Mathematical Biosciences and Engineering

MBE, 20(3): 5117–5134.
DOI: 10.3934/mbe.2023237
Received: 24 October 2022
Revised: 07 December 2022
Accepted: 15 December 2022
Published: 06 January 2023

http://www.aimspress.com/journal/MBE

Research article

Optimal modeling of anti-breast cancer candidate drugs screening based on multi-model ensemble learning with imbalanced data

Juan Zhou[1], Xiong Li[1], Yuanting Ma[2,*], Zejiu Wu[3], Ziruo Xie[1], Yuqi Zhang[4] and Yiming Wei[1]

[1] School of Software, East China Jiaotong University, Nanchang 330013, China
[2] School of Economics and Management, East China Jiaotong University, Nanchang 330013, China
[3] School of Science, East China Jiaotong University, Nanchang 330013, China
[4] School of Foreign Languages, East China Jiaotong University, Nanchang 330013, China

* **Correspondence:** Email: yuantingma@189.cn, zhoujuan@ecjtu.edu.cn.

Abstract: The imbalanced data makes the machine learning model seriously biased, which leads to false positive in screening of therapeutic drugs for breast cancer. In order to deal with this problem, a multi-model ensemble framework based on tree-model, linear model and deep-learning model is proposed. Based on the methodology constructed in this study, we screened the 20 most critical molecular descriptors from 729 molecular descriptors of 1974 anti-breast cancer drug candidates and, in order to measure the pharmacokinetic properties and safety of the drug candidates, the screened molecular descriptors were used in this study for subsequent bioactivity, absorption, distribution metabolism, excretion, toxicity, and other prediction tasks. The results show that the method constructed in this study is superior and more stable than the individual models used in the ensemble approach.

Keywords: ensemble algorithm; imbalanced data; feature selection; estrogen receptor; ADMET

1. Introduction

Breast cancer is a very common female cancer in the world affecting women's health [1,2]. The development of breast cancer is closely related to estrogen receptor. Some studies have shown that estrogen receptor α subtypes are expressed in no more than 10% of normal breast epithelial cells, but about 50% in 80% of breast tumor cells. Therapies targeting estrogen receptor α (ER α) have

transformed the treatment of breast cancer [3]. ERα plays an important role in breast development. In general, compounds that antagonize ER α activity can be used as candidate drugs for breast cancer treatment, many scholars regard it as an important index for screening anti-breast cancer drugs [3,4]. However, in order to become a candidate drug, a compound needs not only good biological activity, but also good pharmacokinetic properties and safety in human body, which is known as ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) [5]. Among them, ADME mainly refers to the pharmacokinetic properties of the compound, describes the law of the concentration of the compound in the organism with time, and T mainly refers to the toxic and side effects that the compound may produce in the human body.

In the process of drug research and development, Quantitative Structure-Activity Relationship (QSAR) model and ADMET property relationship model needs to be constructed to save time and cost [6]. The model was then used to screen new compounds with better bioactivity and ADMET properties. However, in the practical screening process, excessive molecular descriptors will increase the noise of the data and the error of the learning algorithm [7], therefore, critical molecular descriptors should be selected as main indicators of relational models. Moreover, it is a crucial issue that data imbalance is quite common in the field of bioinformatics due to the limited availability of data, and the imbalance caused by the excessive difference in the number of samples from different categories tends to bias the prediction results of the model toward the category with a larger number, so overcoming the effect of imbalanced data is another crucial issue.

In recent years, machine learning methods have gained popularity in bioinformatics [8–11] and have been successful in dealing with issues such as feature selection for high-dimensional data [12,13], error correction [14,15], category imbalanced data [16,17], and other important problems [18–20]. Ensemble learning, as an efficient machine learning method, combines independent feature sub-models and might give a better approximation to the target dataset [7,21]. Whereas ensemble systems have been proven to be competent in reducing the variance of automated decision systems and are very effective and extremely versatile in a broad spectrum of problem domains and real-world applications [22].

Motivated by the idea above, we propose a novel framework for breast cancer drug screening. This framework can correct for errors among different algorithms to handle imbalanced data and can learn decision weights among individual models based on training data, which is described as follows:

1) A novel molecular descriptor filtering approach based on Random Forest, XGBoost, Lasso and neural network is constructed to calculate the importance of each variable for selecting the top 20 molecular descriptors with the most significant impact on biological activity from a large number of molecular descriptors.

2) An ensemble model based on Random Forest, XGBoost, Lasso and neural network is constructed to predict the ER α value, and the weights of sub-models are obtained adaptively by machine learning. Based on which, the pIC_{50} (negative logarithm of IC_{50}) value of compounds were predicted. The bioactivity value of the compound to ER α can be expressed by IC_{50}, which is experimentally measured values in nM. The smaller the IC_{50} is, the greater the biological activity is. In QSAR model, pIC_{50} is generally used to represent the bioactivity value.

3) The ensemble model described in 2) is applied to pharmacokinetic properties and safety (ADMET) prediction. The absorption, distribution, metabolism, excretion and toxicity of compounds were predicted respectfully. The results showed that the model in this study has strong stability and accuracy.

2. Materials and methods

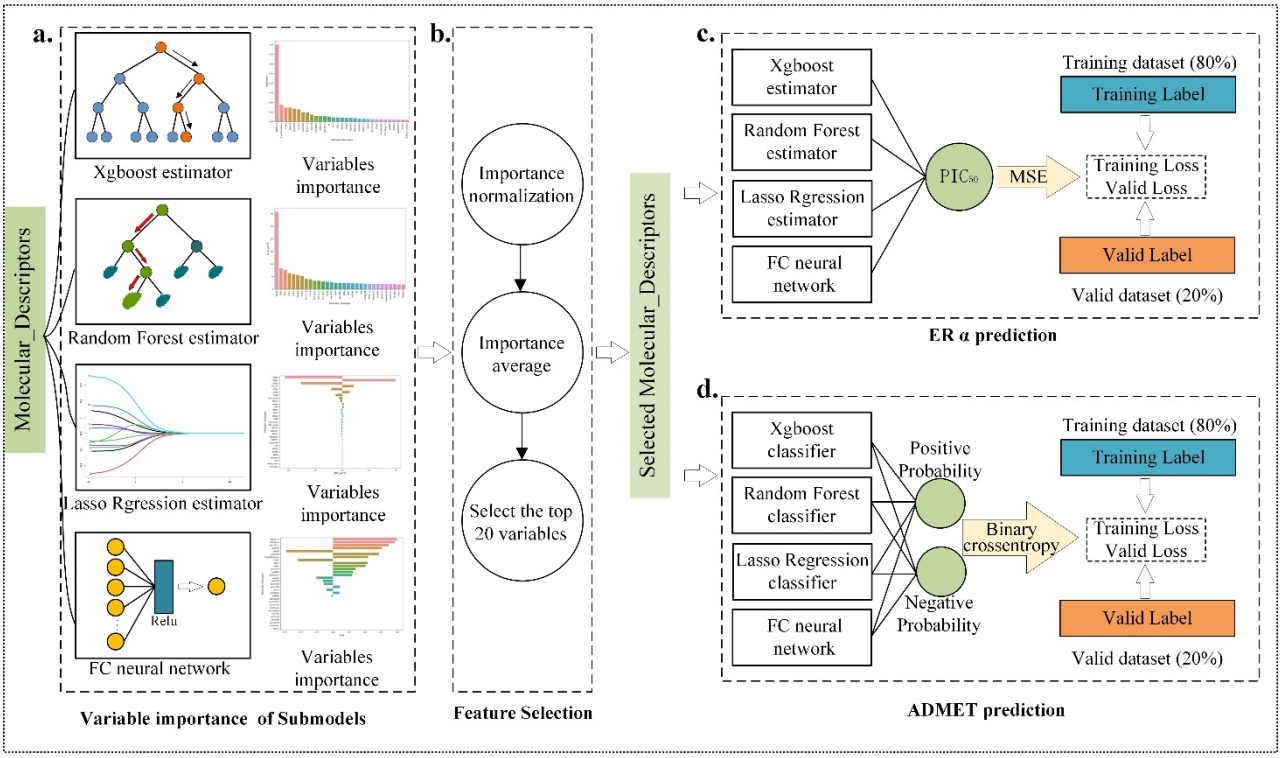

Figure 1. Ensemble framework for screening of therapeutic drugs.

The screening modeling of anticancer drug candidates is mainly to establish the prediction model of biological activity and classification models of ADMET properties. Firstly, key variables need to be identified from numerous molecular descriptors for drug candidates using ensemble method (see Figure 1(b)). Secondly, based on the variables selected before, a pIC_{50} prediction model need to be build (see Figure 1(c)). Finally, the ADMET properties' prediction model of drug candidates need to be built up separately (see Figure 1(d)). Figure 1 shows the algorithm framework for screening of therapeutic drugs, and the details of the entire modeling process are as follows.

2.1. Feature selection methods

Facing large number of variables, different features election and prediction algorithms may yield local optima in the space of feature subsets [7]. An ensemble feature selection algorithm could capture independent features from different sub-models, so we choose XGBoost, Random Forest, Lasso Regression and neural networks as sub-models of the ensemble model. Besides, an ensemble model also can fix the problem of imbalanced data, which is useful in ER α and ADMET prediction. Therefore, we need to screen the most critical input variables in different single models and measure the degree of importance of the variables using feature importance.

2.1.1. XGBoost feature selection

The essence of XGBoost is an effective boosting algorithm, which is a tree structure constructed

by a variety of weak classifiers. Each weak classifier corrects the previously misclassified samples, therefore, XGBoost is more focusing on reducing bias. One of XGBoost metrics for importance is "weights", and each node of XGBoost is "weakly classified" based on one variable, and the final feature score is obtained by adding the gradient The final feature score requires adding up the gradient and second-order gradient statistics on each leaf, and then applying the scoring formula. The higher the weight, the more important the corresponding variable is [23].

$$w_j^* = -\frac{\sum_{i\in I_j} g_j}{\sum_{i\in I_j} h_j+\lambda} \tag{1}$$

In Eq (1), w_j^* is the j-th leaf node of the current tree, g_j is the first−step degree of the ith sample falling on the j-th leaf node, h_j is the second-order gradient of the i-th sample falling on the JTH leaf node, and λ is the regularization coefficient.

2.1.2. Random Forest (RF) feature selection

Random forests are a combination of tree predictors [24], it is also an ensemble bagging algorithm. This algorithm builds up a large number of decision trees and gets the final result by equally voting, therefore, RF pays more attention to reducing variance. The most commonly used variable filtering indicator in RF models is Gini importance, which is the total reduction of criterion (Gini) brought by a feature, Similar to the boost method, the higher the Gini value, the more critical the corresponding input descriptor is.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} p(i|t)^2 \tag{2}$$

In Eq (2), t is on behalf of a given node, i represents any classification of the label, p(i|t) represents the proportion of label i on node t.

2.1.3. Lasso feature selection

Lasso is a classical linear regression method with L1-regularization. It reduces model complexity and makes model more robust by constructing L1 penalty term. At this point, the absolute value of the coefficient before the variable in lasso's regression measures the importance of the variable [25], it represents the marginal impact of the input variables on the final prediction, and a larger value means that the corresponding variable is more critical in the prediction.

$$\beta^* = argmin_\beta(\frac{1}{n} \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1) \tag{3}$$

In Eq (3), n is the number of samples, λ is a self-defined penalty coefficient and β is the coefficient vector of Lasso's regression.

2.1.4. Neural network feature selection

BP neural network was proposed in 1986 and has been proved to be useful in processing high dimensional data in the next decades [26–29], as a nonparametric model, which could deal with data

with complex noise [30]. Mean impact value (MIV) is an indicator to evaluate the importance of variables, and commonly used in measuring the importance of neural network's inputs [31,32].

$$MIV_i = \mathcal{F}(diag(\rho_1, \rho_2, \cdots \rho_i, \cdots \rho_n)X) - \mathcal{F}(diag(\sigma_1, \sigma_2, \cdots \sigma_i, \cdots \sigma_n)X) \tag{4}$$

$$\rho_k = \begin{cases} 1.1, & if\ k = i \\ 1, & if\ k \neq i \end{cases}, \ \sigma_k = \begin{cases} 0.9, & if\ k = i \\ 1, & if\ k \neq i \end{cases}$$

It indicates the difference between the outputs of a model when a variable increases or decreases by 10% independently, similarly, the larger the MIV value, the more prominent the corresponding molecular descriptor.

To correct for sub-method errors and to synthesize the importance of the features calculated by each sub-method, we adopted a fair weighting strategy, i.e., weighting by the standard deviation of the importance metrics calculated by each sub-method. (As shown in Figure 1(a),(b) and Eq (5))

$$Importance_i = \frac{1}{4}\left(\tilde{f}(w_i) + \tilde{f}(\nabla Gini_i) + \tilde{f}(\beta_i) + \tilde{f}(MIV_i)\right) \tag{5}$$

In Eq (5), $\tilde{f}(x_i) = (x_i - \bar{x}_i)/\sigma_i$, σ_i is the standard deviation of the importance value for i-th input variable, $\bar{x}_i$ is the mean value of the importance value for i-th input variable. This also means that in the stage of feature selection, the weight of each sub-method is determined by the distribution of its respective importance.

2.2. Ensemble model for ER α prediction

Multi-model ensemble learning combines multiple learners and can obtain better generalization ability than single learner and can correct errors among multiple models. Firstly, from a statistical point of view, the hypothesis space of task learning is usually large, and there may be multiple hypotheses that achieve the same performance in the training set [33]. Secondly, from the point of view of computation, a single algorithm is more likely to fall into local minima, and some local minima may lead to weak generalization ability of the model, but ensemble learning can effectively improve this situation through multiple runs. Thirdly, from the perspective of representation, the hypothesis space involved in a single learner may not contain the real hypothesis of the task, while ensemble learning can increase the hypothesis space.

Therefore, this study builds the QSAR model based on ensemble learning of two classical tree models, Lasso linear model and neural network model. Through this method, the bioactivity (ER α) of the compound is predicted. Lasso captures the linear relationship between molecular descriptors and biological activity; random forest (bagging) and XGBoost (boosting) aims to reduce variance and bias. Neural network was introduced to measure the complex nonlinear relationship between molecular descriptors and biological activity. Finally, the outputs of the four sub-models are fused by a fully connected layer (as shown in Figure 1(c)), which is also a commonly used method in machine learning [34,35]. The parameters of the full connected layer and the neural network will be trained by optimizing the MSE loss function.

$$\mathcal{L}' = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 \tag{6}$$

ER α QSAR model can also be written as:

$$\hat{y} = \eta_1 XGBoost(x) + \eta_2 RF(x) + \eta_3 \sum \beta x + \eta_4 NN(x) \quad (7)$$

Equation (7) shows the expanded form of the final prediction, where η_1 to η_4 are the decision coefficients for each of the four submodels, and we fit the objective function to assign the values of the four parameters. This means that the weights of each submodel are also adaptively determined by the data in the training set during the prediction phase.

2.3. Ensemble model for ADMET prediction

To classify and predict the properties of ADMET, this paper finishes the binary classification task of Caco-2, CYP3A4, hERG, HOB and MN. Caco-2: '1' represents that the intestinal epithelial cells of this compound have good permeability, and '0' represents that the intestinal epithelial cells of this compound have poor permeability. CYP3A4: '1' means that the compound can be metabolized by CYP3A4, '0' means that the compound cannot be metabolized by CYP3A4; hERG: '1' means that the compound is cardiotoxic, '0' means that the compound is not cardiotoxic; HOB: '1' means that the oral bioavailability of the compound is good, and '0' means that the oral bioavailability of the compound is poor. MN: '1' means that the compound is genotoxic, and '0' means that the compound is not genotoxic. First, the top 20 molecular descriptors with the most significant impact on the five dependent variables were selected. Secondly, based on the most significant molecular descriptors, similar to 2.1, five prediction models for Caco-2, CYP3A4, hERG, HOB and MN were established respectively under the model framework of Figure 1(d).

Different from 2.2, both the neural network sub-model in ADMET prediction and the final output are two 0/1 values, respectively representing the probability of being divided into positive and negative, while other sub-models are added a classifier with threshold value of 0.5 on the basis of Figure 1(c).

Parameters of the full connection layer and neural network are obtained by optimizing the Binary cross-entropy loss function:

$$\mathcal{L}^* = -\frac{1}{n}\sum_x [y ln\hat{y} + (1-y)\ln(1-\hat{y})] \quad (8)$$

In Eq (8), $\hat{y}$ is the conditional estimation of given molecular descriptors, and y is the true label of the compound.

3. Experimental results and decision

3.1. Data description and preprocessing

The data used in this paper is obtained from ChEMBL database (version: ChEMBL27) (www.ebi.ac.uk/chembl/), which Search with 'Estrogen receptor alpha' as the keyword. Click 'Homo sapiens' in the 'Organism' menu of the filter, and the ChEMBL ID of the Target was confirmed as 'CHEMBL206'. Download IC_{50} data related to 'CHEMBL206' in the database. To process the downloaded data, only the data with a clear IC_{50} value (Standard Relation is "=") and the assay is 'Homo sapiens' are retained. The obtained compounds were then further processed using Pipeline Pilot Software 2017 R2 (BIOVIA, USA), including desalinization, weight removal, inorganic removal,

compound standardization, and removal of duplicate compounds. The IC_{50} value of the final compound was treated with negative logarithm, that is, the pIC_{50} value of each compound was obtained. After obtaining the compounds, 1D and 2D molecular descriptors for each compound were calculated using software PaDEL-Descriptor. The dataset provides 729 molecular descriptors of 1974 anticancer drug candidates (including nAcid, ALogP, ALogp2, etc.) and 729 molecular descriptors of an additional 50 candidates with gold standard for testing, pIC_{50} values and five dichotomies of ADMET for each candidate. These parameters describe the structure and property characteristics of compounds, including physicochemical properties (e.g., molecular weight, LogP, etc.), topological characteristics (e.g., number of hydrogen bond donors, number of hydrogen bond receptors, etc.), and so on.

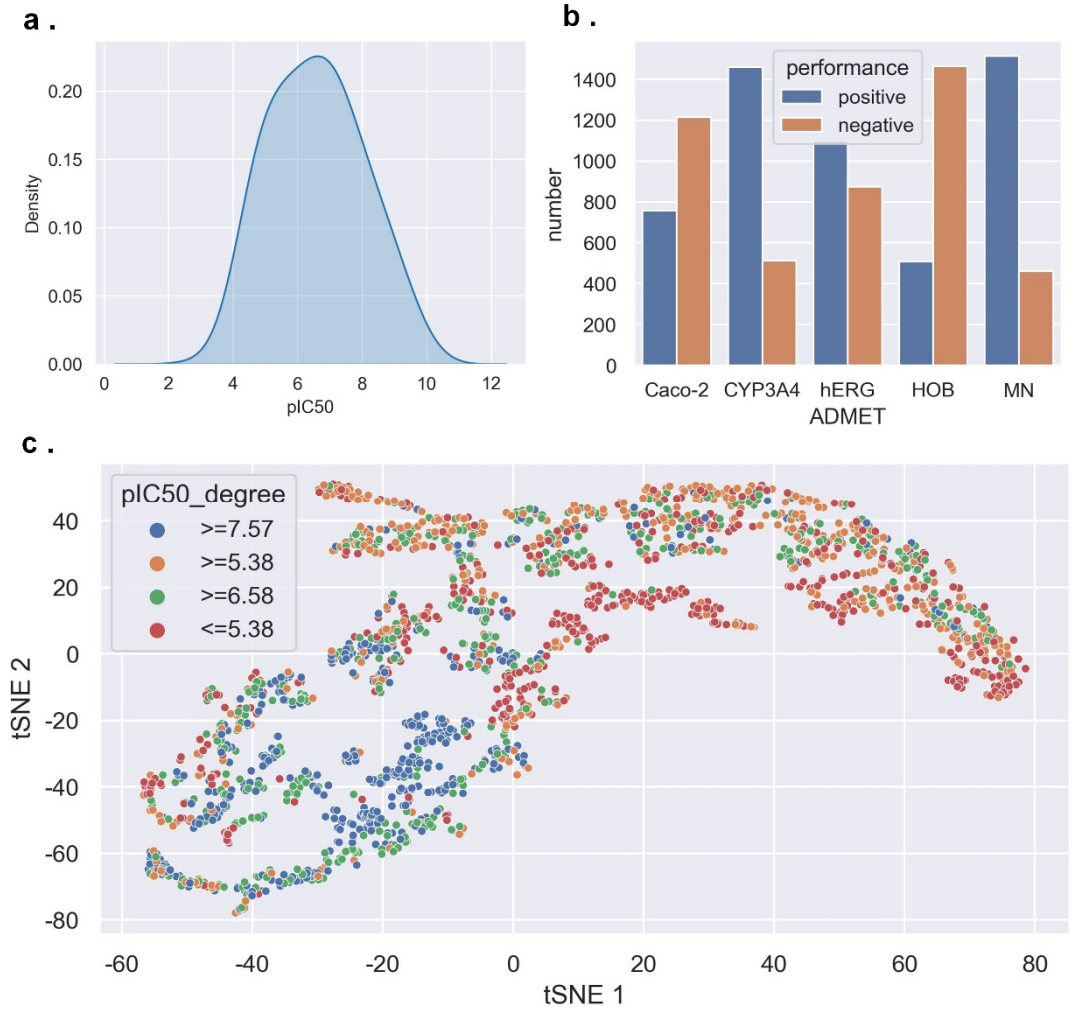

Figure 2. Dataset overview.

The part a and b of Figure 2 indicate the distribution of pIC_{50} (ER α activity) and ADMET, from these two figures, it can be seen that except for hERG, the difference in sample size between the positive and negative classes of other indicators is relatively large, which is called imbalance and a common problem in statistics [36,37]. This imbalance will make the model built biased towards the majority class [33,38], and we must try to overcome this issue, which is one of the reasons we chose the ensemble model for dealing with the task. Besides, to present the chemical structures of the selected

compounds, we used the tSNE method to visualize them, which is better than PCA to characterize the non-linear and complex relationships between the variables. As can be seen in Figure 2(c), the chemical space of the compounds selected for this study covers a wide range.

In dataset, 1215 cases were Caco-24 negative, accounting for 61.55%. There were 513 CYP3A4 negative patients, accounting for about 25.99%; there were 875 hERG negative patients, accounting for 44.32%; there were 1465 HOB positive patients, accounting for about 74.21%; there were 460 MN negative cases, accounting for about 23.30%. For the reliability of ensemble model, the labelled dataset is randomly split into two groups: training subset, validation subset, the ratio of two groups is 8:2 on subjects. During the training phase, the training set is used to optimize the parameters of ensemble model. The validation set is used to validate the model to avoid overfitting. And the additional test dataset with gold standard is used measure the performance of models. It is necessary to clarify that feature selection and subsequent model training are performed in the training set, which is to avoid introducing too many input variables in the prediction task, and the most important variables need to be selected before based on the training set filtering noise.

Before the establishment of bioactivity prediction and classification prediction of ADMET in multi-model ensemble learning, the top 20 molecular descriptors with the most significant influence on ER α， ADMET properties were respectively selected in this paper. Before selecting the molecular descriptors, z-score standardization was performed for them due to their different dimensions:

$$Z_i = \frac{X_i - \mu_i}{\delta_i} \tag{9}$$

3.2. Feature selection

According to the methodology proposed above, the selected 20 input variables for ER α prediction is as Table 1. Table 1 also shows the importance value (normalized) of each molecular descriptor in four sub-models. R means the rank of importance of each molecular descriptor.

Most of the 20 molecular descriptors screened by this model are variables commonly used in organic chemistry to estimate or predict molecular properties. In biochemistry, these indicators also have a theoretical basis for predicting biological activity.

MDEC-23 is the molecular distance between all secondary carbon and tertiary carbon. Carbon atoms in organic matter generally have four atomic bonds to them. The number of hydrogen atoms connected to different carbon atoms can be divided into four types: primary carbon atoms connected to three hydrogen atoms; secondary carbon atoms connected to two hydrogen atoms. A carbon atom connected to one hydrogen atom is called a tertiary carbon, and a carbon atom not connected to a hydrogen atom is called a quaternary carbon. Secondary carbon and tertiary carbon, as “skeleton” carbon atoms in organic compounds, the molecular distance between them may have a great influence on the properties of compounds. MLFER_A is the total solute hydrogen bond acidity, many properties of matter such as boiling point, melting point, viscosity, surface tension and so on are related to it. ATSc index is an autocorrelation descriptor weighted by atomic valence. Lipoaffinity Index is the fat affinity index, which is used to measure the maximum soluble number of compounds in a certain mass of fat and is one of the indicators that can better measure the biological activity of compounds. AMR is the molar refractive index. The molar refraction can be used as a measure of electron polarizability in a molecule. Molar refraction is generally measured by abbe refractometer. Computational methods can also be used to organically combine group contribution methods and topological methods

according to the properties and connectivity of the groups in the molecule, by exploring the quantitative relationship between the molar refraction of alkynes and molecular structure.

Table 1. Contribution of molecular descriptors (normalized) and rank of importance (Top 20).

Molecular Descriptor	MIV		RF		Lasso		XGBoost		Ensemble	
	Im	R	Im	R	Im	R	Im	R	Im	R
MDEC-23	0.0031	44	1.0000	1	1.0000	1	0.0684	23	0.5179	1
BCUTp-1h	0.0008	98	0.0251	32	0.9673	2	0.1075	10	0.2752	2
ATSp4	1.0000	1	0.0050	168	0.0000	21	0.0489	48	0.2635	3
ALogP	0.0003	144	0.0129	62	0.0000	21	1.0000	1	0.2533	4
MLFER_A	0.0000	271	0.0738	11	0.8674	3	0.0521	41	0.2483	5
minsssN	0.0001	228	0.1823	4	0.7615	4	0.0358	89	0.2449	6
ATSp1	0.9280	2	0.0029	237	0.0000	21	0.0261	130	0.2392	7
minHsOH	0.0001	230	0.1295	7	0.7148	6	0.0423	68	0.2217	8
LipoaffinityIndex	0.0006	111	0.2396	2	0.5464	7	0.0391	77	0.2064	9
mindO	0.0022	54	0.0157	54	0.7335	5	0.0423	68	0.1984	10
ATSp5	0.7196	3	0.0077	110	0.0000	21	0.0228	149	0.1875	11
C1SP2	0.0000	355	0.1855	3	0.5450	8	0.0163	184	0.1867	12
ATSc4	0.0000	384	0.0181	46	0.5409	9	0.1270	9	0.1715	13
C3SP2	0.0007	104	0.0159	52	0.5048	10	0.0391	77	0.1401	14
maxsssCH	0.0000	310	0.0077	112	0.3549	11	0.0163	184	0.0947	15
maxHBd	0.0000	317	0.0186	44	0.3111	12	0.0456	59	0.0938	16
minHBint10	0.0007	100	0.0251	33	0.2992	13	0.0456	59	0.0927	17
ATSc3	0.0000	382	0.0520	14	0.0000	21	0.2476	3	0.0749	18
maxHsOH	0.0000	340	0.1680	5	0.1081	17	0.0163	184	0.0731	19
AMR	0.0162	15	0.0050	169	0.0000	21	0.2704	2	0.0729	20

Table 2. The number of variables that are properly selected in each model.

class	variables	MIV	RF	Lasso	XGB	Ensemble
AlogP	AlogP	√			√	√
Carbon types	C3SP2/C1SP2		√	√		√
Autocorrelation (charge)	ATSc1-4		√	√	√	√
Lipoaffinity Index	Lipoaffinity Index		√	√		√
BCUT	BCUTc-1l/1h; BCUTp-1l/1h		√		√	√
XLogP	XLogP					
Molecular distance edge	MDEC-22/23; MDEC-33		√	√	√	√

Continued on next page

class	variables	MIV	RF	Lasso	XGB	Ensemble
Atom type electrotopological state	minssCH2/ minHBa/mindssC			√	√	
Molecular linear free energy	MLFER_A		√	√		√
Crippen logP and MR	CrippenLogP					
Acidic group count	nAcid				√	
sum		**1**	**6**	**6**	**6**	**7**

The variable selection criteria of the final announcement of the competition have 11 categories of variables, and if one of the variables of each category is selected, it will be correct. As shown in Table 2, The ensemble algorithm has a higher performance of feature selection than other methods.

3.3. ER α prediction

The model was trained and run on a laptop computer with i7 6700-HQ CPU + Nvidia GeForce GTX-960m. During training the ensemble model, we use a popular optimizer Adam [39], the detailed parameters of the optimizer are shown in Table 3. By minimizing the binary cross-entropy between label y and estimation $\hat{y}$, the final model could be obtained.

Table 3. Details parameters of the optimizer.

Optimizer	Learning rate	β_1	β_2	ε
Adam	2e-2	0.9	0.999	1e-7

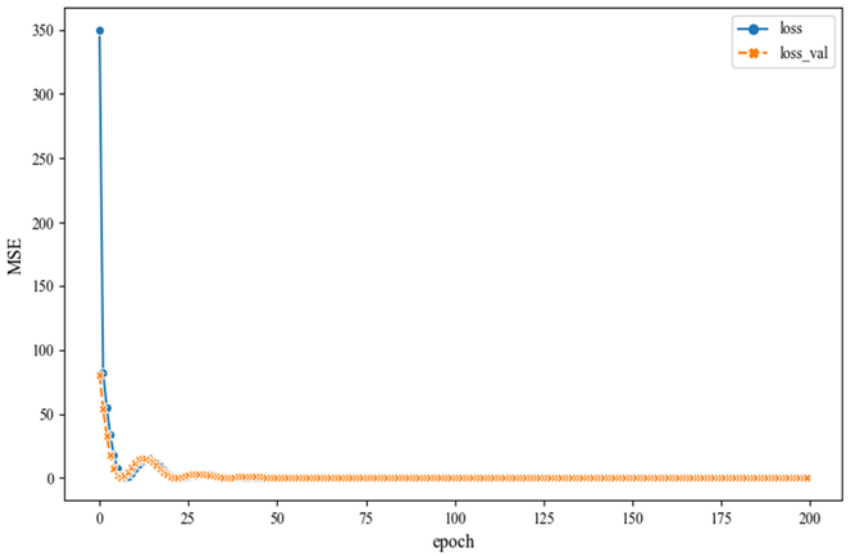

Figure 3. Loss in the training set and validation set during training.

In Table 3, learning rate controls the step length of parameter update, β_1 the exponential decay

rate for the 1st moment estimates, β_2 is exponential decay rate for the 2nd moment estimates, ε is a small constant for numerical stability.

As can be seen from Figure 3, after 200 iterations, loss of the model tends to zero in both training and verification sets, indicating strong generalization ability of the model. In order to compare the prediction effect of the models, XGBoost, random forest, Lasso and compound QSAR model based on ensemble learning are compared in this paper (See Table 4).

Table 4. Comparison of prediction errors (MSE) among different algorithms.

	XGBoost	Lasso	Random Forest	Ensemble
Train_MSE	0.0220	0.9864	0.0855	0.0576
Val_MSE	0.5720	0.9273	0.5172	0.5063
Test_MSE	0.8502	1.2118	0.7994	0.7191

As shown in Table 4, on the validation set, a multi-model ensemble method is slightly less than other models, but as part of methodology said, the advantage of ensemble algorithm is to correct the error between multiple models and dealing with the problem of imbalanced data, plus the accuracy in test set also shows that the ensemble algorithm at the same time weakens the over fitting of sub-models, and also ensure its low MSE on the validation and test set. Therefore, the ensemble algorithm has strong stability and generalization ability in predicting the bioactivity of ER α.

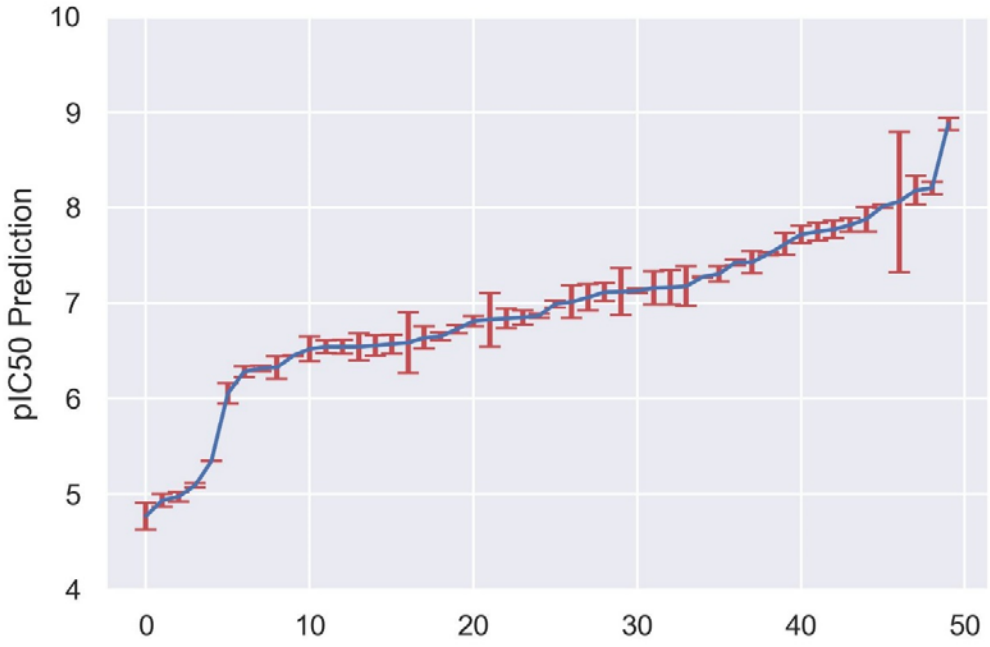

Figure 4. Model performance varied with pIC50.

Furthermore, for assessing the performance of the model under different pIC50, we plotted the error of the model at different prediction values under the test set, and it can also be seen from Figure 4 that the prediction of pIC50 shows a large deviation when the prediction value is greater than 8, and the

error in all other cases is acceptable.

3.4. ADMET prediction

Caco-2, CYP3A4, hERG, HOB and MN are the binary indicators of ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity). According to the algorithm proposed above, we select most important molecular descriptors for Caco-2, CYP3A4, hERG, HOB and MN respectfully, and the ensemble learning models were established according to the selected molecular descriptors. By minimizing the objective function between model outputs and true label, a final prediction model could be obtained (as shown in Figure 1(d)).

To verify idea above more closely, we experimented with the effect of imbalance using stratified sampling, and the results are shown in Table 5. Specifically, we evaluated the model performance in predicting Caco-2 property of different submodels as well as the ensemble model at different imbalance rates $\mathcal{I}$, where $\mathcal{I} = N_{positive}/N_{negative}$, $N_{positive}$ and $N_{negative}$ are the number of positive class samples and negative class samples, respectively. It should be noted that for achieving a comprehensive verification, we downsample the majority class samples to the same number as the minority class samples and then control the number of minority class samples.

Table 5. AUC value of Caco-2 classification under different imbalance rate.

Model	$\mathcal{I} = 1.0$	$\mathcal{I} = 0.7$	$\mathcal{I} = 0.5$	$\mathcal{I} = 0.1$
XGBoost	0.9169	0.9286	0.8571	0.7143
Lasso	0.8571	0.7043	0.7857	0.5000
Random Forest	0.9884	0.9053	0.9169	0.6429
Ensemble	0.9884	0.9286	0.9285	0.7143

Table 6. Accuracy of ADMET model classification.

		XGBoost	Lasso	Random Forest	Ensemble
Caco-2	Training	1.0000	0.8455	1.0000	1.0000
	Validation	0.9030	0.8329	0.9013	0.9038
	Test	0.9400	0.7600	0.9800	0.9800
CYP3A4	Training	1.0000	0.8955	1.0000	0.9994
	Validation	0.9302	0.8810	0.9291	0.9336
	Test	1.0000	0.7600	0.9800	1.0000
hERG	Training	1.0000	0.8271	1.0000	0.9975
	Validation	0.9160	0.8127	0.8987	0.9165
	Test	0.8200	0.7200	0.7000	0.8000
HOB	Training	1.0000	0.7999	1.0000	1.0000
	Validation	0.8405	0.8228	0.8404	0.8405
	Test	0.8400	0.6800	0.7800	0.8600
MN	Training	1.0000	0.8423	1.0000	1.0000
	Validation	0.9419	0.8000	0.9468	0.9468
	Test	0.9400	0.8000	0.9400	0.9400

Table 5 shows that the performance (AUC value) of individual models in the test set deteriorates as the imbalance level increases, and the performance of individual sub-models is not guaranteed to be always optimal at different imbalance levels, but the ensemble of multiple models does enable the error between different sub-models to be "corrected", thus achieving a more stable and superior performance.

Table 6 shows the accuracy in training and validation and test set of five indicators with different models. In most cases, the ensemble learning model proposed in this study performs better than other algorithms, besides, Table 6 also indicates that the ensemble learning model is more advantageous in dealing with unbalanced data. Although the accuracy of our algorithm is not much higher than other algorithms, the over-fitting phenomenon of our algorithm is generally better than that of other models.

With imbalance data label, the accuracy metric is not fully plausible, and we plot the ROC curves and AUC values of the submodel and ensemble model in the five test sets for predicting ADMET properties, and the corresponding results are shown in Figure 5.

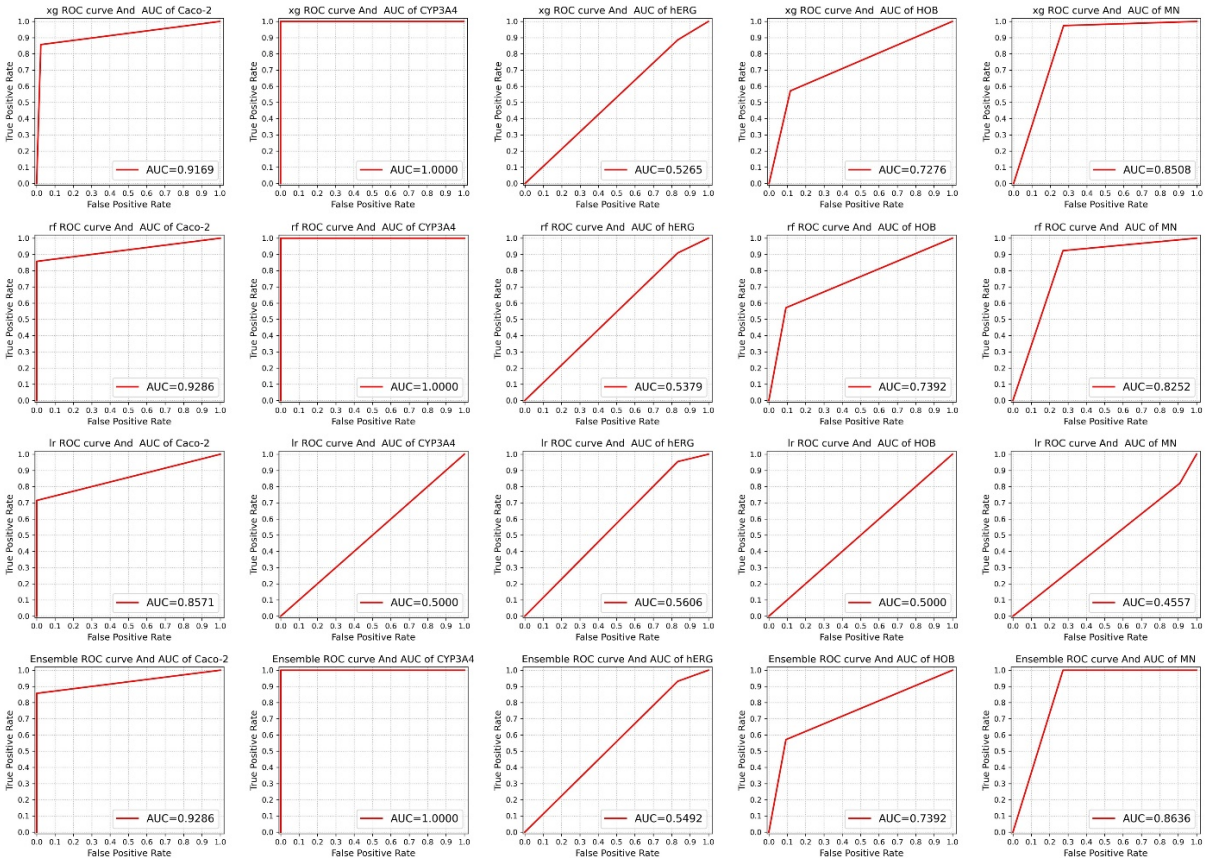

Figure 5. ROC curve and AUC in test dataset.

Combining Tables 5 and 6 and Figure 7, the ensemble model is not always noticeably superior to the other submodels, but the point is that a single submodel does not guarantee that it is the optimal choice in all scenarios, and it may perform worse in some cases, for example, the random forest outperforms in the prediction of Caco-2 and CYP3A4 but poorly in the prediction of hERG and MN, but the ensemble model proposed in this study provides more stable promising even superior predictions.

4. Conclusions

In order to overcome the influence of imbalanced data in the drug candidate screening process and achieve more promising molecular descriptor screening and drug property prediction, we proposed a multi-model ensemble algorithm to select key features from the huge number of molecular descriptors, so that the model could ignore more noise. This innovative ensemble model includes classical models of different types, each with a different perspective on the study data, and finally the different models are fused by an adaptive trainable weight to obtain the final prediction results. With the ensemble of different models, we are capable of finding more plausible molecular descriptors in high-dimensional chemical structures and achieve more promising predictions compared to a single model.

Experimental results based on data of 1974 anti-breast cancer drug candidates containing 729 molecular descriptors and additional 50 candidates with gold standard show that the proposed model can effectively complete the tasks of feature screening, numerical prediction and classification task, and well restrain the occurrence of over-fitting, and the model is more universal in different tasks. Particularly, in terms of feature selection, our proposed approach is significantly able to select a more comprehensive set of important molecular descriptors compared to a single approach. Combining all prediction tasks, a single model is not guaranteed to maintain excellent prediction performance again in all tasks, and it performs well in the Caco-2 and CYP3A4 prediction but poorly in the hERG and MN predictions, while the ensemble model proposed in this study, even if it does not significantly outperform the other submodels, is able to consider all of them together, so that the final output always corrects for the errors from the individual submodels to achieve equal or even superior performance with one of them. This also suggests that the predictions of the ensemble model are more plausible than those of a single model, and the ensemble approach can be considered in other drug screening task.

Although the model proposed in this severe performed well in feature selection and prediction tasks, it still has some limitations. Firstly, this study used dense layers to fuse the outputs of different single models, but this can only represent the linear relationship between feature subsets, and in the future, a multilayer neural network approach can be chosen to fuse different features extracted from single model to represent the nonlinear complex relationship between different feature subsets. Secondly, the focus of this study is biased towards methodological innovation for feature selection and prediction tasks, and the screening of anti-cancer drug candidates needs to rely on more experimental data and validation for further research. Finally, in terms of methodological innovation, we may consider combining with edge computing [40], subspace clustering [41] and other frontier areas [42–47] to expand the algorithms in this study further in the future.

Data availability

The data that support the findings of this study are obtained from ChEMBL database (version: ChEMBL27) (www.ebi.ac.uk/chembl/), which Search with ‘Estrogen receptor alpha’ as the keyword. Click ‘Homo sapiens’ in the ‘Organism’ menu of the filter, and the ChEMBL ID of the Target was confirmed as ‘CHEMBL206’. Download IC_{50} data related to ‘CHEMBL206’ in the database. The processed data can be obtained by contacting the corresponding author.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Serial No. 62062032) and Training Plan for Academic and Technical Leaders of Major Disciplines of Jiangxi Province(20204BCJL23035), Technological Research Project of Education Department in Jiangxi Province (No. GJJ210624) and Special Fund Project for Graduate Innovation in Jiangxi Province (YC2021-S478).

We confirm that all methods were carried out in accordance with relevant guidelines and regulations. and all experimental protocols were approved by the committee of National Post-Graduate Mathematical Contest in Modeling. And informed consent was obtained from all subjects.

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, *Ca-Cancer J. Clin.*, **69** (2019), 7–34. https://doi.org/10.3322/caac.21551
2. C. DeSantis, J. Ma, L. Bryan, A. Jemal, Breast cancer statistics, *Ca-Cancer J. Clin.*, **64** (2014), 52–62. https://doi.org/10.3322/caac.21203
3. G. Giamas, A. Filipović, J. Jacob, W. Messier, H. Zhang, D. Yang, et al., Kinome screening for regulators of the estrogen receptor identifies LMTK3 as a new therapeutic target in breast cancer, *Nat. Med.*, **17** (2011), 715–719. https://doi.org/10.1038/nm.2351
4. Q. Feng, Z. Zhang, M. J. Shea, C. J. Creighton, C. Coarfa, S. G. Hilsenbeck, et al., An epigenomic approach to therapy for tamoxifen-resistant breast cancer, *Cell Res.*, **24** (2014), 809–819. https://doi.org/10.1038/cr.2014.71
5. B. Shaker, K. M. Tran, C. Jung, D. Na, Introduction of advanced methods for structure-based drug discovery, *Curr. Bioinf.*, **16** (2021), 351–363. https://doi.org/10.2174/1574893615999200703113200
6. L. Cai, C. Lu, J. Xu, Y. Meng, P. Wang, X. Fu, et al., Drug repositioning based on the heterogeneous information fusion graph convolutional network, *Briefings Bioinf.*, **22** (2021), bbab319. https://doi.org/10.1093/bib/bbab319
7. A. Ben Brahim, L. Mohamed, Ensemble feature selection for high dimensional data: a new method and a comparative study, *Adv. Data Anal. Classif.*, **12** (2018), 937–952. https://doi.org/10.1007/s11634-017-0285-y
8. L. Meng, N. Masuda, Epidemic dynamics on metapopulation networks with node2vec mobility, *J. Theor. Biol.*, **534** (2022), 110960. https://doi.org/10.1016/j.jtbi.2021.110960
9. D. H. Le, D. Nguyen Ngoc, Drug repositioning by integrating known disease-gene and drug-target associations in a semi-supervised learning model, *Acta Biotheor.*, **66** (2018), 315–331. https://doi.org/10.1007/s10441-018-9325-z
10. R. Su, J. Hu, Q. Zou, B. Manavalan, L. Wei, Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools, *Briefings Bioinf.*, **21** (2020), 408–420. https://doi.org/10.1093/bib/bby124

11. Y. Yang, L. Chen, Identification of drug-disease associations by using multiple drug and disease networks, *Curr. Bioinf.*, **17** (2022), 48–59. https://doi.org/10.2174/1574893616666210825115406
12. Y. Saeys, A. Thomas, Y. Van de Peer, Robust feature selection using ensemble feature selection techniques, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (2008), 313–325. https://doi.org/10.1007/978-3-540-87481-2_21
13. B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: Homogeneous and heterogeneous approaches, *Knowledge-Based Syst.*, **118** (2017), 124–139. https://doi.org/10.1016/j.knosys.2016.11.017
14. S. Zhang, Y. Chen, W. Zhang, R. Feng, A novel ensemble deep learning model with dynamic error correction and multi-objective ensemble pruning for time series forecasting, *Inf. Sci.*, **544** (2021), 427–445. https://doi.org/10.1016/j.ins.2020.08.053
15. H. Liu, Z. Duan, F. Han, Y. Li, Big multi-step wind speed forecasting model based on secondary decomposition, ensemble method and error correction algorithm, *Energy Convers. Manage.*, **156** (2018), 525–541. https://doi.org/10.1016/j.enconman.2017.11.049
16. Z. Zhang, B. Krawczyk, S. Garcìa, A. Rosales-Pérez, F. Herrera, Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data, *Knowledge-Based Syst.*, **106** (2016), 251–263. https://doi.org/10.1016/j.knosys.2016.05.048
17. H. Guo, Y. Li, Y. Li, X. Liu, J. Li, BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification, *Eng. Appl. Artif. Intell.*, **49** (2016), 176–193. https://doi.org/10.1016/j.engappai.2015.09.011
18. A. K. Sharma, R. Srivastava, Protein secondary structure prediction using character bi-gram embedding and bi-LSTM, *Curr. Bioinf.*, **16** (2021), 333–338. https://doi.org/10.2174/1574893615999200601122840
19. F. Weng, H. Zhang, C. Yang, Volatility forecasting of crude oil futures based on a genetic algorithm regularization online extreme learning machine with a forgetting factor: The role of news during the COVID-19 pandemic, *Resour. Policy*, **73** (2021), 102148. https://doi.org/10.1016/j.resourpol.2021.102148
20. Y. Xu, Y. Ma, Z. Zhu, J. Li, T. Lu, Construct comprehensive indicators through a signal extraction approach for predicting housing price crises, *PloS One*, **17** (2022), e0272213. https://doi.org/10.1371/journal.pone.0272213
21. F. Weng, J. Zhu, C. Yang, W. Gao, H. Zhang, Analysis of financial pressure impacts on the health care industry with an explainable machine learning method: China versus the USA, *Expert Syst. Appl.*, **210** (2022), 118482. https://doi.org/10.1016/j.eswa.2022.118482
22. R. Polikar, Ensemble learning, in *Ensemble Machine Learning*, Springer, Boston, MA, (2012), 1–34. https://doi.org/10.1007/978-1-4419-9326-7_1
23. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785
24. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324
25. P. Bühlmann, S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media, 2011. https://doi.org/10.1007/978-3-642-20192-9

26. L. Huang, S. Chen, Z. Ling, Y. Cui, Q. Wang, Non-invasive load identification based on LSTM-BP neural network, *Energy Rep.*, **7** (2021), 485–492. https://doi.org/10.1016/j.egyr.2021.01.040
27. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE*, **86** (1998), 2278–2324. https://doi.org/10.1109/5.726791
28. H. Altun, A. Bilgil, B. C. Fidan, Treatment of multi-dimensional data to enhance neural network estimators in regression problems, *Expert Syst. Appl.*, **32** (2007), 599–605. https://doi.org/10.1016/j.eswa.2006.01.054
29. D. E. Rumelhart, E. H. Geoffrey, R. J. Williams, Learning representations by back-propagating errors, *Nature*, **323** (1986), 533–536. https://doi.org/10.1038/323533a0
30. Y. Nakamura, O. Hasegawa, Nonparametric density estimation based on self-organizing incremental neural network for large noisy data, *IEEE Trans. Neural Networks Learn. Syst.*, **28** (2016), 8–17. https://doi.org/10.1109/TNNLS.2015.2489225
31. W. Sun, Q. Gao, Exploration of energy saving potential in China power industry based on Adaboost back propagation neural network, *J. Cleaner Prod.*, **217** (2019), 257–266. https://doi.org/10.1016/j.jclepro.2019.01.205
32. C. Yan, T. Zhang, Y. Sun, H. Tang, H. Li, A hybrid variable selection method based on wavelet transform and mean impact value for calorific value determination of coal using laser-induced breakdown spectroscopy and kernel extreme learning machine, *Spectrochim. Acta, Part B*, **154** (2019), 75–81. https://doi.org/10.1016/j.sab.2019.02.007
33. N. M. Nasrabadi, Pattern recognition and machine learning, *J. Electron. Imaging*, **16** (2007), 049901. https://doi.org/10.1117/1.2819119
34. P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, T. Lasser, FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification, *Med. Image Anal.*, **76** (2022), 102307. https://doi.org/10.1016/j.media.2021.102307
35. F. Weng, Y. Chen, Z. Wang, M. Hou, J. Luo, Z. Tian, Gold price forecasting research based on an improved online extreme learning machine algorithm, *J. Ambient Intell. Hum. Comput.*, **11** (2020), 4101–4111. https://doi.org/10.1007/s12652-020-01682-z
36. K. Zhang, S. Zhang, Y. Song, L. Cai, B. Hu, Double decoupled network for imbalanced obstetric intelligent diagnosis, *Math. Biosci. Eng.*, **19** (2022), 10006–10021. https://doi.org/10.3934/mbe.2022467
37. J. Wang, Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques, *Math. Biosci. Eng.*, **19** (2022), 10407–10423. https://doi.org/10.3934/mbe.2022487
38. C. Wei, K. Sohn, C. Mellina, A. Yuille, F. Yang, Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 10857–10866.
39. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, preprint, arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980
40. P. Wang, K. Li, B. Xiao, K. Li, Multi-objective optimization for joint task offloading, power assignment, and resource allocation in mobile edge computing, *IEEE Internet Things J.*, **9** (2021), 11737–11748. https://doi.org/10.1109/JIOT.2021.3132080
41. R. Zheng, M. Li, Z. Liang, F. Wu, Y. Pan, J. Wang, SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation, *Bioinformatics*, **35** (2019), 3642–3650. https://doi.org/10.1093/bioinformatics/btz139

42. P. Wang, W. Zhu, B. Liao, L. Cai, L. Peng, J. Yang, Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity, *Front. Microbiol.*, **9** (2018), 2500. https://doi.org/10.3389/fmicb.2018.02500
43. Z. Dimitris, Healthcare access as an important element for the EU's socioeconomic development: Greece's residents' opinions during the COVID-19 pandemic, *Natl. Account. Rev.*, **4** (2022), 362–377. https://doi.org/10.3934/NAR.2022020
44. F. Corradin, M. Billio, R. Casarin, Forecasting economic indicators with robust factor models, *Natl. Account. Rev.*, **4** (2022), 167–190. https://doi.org/10.3934/NAR.2022010
45. D. Panarello, G. Tassinari, The consequences of COVID-19 on older adults: evidence from the SHARE Corona Survey, *Natl. Account. Rev.*, **4** (2022), 56–73. https://doi.org/10.3934/NAR.2022004
46. Z. Li, H. Chen, B. Mo, Can digital finance promote urban innovation? Evidence from China, *Borsa Istanbul Rev.*, **2022** (2022). https://doi.org/10.1016/j.bir.2022.10.006
47. Y. Liu, P. Failler, Y. Ding, Enterprise financialization and technological innovation: Mechanism and heterogeneity, *PLoS One*, **17** (2022), e0275461. https://doi.org/10.1371/journal.pone.0275461

AIMS Press