, 13]
BCUTp-1l	[5.3, 5.7]
ETA_Shape_Y	[0.38, 0.42]
minHBd	[0.7, 0.75]
hmin	[0.3, 0.4]
minsCH3	[2.3, 2.5]
minHBa	[12, 14]
ETA_Shape_P	[0.3, 0.35]
TopoPSA	[840, 900]
MDEC-22	[5, 10]

4. Conclusion

In this paper, we make full use of grey correlation analysis, Spearman correlation coefficient analysis, fusion tree model, and L1, L2 penalty item LightGBM, ResNet, NN, Logistic, Choquet fuzzy integral, and NSGAII algorithm to study the optimal modeling of anti-breast cancer drug candidates. The depth learning model based on NN neural network and ResNet residual network have higher accuracy, better fitting effect, and more accurate prediction results. Through the comparison of the three models, the accuracy of the prediction results of this question is ensured. The fuzzy integral fusion multi-fold verification model is used to improve the accuracy and generalization ability. NSGAII algorithm is used to optimize the value of molecular descriptor, and the value range of molecular descriptor in the case of high biological activity is obtained. However, in the classification model, the distribution of target classes is still unbalanced, and the accuracy used in training may not well reflect the real generalization accuracy.

References

[1] Ni Wenting. Exploring the effect of cryptotanshinone on ERα+ based on BCRP Molecular mechanism of proliferation inhibition in breast cancer [D]. Nanjing University of Chinese Medicine, 2019.

[2] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE transactions on evolutionary computation, 2002, 6(2): 182-197.

[3] He K, Zhang X , Ren S , et al. Residual for Image Recognition[J]. IEEE, 2016.

[4] Meng Q. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2018.

[5] Islam M A, Anderson DT, Pinar A J , et al. Enabling Explainable Fusion in Deep Learning with Fuzzy Integral Neural Networks[J]. IEEE Transactions on Fuzzy Systems, 2019.