Quantitative and Qualitative Prediction Method of Anti-breast Cancer Drug Candidates Integrating Machine Learning and Data-driven Perspective

Chao Min, Ping Wu, *Member*, *IAENG*, Yihua Zhong, *Member*, *IAENG*, Yuan Li, *Member*, *IAENG*, and Ji Li, *Member*, *IAENG*

***Abstract*—Drug candidate research plays an essential role in the current breast cancer treatment, such as the quantitative and qualitative prediction problem for anti-breast cancer drug candidates. Aiming at this crucial problem, a quantitative and qualitative prediction method is proposed for anti-breast cancer drug candidates, integrating machine learning and data-driven perspective. Firstly, a drug molecular descriptor selecting method based on random forest (RF) is established to extract the important information in high-dimensional drug molecular descriptors. It is examined through the correlation analysis and variable representative analysis. After processing the IC_{50} outliers, we convert the IC_{50} value to pIC_{50} value. By adopting pIC_{50} as the ERα bioactivity prediction target, a quantitative prediction method of ERα bioactivity is further built based on machine learning (ML). Then, a SMOTE-based data processing method of unbalanced ADMET classes is developed to balance the imbalance categories in five different ADMET properties. A comprehensive evaluation index score (CEIS) is constructed by combining three different indicators. Both ML and CEIS are further integrated to establish a qualitative prediction method for ADMET. Finally, the feasibility, rationality and effectivity of our method are validated through the relevant experiments. The results show that our method can be useful in promoting the development of anti-breast cancer drug candidates.**

***Index Terms*—Drug candidate, Quantitative and qualitative prediction, Machine learning, Imbalanced data, CEIS**

I. Introduction

Breast cancer stands as one of the most widespread malignancies affecting women health. Its incidence is increasing every year. In [1], 60-80% of individuals diagnosed with breast cancer present with estrogen receptor (ER)-positive tumor phenotypes. [2] have demonstrated a mechanistic association between estrogen receptor α subtype (ERα) activation and mammary carcinogenesis. As a result, therapeutic strategies for mammary carcinoma continue to prioritize ERα due to its central role in tumor progression. The compounds demonstrating ERα antagonism show promising potential as anti-neoplastic leads in mammary carcinoma management. Currently, the potential active compounds are usually selected by predicting the ERα bioactivity. Then, the drug candidates can be further evaluated and identified. For the breast cancer treatment, its drug candidates also need to have some great medicinal kinetics properties and biological safety, such as Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) [3]. So, it is also essential for drug candidates to predict their ADMET properties.

For saving the related time and cost in drug research and development, the current methods focus on establishing the bioactivity prediction model [4] to screen the potential active compounds. Specifically, the data is collected for various compounds targeting pathology-associated sites and their bioactivity, such as ERα and its related bioactivity. Then, a succession of molecular structure features is considered as stand-alone variables. The compound bioactivity metrics can serve as the response variables in establishing quantitative structure-activity relationship (QSAR) modeling, such as the values reflecting the biological activities of such compounds. Besides, the ADMET property also needs to be effectively predicted for determining whether the active compound makes ADMET property better or worse, such as the above five ADMET properties. Predicting the drug bioactivity and property can promote the drug candidate development of anti-breast cancer. Thus, a growing number of scholars have conducted the related research on predicting biological activity and ADMET property.

In the early stages, Niculescu-Duvaz et al [5] formulated a QSAR relationship for predicting the biological activity. It can also illustrate the stereochemical features about cyclic moiety and uninterrupted conjugation in polienic chain. Through the integration of E-state calculations and physicochemical profiling, [8] delineated molecular features governing ERα subtype selectivity. Zekri et al [9] developed an effective QSAR model about ERα positive antagonists. By using the molecular docking, they also investigated the binding characteristics between ERα and antagonists. In [10], a QSAR model is built by the bioactivity data and distribution with different structural features, such as quantitative molecular descriptors. These molecular structural features are further employed to predict the ADMET property. Then, Guan et al [11] formulated an ADMET-score (namely scoring function) for the compound medicinal-likeness. In

Manuscript received February 2, 2024; revised February 27, 2025.

Chao Min is a professor in School of Science, Southwest Petroleum University, Chengdu 610500, China (e-mail: minchao@swpu.edu.cn).

Ping Wu is a Ph.D. student in School of Science, Southwest Petroleum University, Chengdu 610500, China (Corresponding Author, e-mail: wuping2021314@163.com).

Yihua Zhong is a professor in School of Science, Southwest Petroleum University, Chengdu 610500, China (e-mail: zhongyihua@swpu.edu.cn).

Yuan Li is a master's graduate from School of Science, Southwest Petroleum University, Chengdu 610500, China (e-mail: 1431326542@qq.com).

Ji Li is a master's graduate from School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China (e-mail: 1023634878@qq.com).

general, some relevant research [12]-[15] on drug biological activity and ADMET property has primarily been conducted from a traditional perspective. They concentrate on biochemical experiments and molecular structures.

Nevertheless, significant room for enhancement exists in the process of developing drug candidates, given the dramatic increase in the number of available drugs. For example, it can be integrated with ML to accelerate the key stages of research and development [16], thereby reducing costs and saving time. Because ML has a better ability to process and analyze the high-dimensional data. Currently, the accelerated evolution of ML algorithms has introduced a novel computational paradigm to study drug activity and ADMET properties. For instance, Beheshti et al [17] selected some important descriptors with the aid of genetic algorithm (GA). QSAR modeling was also performed to predict the drug antimalarial activity by multiple linear regression (MLR). For predicting the biological activity, Anter et al [18] adopted the extreme learning machine (ELM) to constructed a new QSAR model. Based on ML methods, the allosteric drug bioactivity prediction [19] was further studied from molecular dynamics. According to ensemble learning (EL), Shi et al [20] predicted the ADMET property by two-level stacking algorithm (TLSA). Feinberg et al [21] explicitly represented each molecule as a graph to identify the features, which were the most pertinent to each chemical task. The graph convolutional neural networks (GCNN) were utilized to predict the ADMET property. Then, a quantitative regression model [22] was developed by the graph attention network and GCNN. [22] also provided an optimization model and model explanation with the gradient-weighted class activation map. Furthermore, some other ML models and methods can be further integrated into the research on drug bioactivity and ADMET properties, such as multilayer perceptron (MLP) [23], back propagation neural network (BPNN)[24], adaptive boosting(AdaBoost) [25], light gradient boosting machine (LGBM) [25], extreme gradient boosting (XGBoost) [26], and deep learning (DL) [27]-[29], etc.

In summary, ML has been successfully integrated into the drug candidate research on the basis of traditional methods, such as QSAR. Many examples show that this can reduce the experimental cost and time, effectively screen potential active compounds, and improve the pharmaceutical efficiency. Although they have achieved some good results, there are still some aspects that need to be further improved and optimized. For example, both qualitative and quantitative dimensional studies need to be fully considered for drug candidates. The unbalanced category data processing is also important for different drug candidate categories. Motivated by these observations, this paper is committed to establishing a quantitative and qualitative prediction method of anti-breast cancer drug candidates integrating data-driven perspective and ML. The main contributions and innovations can be concisely stated as below:

(1) A drug molecular descriptor selecting method based on RF is developed to extract the important information from high-dimensional drug molecular descriptors.

(2) From a quantitative perspective, a quantitative prediction method of ERα bioactivity based on ML is proposed for drug bioactivity prediction.

(3) For unbalanced ADMET classes, a SMOTE-based data processing method is also developed to balance the unbalanced classes in different ADMET properties.

(4) CEIS is constructed and a qualitative prediction method for ADMET is established by integrating the CEIS and ML. It aims to predict the ADMET property in drug candidates from a qualitative perspective.

The rest of this paper is principally structured in the subsequent sections. Section. II develops a drug molecular descriptor selecting method based on RF, and introduces its whole process. Based on ML, Section. III advances a bioactivity quantitative prediction method and gives the related solution steps. Section. IV establishes a qualitative prediction method of ADMET based on CEIS and ML. It also designs the specific solution steps. In Section. V, some relevant experiments are implemented to demonstrate the proposed method. Finally, Section. VI condenses the key tasks in this paper and furnishes some prospective work orientations.

II. Drug Molecular Descriptor Selecting Method Based on RF

In the existing drug candidate dataset, there are 729 molecular descriptors with compound information. For some molecular descriptors, the compound values are filled with 0. Hence, they are invalid and should be eliminated. The 225 molecular descriptors are eliminated, and the remaining number of molecular descriptors is 504. Due to the redundant information, the number of zeros exceeds 95% in some molecular descriptors. They have little effect on the ERα biological activity, and also should be eliminated. The other molecular descriptors are further retained as 403 in ERα bioactivity data.

Aiming at removing the impact of data dimension and value range, a minimum-maximum normalization method is utilized to perform a linear transformation on the original data. Its formula is as follows:

$$x_{norm} = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{1}$$

Among them, x is the known data on the molecular descriptor corresponding to each compound, x_{norm} is the normalized value, $\min(x_i)$ is the minimum value in molecular descriptor data, $\max(x_i)$ is the maximum value in molecular descriptor data.

Molecular descriptors are essential for QSAR research. The reliability and validity of QSAR models largely depend on the accuracy of the screened descriptors. After the above data processing, the 403 remaining molecular descriptors are clearly characterized by excessive feature dimensionality and redundancy. So, it is necessary to obtain the most important molecular descriptors for ERα bioactive prediction. Inspired by [30] and data-driven perspective, this paper establishes a molecular descriptor selecting method based on RF.

In the proposed method, it aims to employ each decision tree (DT) within RF to calculate the contribution of each molecular descriptor. Then, we take their average values, and compare the average contribution between different features for molecular descriptor selection. The specific operation is as follows:

(1) Adopt the bootstrap method to randomly select some

samples in the processed dataset as a training dataset.

(2) Select d features (d<403) randomly from the 403 molecular descriptors to form a new feature set.

(3) Divide the sample set based on these d features, and find the best dividing feature by the Gini impurity.

(4) Utilize the feature set in (2) to construct the DT, and further form into the RF. Then, the importance degree of molecular descriptors is calculated and sorted. The specific formula is:

$$VIM_i^{(Gini)} = \frac{1}{n_{tree}} \sum_{v \in S_{x_i}} (GI(v) - w_L GI(v^L) - w_R GI(v^R)) \tag{2}$$

$$GI(v) = \sum_{c=1}^{K} p_c^v \left(1 - p_c^v\right) \tag{3}$$

where $VIM_i^{(Gini)}$ is the importance degree of molecular descriptors. $GI(v)$ is Gini information gain for molecular descriptor x_i at node v, v^L, v^R denote the left and right nodes of v, w_L, w_R are the sample proportion assigned to the left and right nodes, p_c^v denotes the sample proportion for category c in v, S_{x_i} is the partition set for all nodes of x_i in n_{tree} trees.

Based on the above steps, we can obtain the molecular descriptors demonstrating significant bioactivity relevance. The importance degree of preferred molecular descriptors is shown in Figure 1.

In Figure 1, the importance degree of MDEC-23 is particularly notable and prominent. However, the importance degree is relatively close for other neighboring molecular descriptors, such as C1SP2 and LipoaffinityIndex. Moreover, the top 20 molecular descriptors are shown in Table I, involving 10 categories.

TABLE I
STATISTICAL TABLE OF MOLECULAR DESCRIPTOR TYPES

Descriptor type	NUMBER	Descriptor
Atom count	1	nC
Autocorrelation (charge)	1	ATSc3
BCUT	1	BCUTw-11
Carbon types	1	C1SP2
Chi cluster	1	VC-5
Hbond acceptor count	1	nHBAcc
Molecular distance edge	2	MDEC-12, MDEC-23
Topological polar surface area	1	TopoPSA
Molecular linear free energy relation	1	MLFER_A
Atom type electrotopological state	10	LipoaffinityIndex,maxssO, ndssC, etc

Among these categories, the most prominent category is Atom type electrotopological state, and it accounts for 50%. Specifically, it contains 10 molecular descriptors, including LipoaffinityIndex, maxssO, minsssN, maxHsOH, minsOH, minHsOH, minHBint5, SHsOH, SHBint10, and ndssC. Besides, the largest contribution (MDEC-23) belongs to the category of Molecular distance edge.

In order to further prevent the overfitting and verify the rationality of the results in Figure 1, the correlation analysis and variable representative analysis are performed on the selected molecular descriptors. For the correlation analysis, we adopt the Spearman rank correlation coefficient, which is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)}{n\left(n^2 - 1\right)} \tag{4}$$

where R_i denotes a molecular descriptor rank, Q_i is the rank of another molecular descriptor, and $R_i - Q_i$ denotes the rank difference.

According to formula (4), we calculate the Spearman rank correlation coefficient between the selected molecular descriptors. The specific results are shown in Figure 2. By the comparative analysis, the correlation coefficients are small between these molecular descriptors, and so their direct relevance is very low. This can justify the proposed selecting method.

With regard to the variable representative analysis, the entropy weight method (EWM) is adopted to calculate the information entropy and cumulative weight for molecular descriptors. The related formulas are as follows:

$$H_j = -\ln^{-1}(n)\sum_{i=1}^{n}(Y_{ij} / \sum_{j=1}^{m} Y_{ij})\log(Y_{ij} / \sum_{j=1}^{m} Y_{ij}) \tag{5}$$

$$w_j = \frac{1 - H_j}{n - \sum_{j=1}^{m} H_j} (j = 1, \cdots, n) \tag{6}$$

where H_j is the information entropy in the jth molecular descriptor, and n denotes the sample number. Y_{ij} is the bioactivity data in the ith compound at the jth molecular descriptor, and w_j denotes the information entropy weight.

According to the formulas (5) to (6), H_j and w_j are first calculated, and then we further calculate the cumulative weight for information entropy. Figure 3 and 4 present the pertinent outcomes.

In Figure 4, the information entropy gradually increases for the selected molecular descriptors. Then, the cumulative weight is approximately 31.9% for information entropy. This means that the information entropy of the selected molecular descriptors can account for 31.9% in the total information entropy of all 403 variables. But, their average level is only 4.96%. It can indicate that the selected top 20 molecular descriptors have a good representative level. This can further validate the rationality of the proposed selecting method.

III. QUANTITATIVE PREDICTION METHOD OF BIOLOGICAL ACTIVITY BASED ON ML

A. The quantitative prediction model of ERα biological activity based on ML

According to the exiting bioactivity dataset, it has 1974 compounds and the top 20 molecular descriptors against ERα for breast cancer treatment target. Then, the IC_{50} value can characterize the biological activity. It is effective in inhibiting the ERα activity. When the IC_{50} value is smaller, this indicates a higher bioactivity level and a more effective ERα inhibition. Meanwhile, it can be discovered that some IC_{50} values are abnormally large in Figure 5(a). To address this issue, the previous data has been re-examined after removing the outliers, as shown in Figure 5(b).

The compounds associated with the removed outliers are given in Table II.

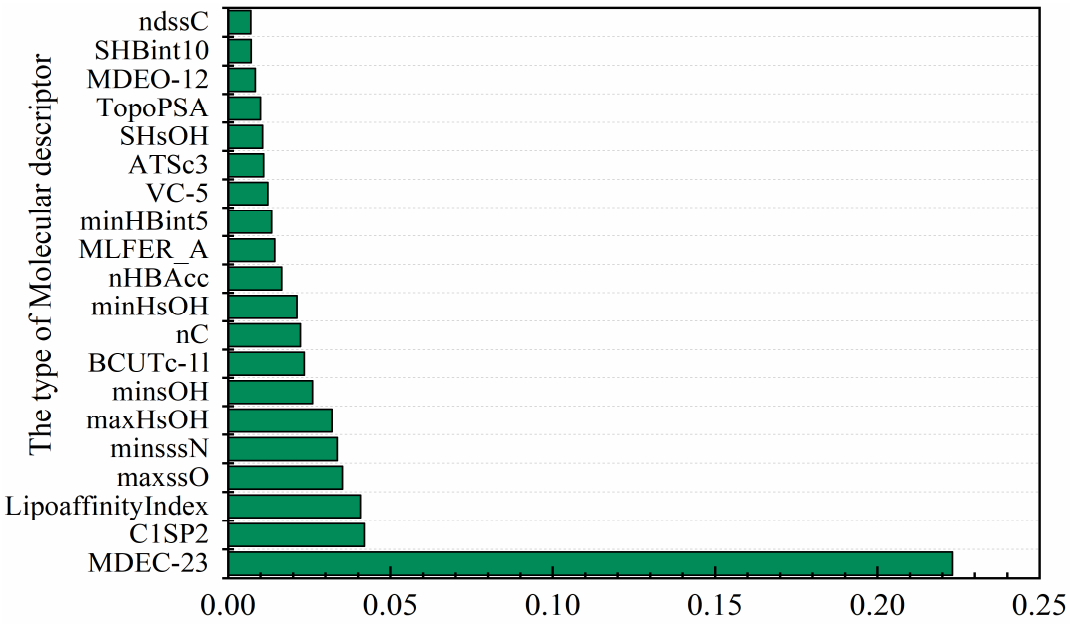


Fig. 1. The importance degree of preferred molecular descriptors

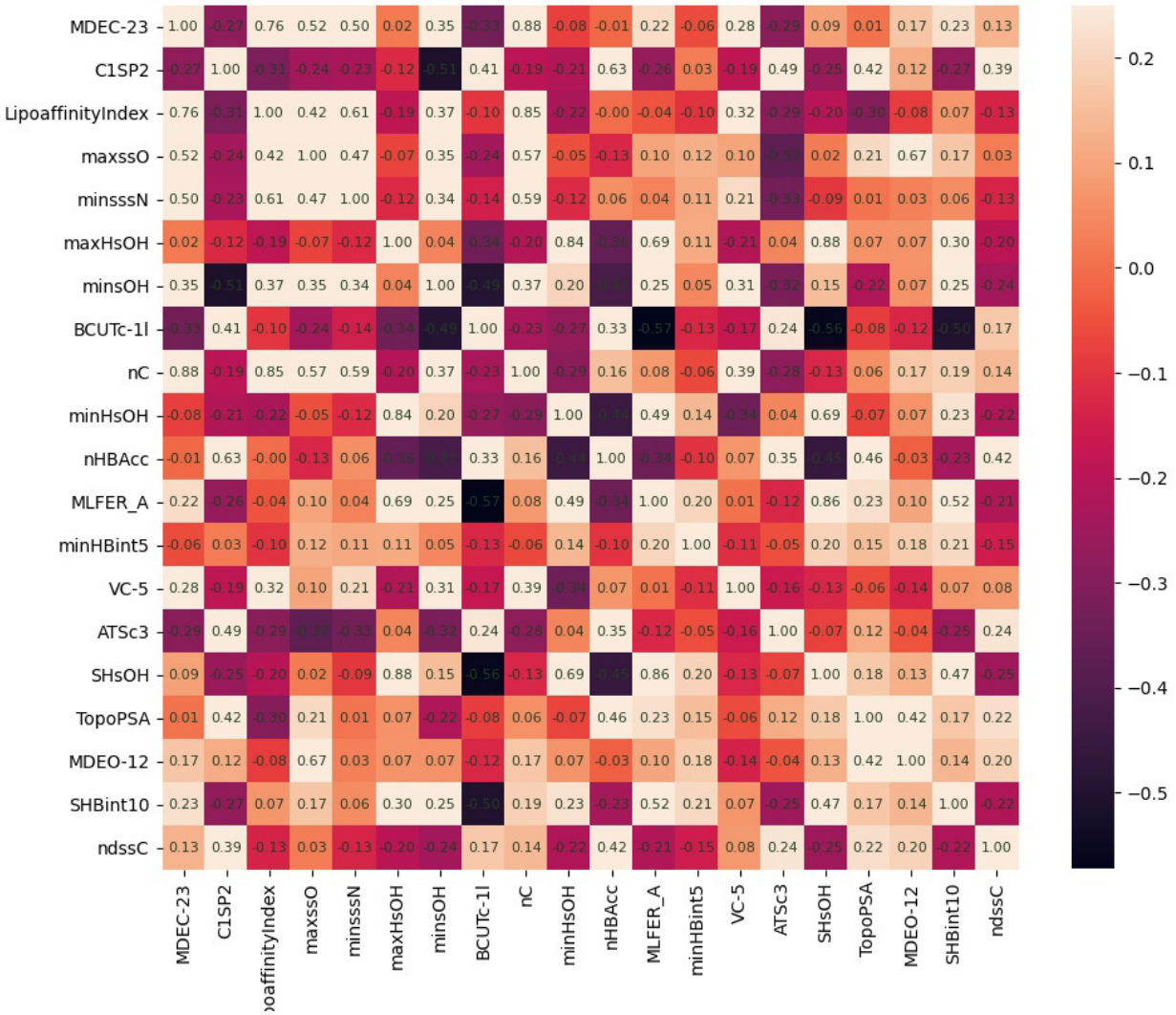


Fig. 2. The correlation graph of molecular descriptors

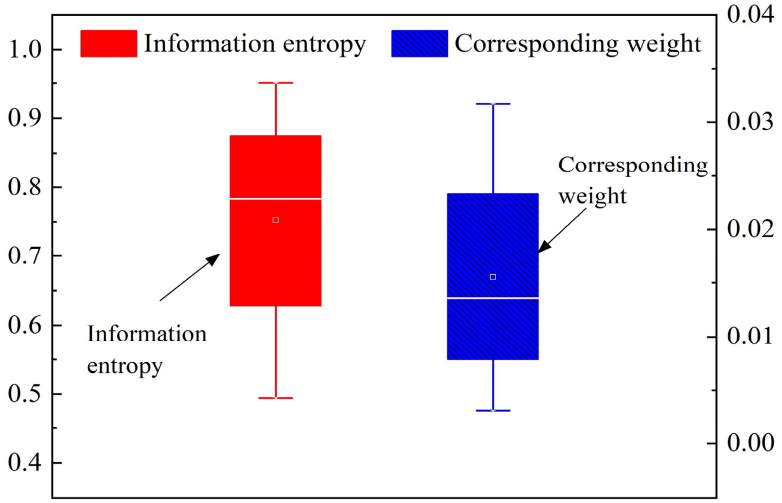


Fig. 3. The distribution of information entropy and corresponding weight

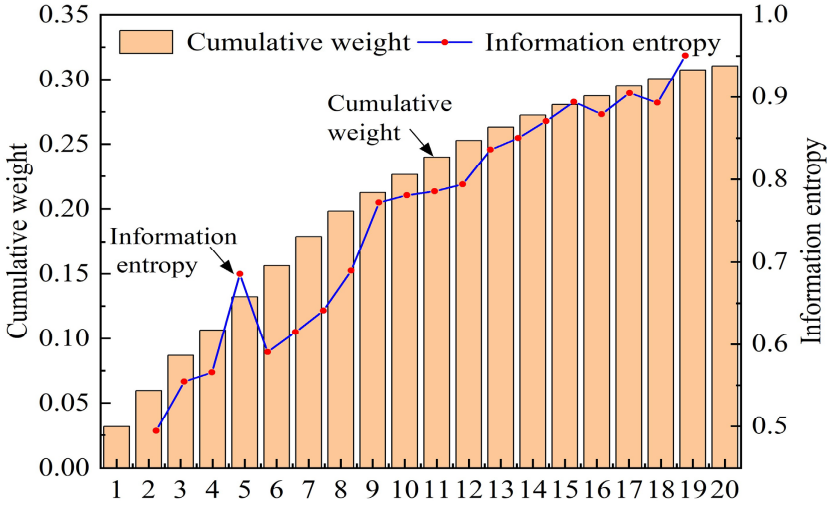


Fig. 4. The cumulative weights of information entropy

TABLE II
STATISTICAL TABLE OF ELIMINATED COMPOUNDS

	SMILES
1	CC(=CCc1c(O)c(O)cc([C@H]2CCc3ccc(O)cc3O2)c1CC=C(C)C)C
2	COc1cc(O)c(\C=C/C(C)(C)O)c2O[C@@H](CC(=O)c12)c3ccccc3
3	COc1cc(O)c(\C=C/C(C)(C)O)c2OC(CC(=O)c12)c3ccc(O)cc3
4	CC(C)[C@]1(O)CC[C@@]2(C)[C@H](O)CCC(=C)[C@@H]2[C@H]1O
5	CC1CC(=O)C2=C(C1)OC3=C(C2c4cc5ccccc5n4Cc6ccccc6)C(=O)CCC3(C)C
6	CC1(C)CC(=O)C2=C(C1)OC3=C(C2c4cc5ccccc5n4Cc6ccccc6)C(=O)CCC3(C)C
7	CC1CC(=O)C2=C(C1)OC3=C(C2c4cc5ccccc5n4Cc6ccccc6)C(=O)CC(C)C3

Combined with [31], the compound bioactivity value against ERα is typically measured by the pIC_{50} value. pIC_{50} is usually positively correlated with biological activity. It can also be regarded as a useful tool in predicting the compound bioactivity against ERα. Meanwhile, the conversion formula between pIC_{50} and IC_{50} is as below:

$$pIC_{50} = 9 - \log_{10} IC_{50} \tag{7}$$

Then, the quantitative prediction model is constructed for ERα biological activity by implementing EL, such as random forest regression (RFR) and gradient boosting regression tree (GBRT).

In RFR, the combination regression of multiple DTs is used to calculate the biological activity as follows:

$$\overline{r_n}(X, D_n) = E_Q\left[r_n(X, Q, D_n)\right] \tag{8}$$

where D_n denotes the total dataset, E_Q is the expectation associated with X, and $X(X_1, X_2, \cdots, X_i, \cdots, X_k, k < n)$ can denote the random parameters of molecular descriptors in D_n. Then, we have:

$$r_n(X, Q, D_n) = \frac{\sum_{i=1}^{n} Y_i 1_{[X_i \in A_n(X,Q)]}}{\sum_{i=1}^{n} 1_{[X_i \in A_n(X,Q)]}} 1_{[E_n(X,Q)]} \tag{9}$$

where Y_i denotes the dependent variable corresponding to X_i in the learning samples, $1_{[\cdot]}$ is the indicator function. If the subscript condition is met, there will be $1_{[\cdot]} = 1$; if not, there is $1_{[\cdot]} = 0$. $A_n(X, Q)$ is a random unit partition with X, and the judgment condition $E_n(X, Q)$ is as below:

$$E_n(X, Q) = \left[\sum_{i=1}^{n} 1_{[X_i \in A_n(X,Q)]} \neq 0\right] \tag{10}$$

By combining with the formula (8) and (9), the bioactivity regression estimation in RFR is:

$$\overline{r_n}(X, D_n) = E_Q\left[\frac{\sum_{i=1}^{n} Y_i 1_{[X_i \in A_n(X,Q)]}}{\sum_{i=1}^{n} 1_{[X_i \in A_n(X,Q)]}} 1_{[E_n(X,Q)]}\right] \tag{11}$$

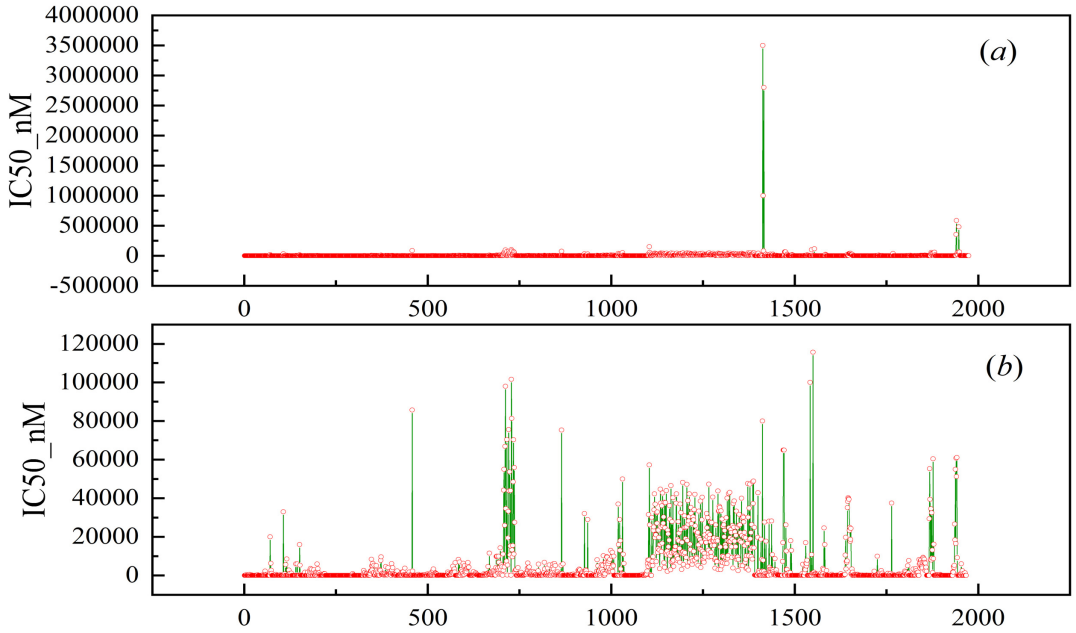


Fig. 5. The comparison of IC_{50} outliers before and after elimination

For GBRT, assume the bioactivity dataset is D, its quantitative bioactivity prediction is denoted as $f(x)$.

Next, we initialize the weak learner as:

$$f_0(x) = \arg\min_c \sum_{i=1}^{n} L(y_i, c) \tag{12}$$

For the iteration $m = 1, 2, \cdots, M$, the negative derivative of samples is calculated as:

$$r_{Mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{13}$$

Then, we use $(x_i, r_{mi}), (i = 1, 2, \cdots, n)$ to obtain the m-th regression tree by fitting a classification and regression tree (CART). $R_{mk, j=1,2,\cdots,J}$ denotes the terminal node partition. J is the terminal counts in regression tree m. Regarding the leaf region, the best bioactivity fitting value can be calculated as:

$$c_{mk} = \arg\min_c \sum_{x_i \in R_{Mk}} L(y_i, f_{M-1}(x_i) + c) \tag{14}$$

Hence, the former learner is further updated as:

$$f(x) = f_{M-1}(x) + f_m(x) = \sum_{m=1}^{M}\sum_{j=1}^{K} c_{mj} I(x \in R_{mj}) \tag{15}$$

B. The solution steps of quantitative prediction model with biological activity

According to the bioactivity quantitative prediction model developed in Section III.A, the detailed solution procedures are structured through the following steps.

Step 1: Further process the IC_{50} outliers in Section III.A based on the processed data in Section II.

Step 2: Randomly shuffle and split the processed data in Step1, and it is partitioned into training (80%) and testing (20%) subsets.

Step 3: Based on the established RFR and GBRT models in Section III.A, pIC_{50} is used as the prediction label. Ten-fold cross-validation is employed to train the models, with optimal parameter selection occurring during the training iterations. The developed model is evaluated on the testing subsets from Step 2. The IC_{50} value is computed by the conversion formula (7) and pIC_{50}.

Step 4: Conduct the model evaluation by calculating the evaluation metrics between the pIC_{50} model-predicted values and its actual values, including calculating the mean absolute error (MAE), the mean absolute percentage error (MAPE), the root mean square error (RMSE), and the goodness of fit (R^2).

IV. Qualitative Prediction Method of ADMET Based on CEIS and ML

A. SMOTE-based data processing method for unbalanced ADMET classes

In the existing biological activity dataset, the five major properties of toxicity or quality will significantly affect the ADMET property. Concretely, the ADMET property mainly needs to consider Caco-2, CYP3A4, human Ether-a-go-go Related Gene (hERG), Human Oral Bioavailability (HOB) and Micronucleus (MN). Thus, they can be effectively classified and predicted to determine whether the active compound makes the ADMET property better or worse. According to the ADMET property data, both 0 and 1 are utilized to denote the category labels. It is converted into the binary classification problem. When the Caco-2 or HoB category labels are 1, it is better for ADMET property. While the CYP3A4, hERG or MN category labels are 0, the ADMET property is better.

Meanwhile, the imbalance categories obviously exist in the ADMET property data. This can significantly affect the construction and performance in predictive models. Inspired by [32] and data-driven perspective, an unbalanced ADMET class data processing method is established by incorporating the synthetic minority oversampling technique (SMOTE). It can reasonably expand the original data and balance the ratio of unbalanced category labels to 1:1. Its operational steps are outlined as below:

(1) Use the idea of k nearest neighbors (KNN) [33],[34] to construct the k nearest neighbors for the ADMET minority categories.

(2) Randomly select the N samples for stochastic linear interpolation among the k nearest neighbors from (1), and the choice of N depends on the final desired equilibrium rate. For

each selected sample point, the subsequent formula (16) is employed to build a new ADMET minority-class sample.

$$x_{new} = x_i + rand(0,1) \times (x_j - x_i), j = 1,2,\cdots,M \quad (16)$$

Where x_j is a sample randomly chosen from the k nearest neighbors, x_i denotes the ADMET minority-class sample, and $rand(0,1)$ denotes the random number between 0 and 1.

(3) Repeat (1) and (2) to iteratively update each of the minority category in five different unbalanced categories for expanding each imbalance category to a ratio close to 1:1.

Based on the above steps, the imbalance categories are processed for five different ADMET properties. The results are given in Table III. Among them, the expanded categories of Caco-2 and HOB are 1. For CYP3A4, hERG and MN, 0 is regarded as their expanded categories.

TABLE III
RESULT TABLE OF UNBALANCED DATA PROCESSING

	Original class ratio	Original dataset size	Current dataset size	Added class
Caco-2	2:1	1974×504	2429×504	1
CYP3A4	1:3	1974×504	3000×504	0
hERG	5:6	1974×504	2192×504	0
HOB	3:1	1974×504	2992×504	1
MN	1:3	1974×504	2894×504	0

B. The qualitative prediction model of ADMET based on CEIS and ML

According to the ADMET property, its processing and analysis, a qualitative prediction model of ADMET property is established by integrating ML and CEIS. Likewise, it is also mainly on the basis of gradient boosting decision tree (GBDT) and random forest classification model (RFC). For GBDT, suppose that the bioactivity dataset is D^*, and the ADMET qualitative prediction is denoted by $f^*(x)$ .Then, a weak learner $F_0(x)$ is initialized as:

$$F_0(x) = \log \frac{P(Y=1|x)}{1-P(Y=1|x)} \quad (17)$$

where $P(Y=1|x)$ is the proportion of $y=1$ in the training dataset. Next, the M^* classification decision trees of ADMET property are constructed by using the priori information. For the iteration $m^* = 1,2,\cdots,M^*$, the negative derivative of cost function corresponding to m^*-th tree is computed:

$$r_{m^*,i} = -\left[\frac{\partial L(y_i, f^*(x_i))}{\partial f^*(x_i)}\right] f^*_{m^*-1}(x) \quad (18)$$

Then, $(x_i, r_{m^*i})(i = 1,2,\cdots,n)$ is used to fit the basic model:

$$f_{m^*}(x) = \sum_{j=1}^{J} c_{m^*j} I(x \in R_{m^*j}) \quad (19)$$

$$c_{m^*j} = \arg\min \sum_{x_i \in R_{m^*j}} \log\left(1 + \exp\left(-y_i\left(f_{m^*-1}(x_i) + c\right)\right)\right) \quad (20)$$

Finally, the above process is repeated, and the final GBDT model is constructed by utilizing the above m^* basic models, which is:

$$f_{M^*}(x) = f_{M^*-1}(x) + f_{m^*}(x) = \sum_{m^*=1}^{M^*}\sum_{j=1}^{J} c_{m^*j} I(x \in R_{m^*j}) \quad (21)$$

For RFC, its basic structure is similar to RFR. The ensemble classifier of RFC can be briefly described as $h(x, \theta_k), k = 1,2,\cdots n$, where x denotes the input bioactivity data, k is the DT number in RFC. θ_k denotes the parameter vector of k-th DT. This is an independently and identically distributed random vector. Meanwhile, it is independently and equally determined by learning on the independent and equally distributed bootstrap sets. The ADMET qualitative predictions for x are determined by voting from each DT in the RFC, and the combined average votes are utilized to determine the corresponding category for ADMET property.

Besides, the CEIS is constructed by combining three different indicators. Moreover, CEIS is also introduced into the qualitative prediction model of ADMET property. It mainly combines the accuracy, precision and recall. When CEIS is larger, there is a better effect for the qualitative prediction of model. The specific formulas are as below:

$$CEIS = Accuracy + Precision + Recall \quad (22)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (23)$$

$$Precision = \frac{TP}{TP + FP} \quad (24)$$

$$Recall = \frac{TP}{TP + FN} \quad (25)$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative in the confusion matrix.

C. The solution steps of qualitative prediction model with ADMET property

According to the qualitative prediction model developed in Section IV.B, the relevant solution processes are devised through the following steps:

Step 1: Use the SMOTE-based data processing method in Section IV.A to deal with the imbalance category in ADMET property data, including the five different ADMET properties: Caco-2, CYP3A4, hERG, HOB, and MN.

Step 2: Randomly disrupt and divide the processed data in Step1, and it is partitioned into an 80% training subset and a 20% testing subset.

Step 3: Train the RFC and GBDT models in Section IV.B, and adopt the grid search to optimize the model parameters for five different ADMET classification targets.

Step 4: Test the optimal model from Step 3 training on the test set for five different ADMET properties, and evaluate the model prediction results by CEIS, involving Accuracy, Precision and Recall.

V. Experimental Analysis and Results

A. The quantitative prediction experiment of ERα biological activity

According to the existing ERα bioactivity dataset, the experiments are implemented by the methods and procedures in Section II and III. The ERα biological activity dataset is randomly disrupted and separated into distinct subsets. Meanwhile, BPNN, generalized regression neural network (GRNN) [35] and support vector machine regression (SVR) [36], [37] are also introduced for comparative experiments. Through the model training and ten-fold cross-validation, the

optimal parameters are obtained and shown in Table IV. All the programs of our experiments are coded in Python.

TABLE IV
TABLE OF OPTIMAL PARAMETERS IN QUANTITATIVE PREDICTION MODELS

Model	Parameter name	Parameter value
RFR	n_estimators	200
	max_depth	12
	max_features	'log2'
GBRT	n_estimators	50
	learning_rate	0.05
	max_depth	4
SVR	Kernel	'rbf'
	n_estimators	100
	learning_rate	0.05
BPNN	leraning_rate	0.1
	Epochs	800
	hide_num	10
GRNN	σ	0.7

Figure 6 shows the relevant prediction results in our experiments. Then, we also adopt four different evaluation indicators to further assess the quantitative prediction results. Based on them, Table V and Figure 7 show the related evaluation results from five different models.

TABLE V
STATISTICAL TABLE OF QUANTITATIVE PREDICTION EVALUATION RESULTS

Model	MAE	MAPE	RMSE	R^2
RFR	0.2107	0.0319	0.3298	0.9428
GBRT	0.2074	0.0331	0.3875	0.9124
GRNN	0.2234	0.0454	0.4545	0.7834
BPNN	0.4002	0.0534	0.5634	0.6434
SVR	0.6845	0.1087	0.8964	0.6046

According to Figure 6, Figure 7 and Table V, it can be discovered that RFR and GBRT have a larger advantage in bioactivity quantitative prediction and generalization ability by comparing the effect of bioactivity prediction.

Furthermore, the test samples are randomly selected to further analyze the quantitative prediction ability of RFR and GBRT for IC_{50}. Figure 8 shows the related prediction results about IC_{50}.

In Figure 8, the IC_{50} predicted values from RFR and GBRT are lower than 200 within a sample range of 0 to 30. In the range of 30 to 40 samples, their IC_{50} predicted values have increased in a certain extent. Then, their IC_{50} predicted values are reduced to less than 150. Thus, the IC_{50} predicted trend is clearly consistent in RFR and GBRT. This can also verify their stability and effectiveness.

In conclusion, these results can validate the rationality and efficacy of the quantitative bioactivity prediction method in Section III. Meanwhile, the molecular descriptors selected method in Section II not only achieves better results for RFR model, but also applies to the GBRT model. This can further validate the reliability of molecular descriptor selection in Section II.

B. The qualitative prediction experiment of ADMET property

According to the existing ADMET property dataset, the experiments are conducted by the methods and procedures in Section II and IV, and Python is also adopted as the primary experiment tool. Then, the ADMET property dataset is also randomly disrupted and separated into distinct subsets. Then, KNN, support vector machine (SVM) [38],[39], DT [40], [41] and logistic regression (LR) [42] are also introduced for comparative experiments. The grid search method is utilized in our model training to determine the optimal parameters, and the accuracy is as an evaluation index for optimal parameter selections.

For RFC, the optimal parameter selection is shown in Figure 9. For other models, their optimal parameters are also obtained through the grid search. They are further given in Table VI. Regarding the five different ADMET properties, CEIS is adopted to evaluate the experiment results from different models. These results are tabulated in Table VII and graphically represented in Figure 10.

According to Table VII and Figure 10, the optimal qualitative prediction model is RF for Caco-2, hERG, and HOB. For CYP3A4 and MN, their best qualitative prediction model is GBDT. Table VIII gives the optimal model for five different ADMET properties. The RFC and GBDT have a significant advantage in the ADMET qualitative prediction. It can also verify the rationality and validity of the qualitative prediction method of ADMET property in Section IV.

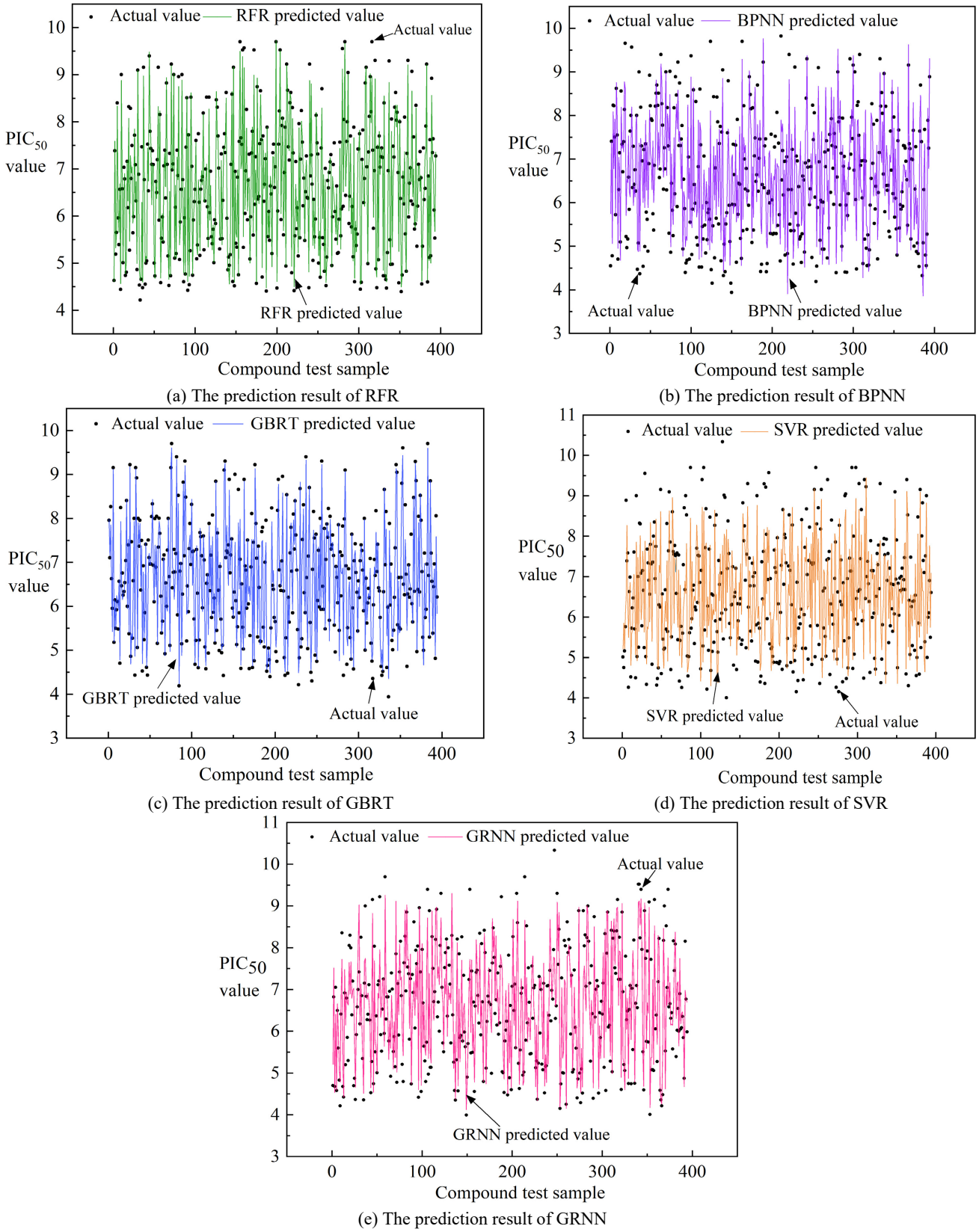
TABLE VI
TABLE OF OPTIMAL PARAMETERS IN QUALITATIVE PREDICTION MODELS

Model	Parameter name	Parameter value
RFC	n_estimators	45
	max_depth	16
	max_features	11
GBDT	n_estimators	200
	learning_rate	0.1
SVM	kernel	'rbf'
	class_weight	'balanced'
KNN	n_neighbors	5
	leaf_size	30
DT	max_depth	10
	min_samples_split	2
	min_samples_leaf	1
LR	max_iter	10
	warm_start	False

TABLE VII
STATISTICAL TABLE OF QUALITATIVE PREDICTION EVALUATION RESULTS

	RFC	SVM	KNN	GBDT	LR	DT
Caco-2	2.810	2.673	2.645	2.800	2.753	2.753
CYP3A4	2.895	2.811	2.807	2.900	2.875	2.875
hERG	2.782	2.645	2.658	2.753	2.679	2.679
HOB	2.774	2.637	2.628	2.769	2.615	2.615
MN	2.922	2.764	2.776	2.964	2.788	2.788

For the five different ADMET properties, the optimal models in Table VIII are further validated by the confusion matrix, as shown in Figure 11.



(a) The prediction result of RFR

(b) The prediction result of BPNN

(c) The prediction result of GBRT

(d) The prediction result of SVR

(e) The prediction result of GRNN

Fig. 6. The quantitative prediction results of ERα biological activity

In Figure 11, the x-axis denotes the predicted values, while the y-axis represents the true values. The yellow diagonal portion denotes the correctly categorized number, and the purple diagonal denotes the incorrectly categorized number. It is evident that the qualitative prediction effects on five different properties are significantly improved by the optimal models in Table VIII. The correctly classified number is much larger than the number of incorrectly classified samples. It can indicate that the optimal qualitative prediction models in Table VIII possess excellent qualitative prediction ability and generalization ability. This can further validate the reasonability and effectivity of the qualitative prediction method established in Section IV.

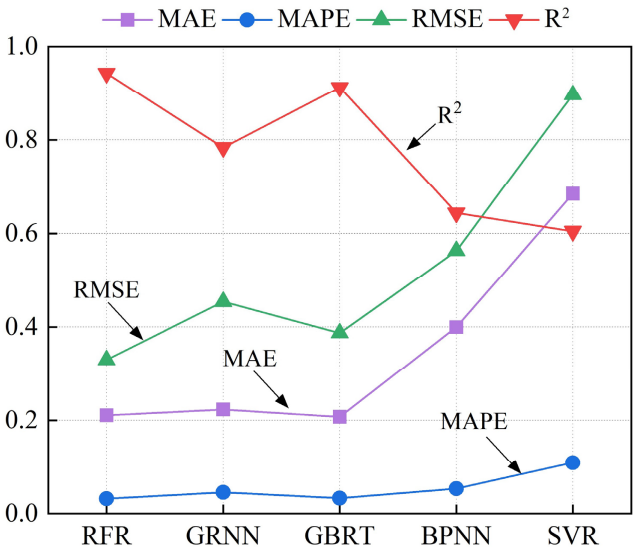


Fig. 7. The four evaluation results in five different models

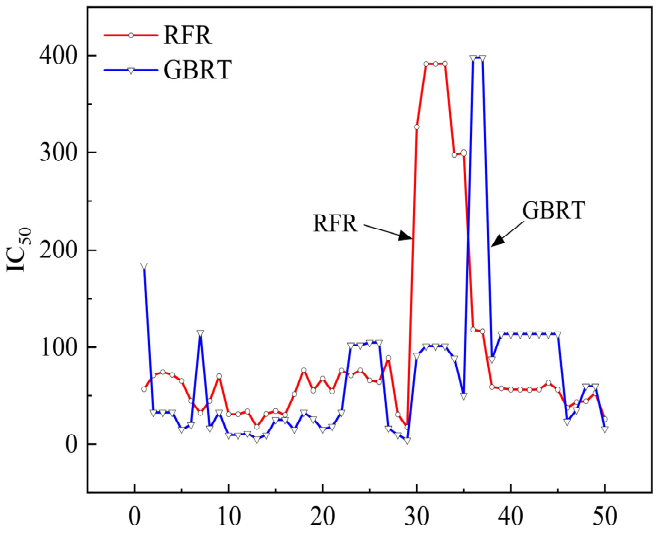


Fig. 8. The IC_{50} prediction results for RFR and GBRT

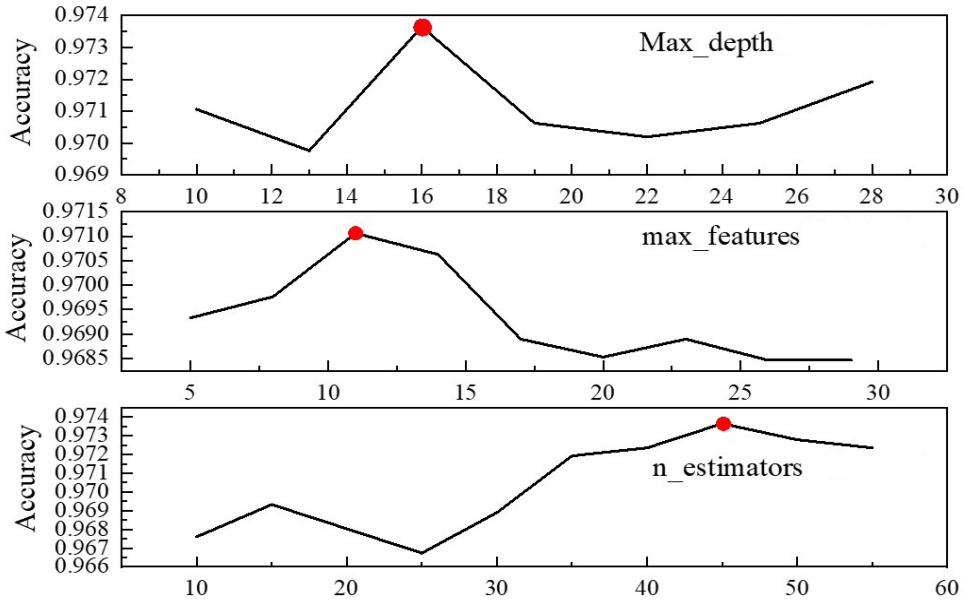


Fig. 9. The optimal parameter selection of RFC

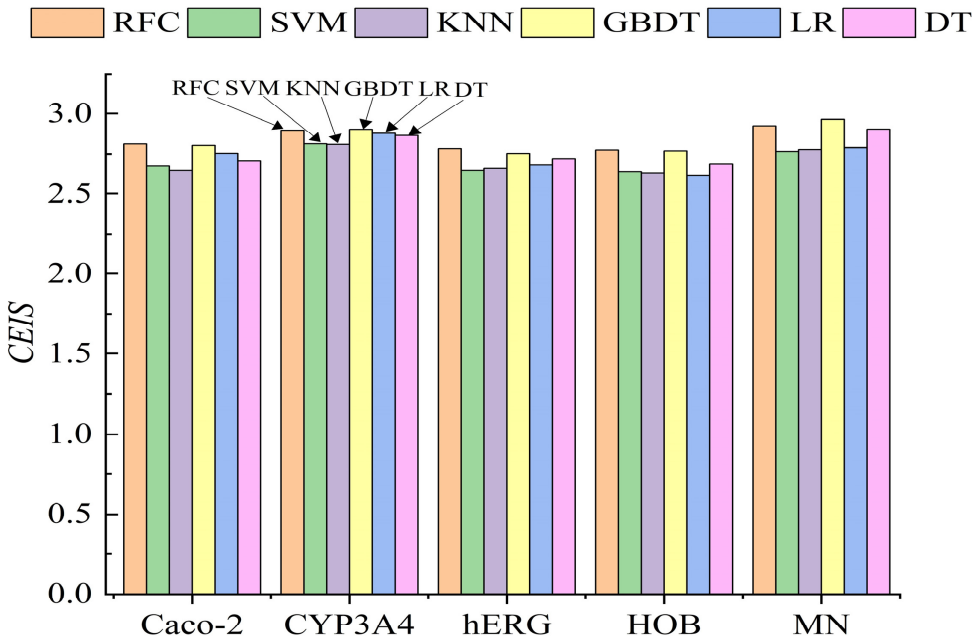


Fig. 10. The evaluation results of different models in five different ADMET properties

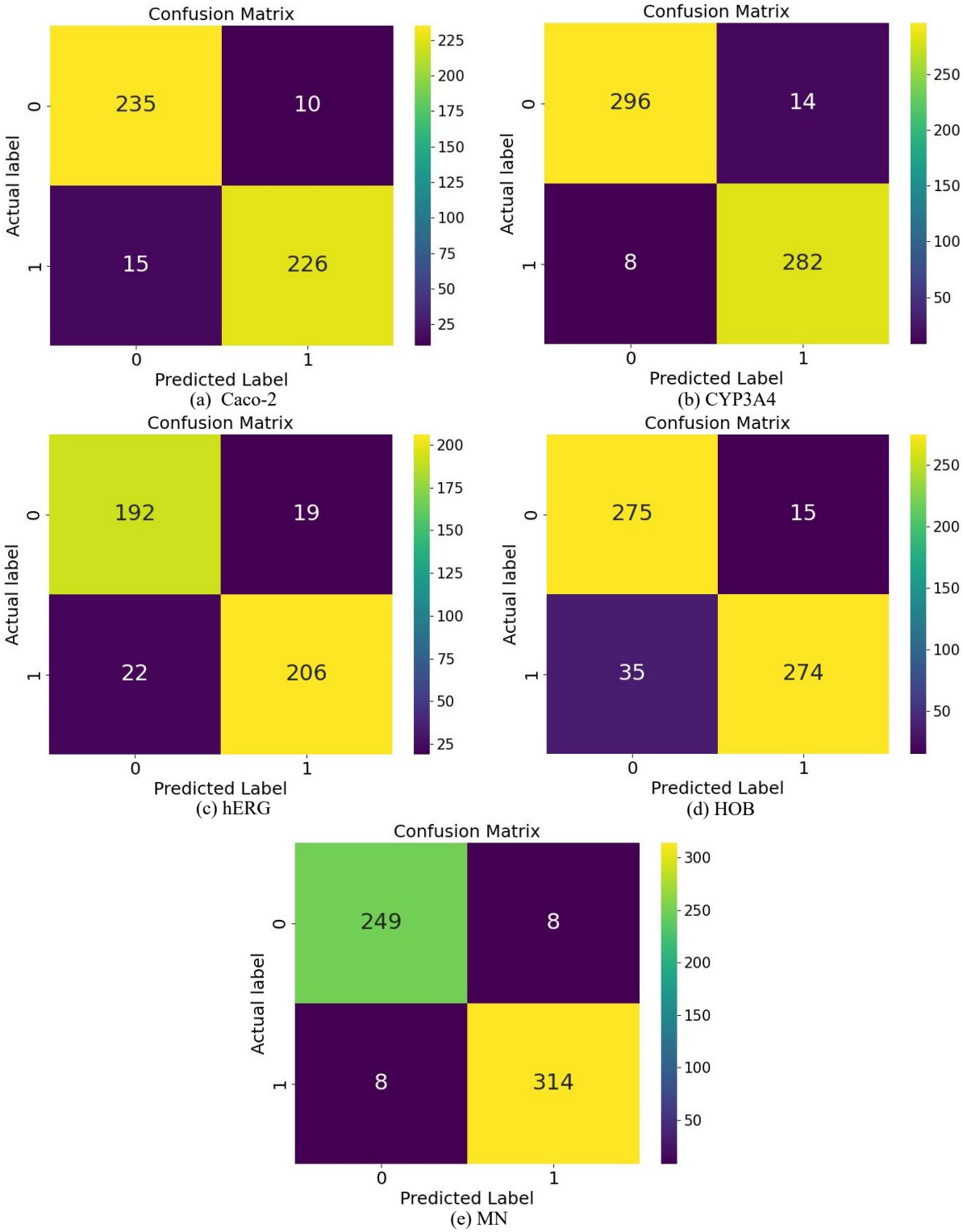


Fig. 11. Confusion matrix for five different ADMET properties

TABLE VIII
STATISTICAL TABLE OF OPTIMAL QUALITATIVE PREDICTION MODEL IN FIVE DIFFERENT ADMET PROPERTIES

ADMET	MODEL
Caco-2	RFC
CYP3A4	GBDT
hERG	RFC
HOB	RFC
MN	GBDT

VI. CONCLUSION

In this paper, we propose a quantitative and qualitative prediction method incorporating data-driven perspective and ML to solve the quantitative and qualitative prediction problem for anti-breast cancer drug candidates. In order to extract the most important molecular descriptors from high-dimensional compound features, a drug molecular descriptor selecting method is established based on RF, and we obtain the top 20 molecular descriptors, involving 10 categories. They have been validated through the correlation

and representative analyses. Then, some important methods are developed and studied, including the SMOTE-based data processing method for unbalanced ADMET classes, the quantitative prediction method of ERα bioactivity based on ML, and the qualitative prediction method of ADMET based on CEIS and ML. Through the relevant experiments and comparative analysis, the ERα biological activity prediction problem is resolved, and the category imbalance problem is also tackled under five different ADMET properties. The optimal models are also identified for five different ADMET properties. For instance, the final qualitative prediction model is RFC for Caco-2, hERG and HOB. Moreover, GBDT is the final qualitative prediction model for CYP3A4 and MN. Finally, the experimental results demonstrate that the proposed approach has superior predictive performance in quantitative and qualitative aspects. This also validates the viability, rationality and efficacy of our method.

In terms of future work, it can be summarized as follows:

(1) For the model parameter optimization, some advanced optimization algorithms [43] are considered to augment our model's generalization ability.
(2) While ensuring the model generalization ability, the model interpretability is also very crucial, and so the interpretable methods [44] can be further introduced into this paper.
(3) The existing dataset in this paper is only analyzed as a case study, but it is necessary to further improve our method in a more complex medical environment, such as combining with DL [45], [46].

REFERENCES

[1] M. J. Duffy, "Estrogen receptors: role in breast cancer," *Critical reviews in clinical laboratory sciences*, vol. 43, no. 4, pp. 325-347, 2006.

[2] M. C. Abba, Y. Hu, H. Sun, J. A. Drake, S. Gaddis, K. Baggerly, A. Sahin, and C. M. Aldaz, "Gene expression signature of estrogen receptor α status in breast cancer," *BMC genomics*, vol. 6, pp. 1-13, 2005.

[3] K. Tsaioun, B. J. Blaauboer, and T. Hartung, "Evidence-based absorption, distribution, metabolism, excretion (ADME) and its interplay with alternative toxicity methods," *ALTEX-Alternatives to Animal Experimentation*, vol. 33, no. 4, pp. 343-358, 2016.

[4] J. Balfer, Y. Hu, and J. Bajorath, "Compound structure independent activity prediction in high-dimensional target space," *Molecular Informatics,* vol. 33, no. 8, pp. 544-558, 2014.

[5] I. Niculescu-Duvaz, Z. Simon, and N. Voiculet, "QSAR application in chemical carcinogenesis. ii. QSAR analysis of a class of carcinogenesis inhibitor: retinoids," *Carcinogenesis*, vol. 6, no. 4, pp. 479-486, 1985.

[6] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini et al., "QSAR modeling: where have you been? where are you going to?," *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977-5010, 2014.

[7] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg et al., "QSAR without borders," *Chemical Society Reviews*, vol. 49, no. 11, pp. 3525-3564, 2020.

[8] S. Mukherjee, A. Saha, and K. Roy, "QSAR of estrogen receptor modulators: exploring selectivity requirements for ERα versus ERβ binding of tetrahydroisoquinoline derivatives using E-state and physicochemical parameters," *Bioorganic & Medicinal Chemistry Letters*, vol. 15, no. 4, pp. 957-961, 2005.

[9] A. Zekri, D. Harkati, S. Kenouche, and B. A. Saleh, "QSAR modeling, docking, ADME and reactivity of indazole derivatives as antagonizes of estrogen receptor alpha (ER-α) positive in breast cancer, "*Journal of Molecular Structure*, vol. 1217, p. 128442, 2020.

[10] M. T. Khan and I. Sylte, "Predictive QSAR modeling for the successful predictions of the ADMET properties of candidate drug molecules," *Current Drug Discovery Technologies*, vol. 4, no.3, pp.141-149, 2007.

[11] L. Guan, H. Yang, Y. Cai, L. Sun, P. Di, W. Li, G. Liu, and Y. Tang, "ADMET-score–a comprehensive scoring function for evaluation of chemical drug-likeness," *Medchemcomm*, vol. 10, no. 1, pp. 148-157, 2019.

[12] N. V. Puranik, P. Srivastava, G. Bhatt, D. J. S. John Mary, A. M. Limaye, and J. Sivaraman, "Determination and analysis of agonist and antagonist potential of naturally occurring flavonoids for estrogen receptor (ERα) by various parameters and molecular modelling approach," *Scientific reports*, vol. 9, no.1, p. 7450, 2019.

[13] L. L. Ferreira and A. D. Andricopulo, "ADMET modeling approaches in drug discovery,"*Drug Discovery Today*, vol.24, no.5, pp.1157-1165, 2019.

[14] M. Cipolletti, S. Bartoloni, C. Busonero, M. Parente, S. Leone, and F. Acconcia, "A new anti-estrogen discovery platform identifies FDA-approved imidazole anti-fungal drugs as bioactive compounds against ERα expressing breast cancer cells," *International Journal of Molecular Sciences*, vol. 22, no. 6, pp.2915, 2021.

[15] M. Ganesan, J. Sekar, S. P. Kandasamy, and P. Srinivasan, "Design, synthesis, spectral characterization, in silico ADMET studies, molecular docking, antimicrobial activity, and anti-breast cancer activity of 5, 6-dihydrobenzo [H] quinazolines," *Journal of Molecular Structure*, vol. 1296, pp. 136771, 2024.

[16] P. Carracedo-Reboredo, J. Linares-Blanco, N. Rodrguez-Fernandez, F. Cedron, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos, and C. Fernandez-Lozano, "A review on machine learning approaches and trends in drug discovery,"*Computational and Structural Biotechnology Journal*, vol. 19, pp. 4538–4558, 2021.

[17] A. Beheshti, E. Pourbasheer, M. Nekoei, and S. Vahdani, "QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm-multiple linear regressions," *Journal of Saudi Chemical Society*, vol. 20, no. 3, pp. 282-290, 2016.

[18] A. M. Anter, Y. S. Moemen, A. Darwish, and A. E. Hassanien, "Multitarget QSAR modelling of chemo-genomic data analysis based on extreme learning machine," *Knowledge-Based Systems*, vol. 188, p. 104977, 2020.

[19] F. Marchetti, E. Moroni, A. Pandini, and G. Colombo, "Machine learning prediction of allosteric drug activity from molecular dynamics," *The Journal of Physical Chemistry Letters*, vol. 12, no. 15, pp. 3724-3732, 2021.

[20] L. Shi, F. Yan, and H. Liu, "Screening model of candidate drugs for breast cancer based on ensemble learning algorithm and molecular descriptor,"*Expert Systems with Applications*, vol. 213, p. 119185, 2023.

[21] E. N. Feinberg, E. Joshi, V. S. Pande, and A. C. Cheng, "Improvement in ADMET prediction with multitask deep featurization," *Journal of Medicinal Chemistry*, vol. 63, no. 16, pp. 8835-8848, 2020.

[22] Y. Wei, S. Li, Z. Li, Z. Wan, and J. Lin, "Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation," *Bioinformatics*, vol. 38, no. 10, pp. 2863-2871, 2022.

[23] Y. Qin, C. Li, X. Shi, and W. Wang, "MLP-based regression prediction model for compound bioactivity," *Frontiers in Bioengineering and Biotechnology*, vol. 10, p. 946329, 2022.

[24] L. Yu, W. Jin, J. Zhou, X. Li, and Y. Zhang, "Optimal extraction bioactive components of tetramethylpyrazine in Chinese herbal medicine jointly using back propagation neural network and genetic algorithm in R language," *Pakistan Journal of Pharmaceutical Sciences*, vol. 33, no. 1, pp. 95-102, 2020.

[25] X. Li, L. Tang, Z. Li, D. Qiu, Z. Yang, and B. Li, "Prediction of ADMET properties of anti-breast cancer compounds using three machine learning algorithms," *Molecules*, vol. 28, no. 5, p. 2326, 2023.

[26] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme gradient boosting as a method for quantitative structure -activity relationships," *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2353-2360, 2016.

[27] Z. Wu, T. Lei, C. Shen, Z. Wang, D. Cao, and T. Hou, "ADMET evaluation in drug discovery.19. reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches," *Journal of Chemical Information and Modeling*, vol. 59, no. 11, pp. 4587-4601, 2019.

[28] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea et al., "Analyzing learned molecular representations for property prediction," *Journal of Chemical Information and Modeling*, vol. 59, no. 8, pp. 3370-3388, 2019.

[29] Y. Gao, S. Chen, J. Tong, and X. Fu, "Topology-enhanced molecular graph representation for anti-breast cancer drug selection," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1-21, 2022.

[30] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia, H. PerezSanchez, J. A. Benediktsson, A. Thapa, and A. Barr, "Automatic selection of molecular descriptors using random forest: Application to drug discovery," *Expert Systems with Applications*, vol. 72, pp. 151-159, 2017.

[31] R. D. Cramer, P. Cruz, G. Stahl, W. C. Curtiss, B. Campbell,B. B. Masek, and F. Soltanshahi, "Virtual screening for R-groups, including predicted pIC50 contributions, within large structural databases, using topomer comfa," *Journal of Chemical Information and Modeling*, vol. 48, no. 11, pp. 2180-2195, 2008

[32] K. Zhou, C. Zhang, Y. Yu, S. Cong, and X. Yue, "Improving SMOTE technology for credit card fraud detection category imbalance issues," *Engineering Letters*, vol. 31, no. 4, pp. 1780-1785, 2023.

[33] D. C. E. Saputra, A. Azhari, and A. Ma'arif, "K-nearest neighbor of beta signal brainwave to accelerate detection of concentration on student learning outcomes," *Engineering Letters*, vol. 30, no. 1, pp. 318-324, 2022.

[34] D. Esan, P. A. Owolawi, and C. Tu, "Anomalous detection in noisy image frames using cooperative median filtering and KNN," *IAENG International Journal of Computer Science*, vol. 49, no. 1, pp. 1-10, 2022.

[35] A. Dwivedi and A. Tajer, "GRNN-based real-time fault chain prediction," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 934-946, 2024.

[36] S. Ding, Y. Xu, T. Sun, J. Yu, L. Wang, and R. Zhu, "Roadside unit visibility prediction method based on SVR," *Engineering Letters*, vol. 31, no. 1, pp. 419-434, 2023.

[37] X. Zhang, Z. Yu, Y. Hu, and J. Yang, "Milling force prediction of titanium alloy based on support vector machine and ant colony optimization," *IAENG International Journal of Computer Science*, vol. 48, no. 2, pp. 223-235, 2021.

[38] L. Zhang, L. Luo, L. Hu, and M. Sun, "An SVM-based classification model for migration prediction of Beijing," *Engineering Letters*, vol. 28, no. 4, pp. 1023-1030, 2020.

[39] A. A. Syed, Y. Lukas, and A. Wibowo, "A comparison of machine learning classifiers on laptop products classification task," in *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2021*, 20-22 October, 2021, Hong Kong, pp. 104-110.

[40] A. F. Jamali, A. Mustapha, and S. A. Mostafa, "Prediction of sea level oscillations: Comparison of regression-based approach," *Engineering Letters*, vol. 29, no. 3, pp. 990-995, 2021.

[41] A. Nugroho, H. L. H. S. Warnars, F. L. Gaol, and T. Matsuo, "Trend of stunting weight for infants and toddlers using decision tree," *IAENG International Journal of Applied Mathematics*, vol. 52, no.1, pp. 144-148, 2022.

[42] Y. Zhong, L. Luo, X. Wang, and J. Yang, "Multi-factor stock selection model based on machine learning," *Engineering Letters*, vol. 29, no. 1, pp. 177-182, 2020.

[43] L. Wang, Y. Yao, X. Luo, C. D. Adenutsi, G. Zhao, and F. Lai, "A critical review on intelligent optimization algorithms and surrogate models for conventional and unconventional reservoir production optimization," *Fuel*, vol. 350, p.128826, 2023.

[44] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206-215, 2019.

[45] R. S. El-Sayed, "A hybrid CNN-LSTM deep learning model for classification of the Parkinson disease," *IAENG International Journal of Applied Mathematics*, vol. 53, no. 4, pp. 1427-1436, 2023.

[46] L. Shen, J. Yang, M. Xu, and B. Yang, "TS-DRN: an EEG recognition algorithm for art design decisions making," *IAENG International Journal of Computer Science*, vol. 51, no. 2, pp. 130-142, 2024.