IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 3, 2020, accepted December 24, 2020, date of publication December 28, 2020,
date of current version January 19, 2021.

*Digital Object Identifier 10.1109/ACCESS.2020.3047852*

# Prediction of Blood-Brain Barrier Permeability of Compounds by Fusing Resampling Strategies and eXtreme Gradient Boosting

**ZHIWEN SHI[1,2,3], YANYI CHU[1], YONGHONG ZHANG[4], YANJING WANG[1], AND DONG-QING WEI[1,2,3]**

[1]State Key Laboratory of Microbial Metabolism, School of Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Joint Laboratory of International Cooperation in Metabolic and Developmental Sciences, Ministry of Education, Shanghai Jiao Tong University (SJTU), Shanghai 200240, China
[3]Peng Cheng Laboratory, Shenzhen 518055, China
[4]Chongqing Research Center for Pharmaceutical Engineering, College of Pharmacy, Chongqing Medical University, Chongqing 400016, China

Corresponding authors: Dong-Qing Wei (dqwei@sjtu.edu.cn) and Yanjing Wang (wangyanjing@sjtu.edu.cn)

The work of Dong-Qing Wei was supported in part by the Key Research Area Grant of the Ministry of Science and Technology of China under Grant 2016YFA0501703; in part by the National Science Foundation of China under Grant 32070662, Grant 61832019, and Grant 32030063; in part by the Science and Technology Commission of Shanghai Municipality under Grant 19430750600; in part by the Natural Science Foundation of Henan Province under Grant 162300410060; and in part by the SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University under Grant YG2017ZD14.

**ABSTRACT** Computer-aided drug design is an efficient method to analyze the development of disease-related drugs. However, developed as binding targets, medicines perform well in cell models and animal models but fail in human models. One main reason for this failure is that the human body has natural barriers, such as the blood-brain barrier, to block exogenous macromolecules. Thus, efficient and accurate predictions of drug molecules that can effectively pass the blood-brain barrier is necessary in developing drug treatments for brain tissue diseases. In this study, 7658 molecular structure features were extracted from 2354 drug molecule SMILE strings using computational methods. By integrating three feature selection algorithms of machine learning, 33 chemical structure features with significantly discriminant performance were screened out and used to construct multiple discriminant models. After a comprehensive comparison, the XGBoost model was selected as the final prediction model. After data preprocessing and parameter optimization, the model achieved 95% accuracy on the training set. To verify the model's stability, we introduced an external data set, which reached 96% accuracy of the model. This study applies new resampling methods and machine learning algorithms, and adjusts the application of resampling methods to obtain new chemical features to construct machine learning predictors. The features may contribute to the significant drug development that integrates biological analysis and machine learning algorithms.

**INDEX TERMS** Blood-brain barrier, data imbalanced, machine learning, eXtreme Gradient Boosting (XGBoost), computational biology, resample methods.

## I. INTRODUCTION

Although drug design and discovery result in various potential drugs candidates, most drug candidates cannot be finalized and marketed due to varying problems related to absorption, distribution, metabolism, addition, and toxicity (ADMET) [1]. Therefore, to reduce loss associated with the drugs, the low marketability of which caused by poor ADMET properties, it is necessary to develop a method that maximizes efficiency in developing novel drugs [2]–[4]. Applying computational technology to drug screening can significantly reduce cost and save time in studying drug ADMET properties in vivo or in vitro [5]–[8].

The blood-brain barrier refers to the barrier formed by brain capillary walls and glial cells between plasma and brain cells. The barrier is formed by the choroidal plexus between plasma and cerebrospinal fluid [9], [10]. The barrier can prevent most harmful substances in the blood from entering

The associate editor coordinating the review of this manuscript and approving it for publication was Trivikram Rao Molugu.

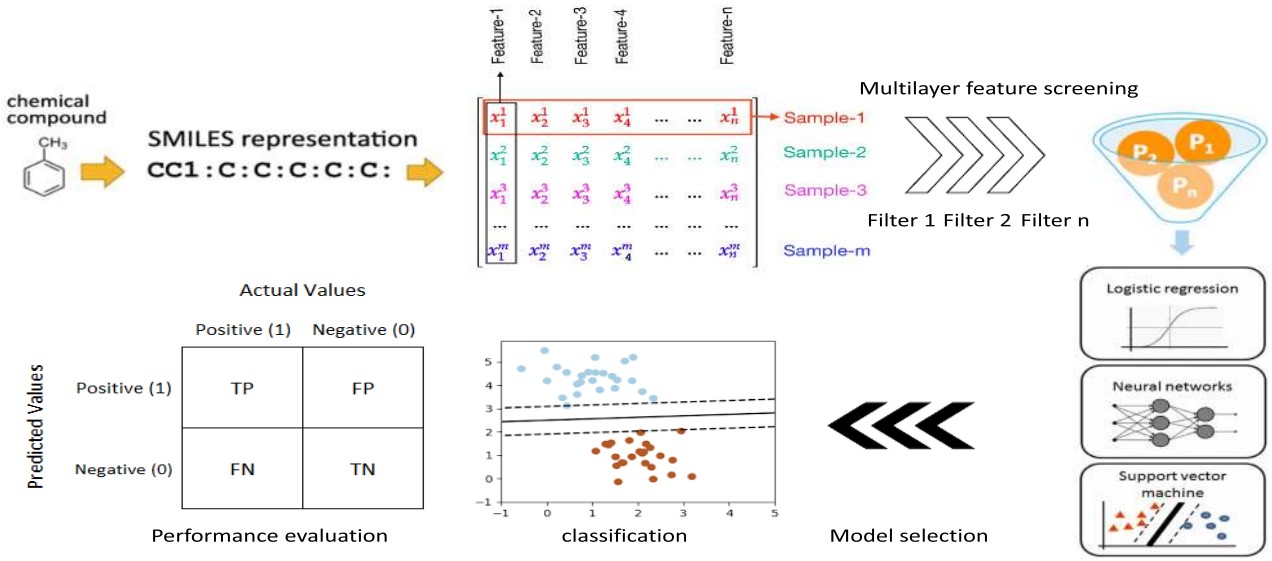


**FIGURE 1.** Workflow of optimized model The general workflow is shown in the illusive graph. We began with the parsing the smile strings and obtained 7658 chemical features in total. Then we applied the feature selection methods to screen the most significant signatures. Eventually, we received 33 features that present significant performance distinguishing positive and negative molecules. We constructed multiple classifiers and compared different machine learning algorithms. Finally, we selected the XGBoost model as the best model according to the F1-score. The trained model was validated using an independent external data.

into the brain tissue. It is different for the various solutes in the blood to enter the brain tissue from the brain capillaries, where some pass quickly, some are slow, and some cannot pass at all. This selective osmosis phenomenon makes people imagine that there may be a particular structure that can restrict solute penetration [11]. This structure can reduce or even protect the brain tissue from damage by harmful substances circulating blood, thus protecting the brain tissue. The stability in the brain that is maintained by the blood-brain barrier has crucial biological significance for maintaining the normal physiological state of the central nervous system [12].

This stability protects the brain, but serious diseases disrupt this environment ignoring the protection of the stability. Central nervous system (CNS) has become the second-largest disease after cardiovascular diseases [13]. Despite the rise in quantity of CNS diseases, the success rate in developing drugs related to these illnesses is shallow. Compared with the 20% success rate of prescriptions for cardiovascular diseases, the success rate of prescriptions for CNS diseases is only 8%. A significant factor affecting the success of CNS drug development is the blood-brain barrier (BBB), which blocks nearly 100% of large molecule drugs and more than 98% of small molecule drugs [14].

Therefore, in addition to overcoming ADMET issues, CNS drugs must also overcome the BBB and achieve sufficient exposure in its targeted region. This is the key to the success of CNS drug development [15].

Current experimental methods with the best accuracy to screen drug candidates that can pass the BBB are limited due to high costs and labor [16]. Because it is impossible to determine all potential candidates that can cross the blood-brain barrier through current experimental methods. it is imperative and even desirable to apply an evaluation method based on a computational model [17]–[19].

The rapid development of prediction models has divided into statistical methods and machine learning algorithms [20]. In this study, we use traditional machine learning such as Logistic regression [21],support vector machine (SVM) [22], Naive Bayes [23], random forest(RF) [24], XGBoost [25] and multilayer perceptron(MLP) [26] to build supervised regression or classification models. Compounds that can cross the BBB (BBB+) and compounds that cannot cross the BBB (BBB-) are used as the label of the models. SVM generally performs well in higher dimension and SVM is the best algorithm when classes are separable. What's more, outliers in SVM models have less impact to the prediction. However, disadvantages in SVM models, such as the long time to process a larger dataset, existed in model training. SVM with overlapped classes may have poor performance. Selecting appropriate hyperparameters is important for the performance. Selecting the appropriate kernel function can be tricky. Naive Bayse is very fast and can be used in real time. It's scalable with large datasets and insensitive to irrelevant features. Multi class prediction is effectively done in Naive Bayes. It generally has good performance with high dimensional data. The disadvantages of Naive Bayes are the independence of features which can not be guaranteed, and Naive Bayes is a bad estimator. The logistic regression is simple to implement. Feature scaling is not needed. Tuning of hyperparameters is not needed, either. The disadvantage of logistic regression is the poor performance on non-linear data. The poor performance with irrelevant and highly correlated features may be involved. The logistic regression is not very powerful algorithm and can be easily outperformed by

other algorithms. It also has high dependancy on the proper presentation of data. Random forest has outperformed on imbalanced datasets, and can handle huge amount of data and missing data. Outliers have little impact on random forest. It's useful to extract feature importance. One of the shortcomings is that features need to have some predictive power, else, they won't work. The predictions of the trees need to be uncorrelated. XGBoost requires less feature engineering. Feature importance can be found out. Outliers have minimal impact. XGBoost handles large-sized datasets well and is fast to interpret. XGBoost is less prone to overfitting. However, the XGBoost is difficult to be interpreted, and tough in visualization. Overfitting is possible if parameters have not be tuned properly which is because of the large number of hyperparameters.

Previous studies that utilized traditional models did not consider data imbalance, leading to problems with model precision and poor performance in independent testing. Most of the existing data lack BBB- samples, which poses a problem for accurate predictions on independent data. One such model was built by Dmitry A *et al.* and Konovalov [27] where 328 compounds in vivo were tested for BBB permeability and predicted by logBB value, which calculated a classification accuracy at only 0.766. The best model developed by Andrey A.Toropov *et al.*, which tested 41 compounds, only had an accuracy of 0.896 accuracy [28]. Martins *et al.* used a series of support vector machines (SVM) and RF classification models based on the Bayesian method. The results from these models has an accuracy of 0.947 [29].

To reconcile the discrepancies created by data imbalance, our study uses several resampling methods, such as upsampling [30], adaptive synthetic sampling(ADASYN) [31], Ra-ndomUnder Sampler(RUS) [32], SMOTE+ENN(edited nearest neighbor) [33] and Synthetic Minority Oversampling Technique(SMOTE) [34]. In this study, the data set was obtained from a recent study [35], which contains 2354 compounds, including 1807 BBB+ samples and 547 BBB- samples for model construction and 92 CNS+ chemicals retrieved from a related article to the external data set for the independent evaluation. Our model has dramatically improved the accuracy of classification predictions for 2356 compounds by using these resampling methods. Although the resampling method dramatically improves the model's accuracy in the test set, its prediction performance on independent data sets needs to be evaluated separately to avoid issues of over fitting.

## II. MATERIAL AND METHODS

### A. DATASETS

In this study, the data set was obtained from the recent study [35], which was integrated by the last four studies [36]–[39], In dataset, if it includes the compounds such as noncovalent, inorganic, mixtures, salt, or those with MW greater than 1000 Da, they were removed from the dataset depends on real cases. The researchers used LogBB as the criterion to divide compounds into BBB + and BBB- if LogBB $>1$ and LogBB $< -1$, respectively [35]. The data set contains 2354 compounds for model constructed and 92 CNS+ chemicals retrieved from a related article to the external data set for the independent evaluation. The full data set includes a self-generated set ID, the generic name as referred to in the literature, the canonical SMILES derived from the literature, the canonical SMILES for each compound, the binary classification of the BBB (''p'' stands for BBB+ and ''n'' for BBB-), and the reference ID of the related articles. The added 92 CNS+ samples contain their PubChem CID, name, and canonical SMILES.

### B. FEATURE SET

We used the java packages called PaDEL-Descriptor [40] to generate 4 different molecular fingerprints, namely, MACCS(166 features), PubChem(881 features), Substructure(307 features), and Klekota Roth(4860 features) and 2D descriptor(1444 features). Finally, we extracted 7658 chemical features in total. We initially removed the 25% features with the lowest variance. The variance represents the average of the squared differences from the mean value of each feature. The features with smaller variance are supposed to contribute less than other features. These features are not representative between two groups and thus are considered as non-significant.

### C. DATA PREPROCESSING

There are several issues to be addressed before the downstream analysis. First, the quantification of each chemical feature is scaled in different intervals. The Z test was applied to normalize all the chemical features, and thus all the features were scaled into the same distribution with a mean value of zero and standard deviation of one. We first calculated the mean value and standard deviation of each feature. Then we calculated the Z score of each feature using the difference between original value and mean value by the standard deviation. Second, outliers may exist and have a considerable impact on the classifiers. Therefore, we computed a confidence interval using the mean value and three times the standard deviation for each feature and removed all the outliers beyond the corresponding confidential interval. The last issue is about the imbalanced samples in two groups. The number of positive molecules is over three times larger than that of negative molecules. We resample the data before feature selection, which is different from the existing research. We think it is more reasonable. We used six types of methods to solve the problem of imbalanced data sets.

1. **Upsampling:** Upsampling is the process of randomly duplicating observations from the minority class to reinforce its signal.
2. **RUS:** The idea of RUS is also relatively simple, which is to select some randomly from most classes of samples and eliminate them
3. **Weight parameter:** Set the weight parameters for each sample and adjust the prediction error loss of a few

classes to be greater than that of most classes. When we use the machine learning model in the *sklearn* package, we set the best weight parameters to get the best results.

4. **SMOTE:** The SMOTE algorithm's basic idea is to analyze the minority samples and synthesize new samples according to the minority samples to add them to the data set. For each sample in the minority class, the Euclidean distance is used as the standard to calculate the distance from it to all the samples in the minority sample set and get its k-nearest neighbor [41]. According to the unbalanced proportion of samples, a sampling ratio is set to determine the sampling rate n. for each minority sample, several samples are randomly selected from its k-nearest neighbors, assuming that the selected nearest neighbor is [34].
5. **SMOTE+ENN:** SMOTE method is used to generate a new minority class sample to obtain the expanded dataset. For each sample, if more than half of its K nearest neighbors do not belong to the majority class, the sample will be rejected [33].
6. **ADASYN:** It uses a mechanism to automatically determine how many composite samples each minority class sample needs to produce, instead of synthesizing the same number of samples for each minority class sample like SMOTE [31].

### D. FEATURE SELECTION

Given the large number of features, we compiled a pipeline of different feature selection methods. In other words, we use the following procedure method to filter features layer by layer, instead of using the best way to filter once, which is why we can achieve better accuracy with fewer (33) features. We applied a multilayer screening process composed by four independent feature selection methods. And this screening process was performed layer by layer. As different feature selection methods focus on diverse aspects and eventually screen different features, we combined multiple feature selection algorithms. After a multilayer screening, we can guarantee that all the retained features are considered to be significant by all feature selection methods. In this way, the feature set is supposed to be more robust and stable any using any of the single feature selection method.

1. **Variance threshold (VT):**For each value of feature variance, we deleted all features whose variance was not higher than the threshold. The features were ranked in descending order based on the variance. Finally, we removed all but the top 200 highest scoring features.
2. **Tree-based feature selection:** The prediction model based on a decision tree can be used to calculate the importance of features, so it can be used to remove irrelevant features.
3. **Univariate feature selection:** Univariate feature selection selects the best feature based on univariate statistical testing. It can be regarded as a preprocessing of the prediction model.
4. **Recursive feature elimination (RFE):** The recursive feature elimination method uses a machine learning model to carry out multiple rounds of training rounds. After each round of training, the features corresponding to some weight coefficients are eliminated, and then the next round of training is carried out based on the new feature set[42]

### E. CLASSIFIER CONSTRUCTION

We investigated different machine learning classifiers, including the Logistic regression, Naive Bayes, SVM, RF, and XGBoost. All the models were constructed using the *sklearn [43]* python package. Then we compared all models' performance in terms of precision, recall, F1 score, and other criteria, including AUC, ACC and G-means. All the models were initially constructed with default parameters. The best model was selected based on the performance, and afterwards we applied the parameter optimization using grid search.

1. **Logistic regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist [21], [44], [45].
2. **Naive Bayes:** Naive Bayes classifiers are a family of simple ''probabilistic classifiers'' based on applying Bayes' theorem [46] with strong independence assumptions between the features. They are among the simplest Bayesian network models [23].
3. **SVM:** It is a supervised learning model with associated learning algorithms that analyze data for classification and regression analysis. In addition to performing linear classification, SVM can efficiently perform a non-linear classification using the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [22].
4. **RF:** RF is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees [24], [47]–[53].
5. **XGBoost:** XGBoost is an optimized distributed gradient enhancement library designed to be efficient, flexible and portable, and is a tree integration model. It sums up the K (tree number) tree results as the final prediction value [25], [54]–[56].
6. **MLP:** MLP is a kind of artificial neural network with a forward structure, which maps a set of input vectors to a set of output vectors. MLP can be regarded as a directed graph, which is composed of several node layers. Each layer is connected to the next layer. Except for the input node, each node is a neuron with a nonlinear activation function [26].

### F. PARAMETER OPTIMIZATION

To achieve better fitting, we applied the grid search algorithm [57] to optimize the best model parameters. In this

study, we chose the essential parameters of models and searched for the best combination. The optimized parameters include the gamma, learning rate, and max depth. The parameter optimization process was done using the *GridSearchCV* function in the python *sklearn[43]* package.

### G. MODEL EVALUATION

To comprehensively analyze models' performance under different conditions, we investigated the model's performance based on all possible combinations. The combination components include additional models, other resampling methods.

We use our test set to evaluate the performance of different feature selection and unbalanced learning methods. Besides, 10-fold cross-validation [58]–[66] is used to verify the accuracy of the model to find the best algorithm to improve test data classification. In this study, we selected 10-fold instead of 5-fold in order to obtain robust performance. The area under of the Receiver Operating Characteristic curve (AUC), Accuracy (ACC), recall, F1-score were used to evaluate the model's accuracy. These metrics have been widely used in the recent bioinformatics studies [67]–[75]. To validate the performance of our selected chemical features and the precision of the model, we applied an independent external data for validation. The external data are drugs of 92 CNS+. Since the external data only contains CNS+ samples, we simulated the CNS- samples based on the smote algorithm. Eventually, we obtained balanced CNS+ and CNS- samples in the external data.

The general workflow is shown in the illusive graph. We began with the parsing the smile strings and obtained 7658 chemical features in total. Then we applied the feature selection methods to screen the most significant signatures. Eventually, we received 33 features that present significant performance distinguishing positive and negative molecules. We constructed multiple classifiers and compared different machine learning algorithms. Finally, we selected the XGBoost model as the best model according to the F1-score. The trained model was validated using an independent external data.

## III. RESULTS

### A. DATA PREPROCESSING

The use of different resampling methods directly impacts the number of BBB+ and BBB- samples in the data. The total increased number of positive and negative samples in the training set after using resampling methods is shown in Fig. 2.

We choose to resample the data before feature selection, which is more reasonable in our opinion. In this study, we used 75% of all samples as the training set and the rest 25% as the test set, and 92 CNS+ chemicals retrieved from a related article to the external data set for the independent evaluation.

### B. THREE-STEP FEATURE SELECTION PROCESS

A three-filters composed feature selection pipeline was introduced to screen the most significant features, and the distribution can be seen in Fig 3A.

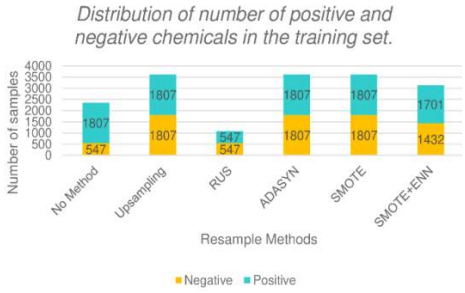


**FIGURE 2.** Distribution of the number of positive and negative chemicals in the training set. The x-axis represents the preprocessing method composed of no method, upsampling, RUS, ADASYN, SMOTE and SMOTE+ENN. The y-axis represents the number of samples in each group.

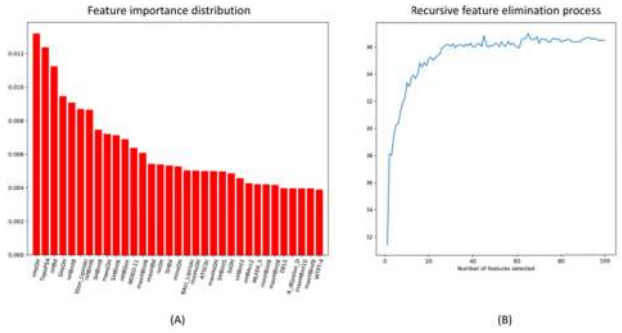


**FIGURE 3.** Feature selection process. (A) The top 30 features and corresponding importance scores are shown. (B) The process of feature selection using a recursive feature selection algorithm.

The features were ranked according to their importance in descending order. It suggests that the feature importance follows a long tail distribution, which means small partial features have contributions, while most features do not make sense.

Fig. 4 shows the distribution of features in three datasets. The x-axis represents the feature value, and the y-axis represents the corresponding probability. The blue bars represent the overall distribution of each feature. We used lines of different colors to indicate the distribution in each dataset accordingly. As seen from the figure, given a specific feature, positive and negative distributions have significant differences. Meanwhile, the training dataset and external dataset also present diverse distribution. This may be attributed to the batch effect

To illustrate the two groups of molecules present various distributions on these features, we visualized the distribution of molecules from three datasets, as shown in Fig.5A. The positive and negative molecules have significant differences. Some features also present interactions, and thus, we computed the internal correlation between any pair of features, as shown in Fig. 5B. The heatmap indicates diverse intensity of interaction. As expected, some features have significant interactions and present darker colors. Moreover, we found that our significant features are different from the existing studies, which may be due to the resampling before feature selection and the use of triple-feature screening to obtain essential features.

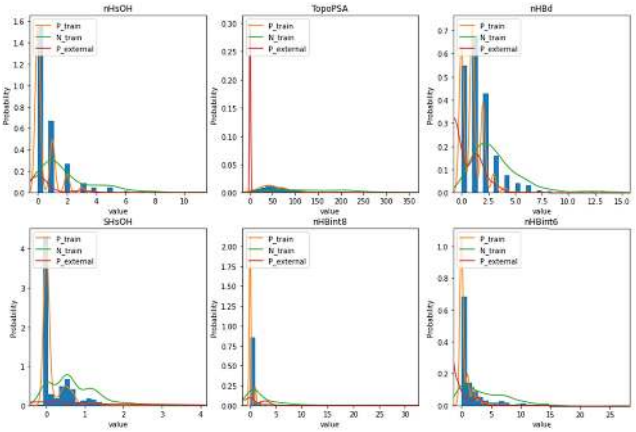


**FIGURE 4.** Distribution of features in three datasets. The positive and negative samples in the train set are marked using orange and blue lines, respectively. The red line indicates the positive samples in the external dataset. The barplot represents the overall distribution of all samples in the train set.

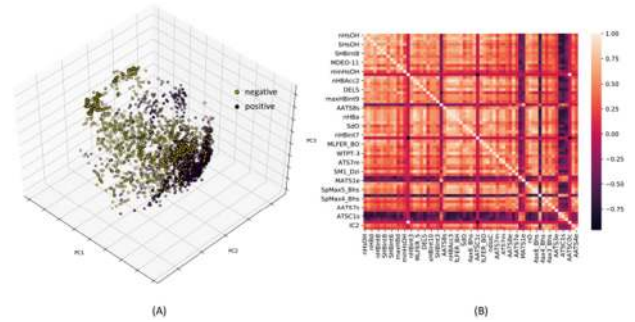


**FIGURE 5.** Significant features interaction (A) The positive (purple) and negative (yellow) molecules are marked in different colors. Using principal component analysis, the two groups of molecules are grouped into separate clusters. (B). The correlation heatmap shows the internal correlation between each pair of features.

Then we applied the selected features to train the classifiers based on different algorithms. The construction process was done using the *sklearn* with default parameters. 75% of all samples were randomly selected as a train set.

### C. DATA PREPROCESSING

Table 1 shows the six machine learning models' performance on the training set without any resampling method, and n stands for BBB- dataset, p stands for BBB+ dataset. The global accuracy like ACC and AUC only reflect the overall performance rather than independent accuracy. Therefore, we investigated the models' performance on each group, in terms of precision, recall and F1 score. In this way, it's clear that whether a model is more sensitive to predict each group. Table 1 shows the performance of different models regarding positive and negative groups, independently. We find that the prediction accuracy of the six models for BBB + compounds is much lower than that for BBB- compounds. And the performance of XGBoost is the best among the six models. However, this accuracy is not as high as we expected, and the prediction accuracy on BBB + is not as good as expected, so it is necessary to use the resampling method.

**TABLE 1.** Performance of all models.

| Model | | Precision | Recall | F1-score |
|---|---|---|---|---|
| **Logistic** | n | 0.89 | 0.96 | 0.92 |
| | p | 0.83 | 0.62 | 0.71 |
| | macro avg | 0.86 | 0.79 | 0.81 |
| | weighted avg | 0.87 | 0.88 | 0.87 |
| **Naive Bayes** | n | 0.91 | 0.86 | 0.88 |
| | p | 0.62 | 0.73 | 0.67 |
| | macro avg | 0.77 | 0.79 | 0.78 |
| | weighted avg | 0.84 | 0.83 | 0.83 |
| **SVM** | n | 0.87 | 0.96 | 0.92 |
| | p | 0.83 | 0.57 | 0.67 |
| | macro avg | 0.85 | 0.76 | 0.79 |
| | weighted avg | 0.86 | 0.87 | 0.86 |
| **Random Forest** | n | 0.91 | 0.96 | 0.94 |
| | p | 0.90 | 0.70 | 0.77 |
| | macro avg | 0.88 | 0.83 | 0.85 |
| | weighted avg | 0.90 | 0.90 | 0.89 |
| **XGBoost** | n | 0.92 | 0.96 | 0.94 |
| | p | 0.87 | 0.75 | 0.80 |
| | macro avg | 0.90 | 0.86 | 0.87 |
| | weighted avg | 0.91 | 0.91 | 0.91 |
| **Multilayer perceptron** | n | 0.90 | 0.93 | 0.92 |
| | p | 0.76 | 0.68 | 0.72 |
| | macro avg | 0.83 | 0.80 | 0.83 |
| | weighted avg | 0.87 | 0.87 | 0.87 |

### D. PERFORMANCE UNDER DIFFERENT COMBINATIONS

Besides, we also wanted to further investigate the model's differential performance under different combinations of machine learning models and resampling methods. We pre-processed the train set in each variety using a specified method, including ADASY, RUS, SMOTE, etc. Then we selected a model from the Logistic, random forest, Naive Bayes, SVM and XGBoost. Finally, the classifier can be weighted or not to calibrate the imbalance between the two groups. A series of criteria were adopted to estimate the model's performance, such as AUC, ACC, F1score and ACC on external data.

After combining the six machine learning algorithm models with resampling methods, we find that most of them improve the test set's accuracy, and only a few reduce the accuracy. In the combination of accuracy improvement, many of the independent test accuracy of external data is extremely low, overfitting. However, when the independent test accuracy of external data is high enough and similar, we prefer to choose the combination that performs better on the test set because it may have better performance for other external data.

As shown in Table 2, the combination of smote + ENN and logistic regression performs well on the test set. Still, the accuracy of external data is inferior, which is the apparent overfitting, and so is the combination of smote and RF. The combination of ADASY and Naive Bayes does not perform well in the test set but has good results in the accuracy of external data. Still, this combination is also unreliable because it can not guarantee that it has such high accuracy in the case of low test set accuracy for other external data.

Finally, our best combination is upsampling and XGBoost, which has good performance in both test sets and external data.

**TABLE 2.** The combination of machine learning models and resampling methods with typical results.

| Combination | AUC | ACC | f1-score | Ex_acc |
|---|---|---|---|---|
| **SMOTE+ENN, Logistic** | **0.97** | 0.92 | 0.92 | **0.37** |
| **ADASY, Naive Bayes** | 0.73 | 0.66 | **0.61** | **0.90** |
| **RUS, SVM** | 0.92 | 0.84 | 0.82 | **0.95** |
| **Upsampling,XGBoost** | 0.98 | 0.95 | 0.95 | **0.96** |
| **SMOTE, RF** | **0.98** | 0.94 | 0.94 | **0.66** |
| **ADASY, SVM** | 0.95 | 0.89 | 0.89 | 0.91 |

**TABLE 3.** Performance of the four models.

| MODEL | COMBINATION | AUC | ACC | F1-SCORE | EX_ACC |
|---|---|---|---|---|---|
| **A** | No resample, XGBoost | **0.90** | 0.89 | 0.74 | **0.77** |
| **B** | Weight parameter, XGBoost | **0.93** | 0.91 | **0.80** | **0.80** |
| **C** | Upsampling, XGBoost | **0.98** | 0.95 | 0.95 | **0.96** |
| **D** | SMOTE+ENN, XGBoost | **1.00** | 0.97 | 0.97 | **0.64** |

### *E. OPTIMIZED MODEL AND VALIDATION ON EXTERNAL DATASETS*

We find that although the resampling method can get a high-precision model in the training set, it may have poor performance in external data testing, which may be caused by data overfitting. Even in some cases, the resampling method will reduce its performance on the training set, resulting underfitting in the final model.

Finally, the obtained best model is to use the method of upsampling to deal with data imbalance and build a model based on XGBoost. Because the combination of the XGBoost model and upsampling achieves the best performance, as we compare the results of other resampling methods based on XGBoost.

Four representative models are used to model, and the performance on the training data is shown in Table3. As shown in Table 3, model A and model B's accuracy on the test set is not high, but model, C and model D, have high precision, especially model D.

To facilitate describing the four models, we represented them as A, B, C and D. The accuracy of model A and model B is not satisfactory in the external data set, but model D's accuracy is also very low. Considering the excellent performance of model D in the training set, model D is supposed to be overfitted, and model D has good performance in both the test set and external data set. This consequence suggests the potential negative effect of SMOTE and ENN preprocessing. Any additional data preprocessing arises the chance of overfitting.

The optimized model is model C. The model can finally get 0.95 accuracy on the training set. In order to verify the stability of the model, we also introduce an external data, the accuracy of the model on the external data is 0.96. Our model's performance is comparable to the previously published methods. [18], [27], [29], [35]

## IV. DISCUSSION

To find the most optimized model, we emphasized resampling over feature selection. In feature selection, several feature selection methods (VT, Tree-based, Univariate feature selection, Recursive feature elimination) are used to filter layer-by-layer. The combination of each model and resampling method was then used to find the optimized model.

In this study, we applied a multilayer screening process composed by different feature selection methods. The features obtained after multilayer screening are different from those obtained by existing studies. This may be since our method emphasized resampling first over feature selection. Moreover, we found that many models in training set greatly improved in accuracy after resampling, but the accuracy rate in the independent validation on external data remained extremely low. We use SMOTE usage data based on SVM and obtained good results in the test set with an AUC equal to 0.95, ACC equal to 0.9, and F1 score of 0.99. However, in the independent verification of external data, ACC was only 0.44 and g-means was only 0.29 due to overfitting. This discrepancy was not rare. It's worth mentioning that, we also observed that after using resampling methods, the risk of overfitting is also increased. It implies that the artificial interfere may introduce unexpected negative effect on the model fitting. Therefore, some algorithms perform well with unbalanced data such as Bayesian. These models generally provide an additional parameter like 'weighted class'. The weighted models are also included in our combination comparison.

Additionally, some models' accuracy on the test set was very low, but their performance in independent testing was excellent. For example, the ADASY + Naive Bayes model's accuracy on the test set is not satisfactory, with AUC calculated at 0.73 and ACC estimated at 0.66. Still, the accuracy of external data was as high as 0.82. This model is supposed to be accurate due to its performance on external data for independent testing. This model may perform well on this external data but may not achieve the expected effect due to its lack of fitting in other data sets.

When compared to the existing research on predicting BBB permeability, our study applied the resampling method to solve data imbalance. We adjusted the resampling method's timing, chose the most critical 33 features by multiple feature selection, and paid attention to the problem of model overfitting. Still, these studies rarely involve the imbalance of BBB data samples, or if they did take into account the problem of inequality, these methods were not effective in their solution. The advantage of our method comprehensively considered the imbalance of data samples, a large number of samples were appropriately synthesized, and all original samples were deleted. Moreover, existing research that uses a resampling method for modeling poses several problems, including ignoring the issue of overfitting and inaccuracy due

to filtering features before resampling, which may not conform to real situations and is not reasonable. Current research uses one feature regarding feature selection compared with various feature selection methods, while we integrate several methods for layer-by-layer selection. This multi-layer approach to feature selection might be one reason why we were able to achieve high accuracy with few features.

Finally, our best model has the best performance in the independent verification of training set and external data and achieved an accuracy rate of 0.96 on the test set and 0.96 in the independent verification of external data. The optimized parameters of the best model are base_score = 0.5, booster = 'gbtree', colsample_bylevel = 1, gamma = 0.001, gpu_id = −1, importance_type = 'gain', interaction_constraints = '', learning_rate = 0.1, max_delta_step = 0, max_depth = 50.

## V. CONCLUSION

We used six machine learning algorithms, four feature selection methods, and six resampling methods where the machine learning method was combined with the resampling method. The timing of resampling was adjusted accordingly. The accuracy of the model on the test set is verified by 10-fold cross-validation. Finally, it was found that the best model can be obtained by using a combination of upsampling and the XGBoost model. Furthermore, we only used 33 features to build the model, which is less than other research needs and different from other features.

To ensure the model's accuracy, we used 92 CNS + to verify performance after obtaining our best model independently. Finally, we found that the accuracy rate was as high as 0.96. Its performance was also excellent in the training set, with ACC equal to 0.95, AUC equal to 0.98, sensitivity similar to 0.983 and specificity equal to 0.93.

Our model can screen the drug candidates obtained from the experiment, greatly reducing the cost and time consumption of drug development.

We are confident that our model developed from the research is helpful to predict whether drugs can penetrate the BBB. We can use our model to quickly and accurately predict all types of small molecular compounds with MW less than 1000 Da.

## ACKNOWLEDGMENT

The computations were partially performed at the Peng Cheng Laboratory and the Center for High-Performance Computing, Shanghai Jiao Tong University.

## REFERENCES

[1] H. van de Waterbeemd and E. Gifford, "ADMET in silico modelling: Towards prediction paradise?" *Nature Rev. Drug Discovery*, vol. 2, no. 3, pp. 192–204, Mar. 2003.

[2] A. Yan, H. Liang, Y. Chong, X. Nie, and C. Yu, "In-silico prediction of blood-brain barrier permeability," *Sar Qsar Environ. Res.*, vol. 24, no. 1, pp. 61–74, 2013.

[3] D.-Q. Wei, J.-F. Wang, C. Chen, Y. Li, and K.-C. Chou, "Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design," *Protein Peptide Lett.*, vol. 15, no. 1, pp. 27–32, Jan. 2008.

[4] D.-Q. Wei, S. Sirois, Q.-S. Du, H. R. Arias, and K.-C. Chou, "Theoretical studies of Alzheimer's disease drug candidate 3-[(2, 4-dimethoxy) benzylidene]-anabaseine (GTS-21) and its derivatives," *Biochem. Biophysical Res. Commun.*, vol. 338, no. 2, pp. 1059–1064, Dec. 2005.

[5] C. Merlot, "Computational toxicology—A tool for early safety evaluation," *Drug Discovery Today*, vol. 15, nos. 1–2, pp. 16–22, Jan. 2010.

[6] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P. W. Lee, and Y. Tang, "admetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties," *J. Chem. Inf. Model.*, vol. 52, no. 11, p. 3099, 2012.

[7] J.-F. Wang, D.-Q. Wei, and K.-C. Chou, "Drug candidates from traditional Chinese medicines," *Current Topics Medicinal Chem.*, vol. 8, no. 18, pp. 1656–1665, Dec. 2008.

[8] Z. Li, T. Zhang, H. Lei, L. Wei, Y. Liu, Y. Shi, S. Li, B. Shen, H. Guo, Z. Chen, X. Yi, and H. Zhang, "Research on gastric cancer's drug-resistant gene regulatory network model," *Current Bioinf.*, vol. 15, no. 3, pp. 225–234, May 2020, doi: 10.2174/1574893614666190722102557.

[9] S. A. Hitchcock, "Structural modifications that alter the P-glycoprotein efflux properties of compounds," *J. Medicinal Chem.*, vol. 55, no. 11, pp. 4877–4895, Jun. 2012.

[10] B. Shaker *et al.*, "LightBBB: Computational prediction model of blood–brain-barrier penetration based on LightGBM," *Bioinformatics*, 2020, doi: 10.1093/bioinformatics/btaa918.

[11] J. Bicker, G. Alves, A. Fortuna, and A. Falcão, "Blood–brain barrier models and their relevance for a successful development of CNS drug delivery systems: A review," *Eur. J. Pharmaceutics Biopharmaceutics*, vol. 87, no. 3, pp. 409–432, Aug. 2014.

[12] J. A. Lenhart, X. Ling, R. Gandhi, T. L. Guo, P. M. Gerk, D. H. Brunzell, and S. Zhang, "'Clicked' bivalent ligands containing curcumin cholesterol as multifunctional A$\beta$ oligomerization inhibitors: Design, synthesis, and biological characterization," *J. Medicinal Chem.*, vol. 53, no. 16, pp. 6198–6209, 2010.

[13] N. G. Grois, B. E. Favara, G. H. Mostbeck, and D. Prayer, "Central nervous system disease in Langerhans cell histiocytosis," *Hematol./Oncol. Clinics North Amer.*, vol. 12, no. 2, pp. 287–305, Apr. 1998.

[14] N. J. Abbott, L. RöNnbäCk, and E. Hansson, "Astrocyte–endothelial interactions at the blood–brain barrier," *Nature Rev. Neurosci.*, vol. 7, no. 1, pp. 41–53, 2006.

[15] H. Tang, L. Z. Zhao, H. T. Zhao, S. L. Huang, S.-M. Zhong, J.-K. Qin, Z.-F. Chen, Z.-S. Huang, and H. Liang, "Hybrids of oxoisoaporphine-tacrine congeners: Novel acetylcholinesterase and acetylcholinesterase-induced $\beta$-amyloid aggregation inhibitors," *Eur. J. Medicinal Chem.*, vol. 46, no. 10, pp. 4970–4979, 2011.

[16] T. Zhang, Q. Chen, L. Li, L. Angela Liu, and D.-Q. Wei, "In silico prediction of cytochrome P450-mediated drug metabolism," *Combinat. Chem. High Throughput Screening*, vol. 14, no. 5, pp. 388–395, Jun. 2011.

[17] A. Munir, S. I. Malik, and K. A. Malik, "Proteome mining for the identification of putative drug targets for human pathogen *clostridium tetani*," *Current Bioinf.*, vol. 14, no. 6, pp. 532–540, Jul. 2019, doi: 10.2174/1574893613666181114095736.

[18] A. A. Toropov, A. P. Toropova, M. Beeg, M. Gobbi, and M. Salmona, "QSAR model for blood-brain barrier permeation," *J. Pharmacol. Toxicol. Methods*, vol. 88, pp. 7–18, Nov. 2017.

[19] Z. Chen, P. Zhao, F. Li, Y. Wang, A. I. Smith, G. I. Webb, T. Akutsu, A. Baggag, H. Bensmail, and J. Song, "Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences," *Briefings Bioinf.*, vol. 21, no. 5, pp. 1676–1696, Sep. 2020, doi: 10.1093/bib/bbz112.

[20] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, "Cancer diagnosis through IsomiR expression with machine learning method," *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, Feb. 2018, doi: 10.2174/1574893611666160609081155.

[21] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2013.

[22] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[23] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. Workshop Empirical Methods Artif. Intell. (IJCAI)*, 2001, vol. 3, no. 22, pp. 41–46.

[24] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*, 2016, pp. 785–794.

[26] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, nos. 14–15, pp. 2627–2636, Aug. 1998.

[27] D. A. Konovalov, D. Coomans, E. Deconinck, and Y. V. Heyden, "Benchmarking of QSAR models for blood-brain barrier permeation," *J. Chem. Inf. Model.*, vol. 47, no. 4, p. 1648, 2007.

[28] A. A. Toropov, A. P. Toropova, M. Beeg, M. Gobbi, and M. Salmona, "QSAR model for blood-brain barrier permeation," *J. Pharmacol. Toxicol. Methods*, vol. 88, pp. 7–18, Nov. 2017.

[29] I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, "A Bayesian approach to in *silico* blood-brain barrier penetration modeling," *J. Chem. Inf. Model.*, vol. 52, no. 6, pp. 1686–1697, Jun. 2012.

[30] J. S. Yazdi, F. Kalantary, and H. S. Yazdi, "Investigation on the effect of data imbalance on prediction of liquefaction," *Int. J. Geomech.*, vol. 13, no. 4, pp. 463–466, Aug. 2013.

[31] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 1322–1328.

[32] J. Cox, C. A. Harper, and A. de Waard, "Optimized machine learning methods predict discourse segment type in biological research articles," in *Semantics, Analytics, Visualization*. Cham, Switzerland: Springer, 2017, pp. 95–109.

[33] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling," *Soft Comput.*, vol. 15, no. 10, pp. 1909–1936, Oct. 2011.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[35] W. Zhuang, H. Yang, Z. Wu, T. Wang, W. Li, Y. Tang, and G. Liu, "In silico prediction of blood–brain barrier permeability of compounds by machine learning and resampling methods," *ChemMedChem*, vol. 13, no. 20, pp. 2189–2201, 2018.

[36] I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, "A Bayesian approach to in silico blood-brain barrier penetration modeling," *J. Chem. Inf. Model.*, vol. 52, no. 6, pp. 1686–1697, Jun. 2012.

[37] M. Muehlbacher, G. M. Spitzer, K. R. Liedl, and J. Kornhuber, "Qualitative prediction of blood–brain barrier permeability on a large and refined dataset," *J. Comput.-Aided Mol. Des.*, vol. 25, no. 12, pp. 1095–1106, Dec. 2011.

[38] J. Shen, F. Cheng, Y. Xu, W. Li, and Y. Tang, "Estimation of ADME properties with substructure pattern recognition," *J. Chem. Inf. Model.*, vol. 50, no. 6, pp. 1034–1041, Jun. 2010.

[39] W. Wang, M. T. Kim, A. Sedykh, and H. Zhu, "Developing enhanced blood–brain barrier permeability models: Integrating external bio-assay data in QSAR modeling," *Pharmaceutical Res.*, vol. 32, no. 9, pp. 3055–3065, Sep. 2015.

[40] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466–1474, May 2011.

[41] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Aug. 1985.

[42] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sens. Actuators B, Chem.*, vol. 212, pp. 353–363, Jun. 2015.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Oct. 2011.

[44] J. Tolles and W. J. Meurer, "Logistic regression: Relating patient characteristics to outcomes," *Jama*, vol. 316, no. 5, p. 533, Aug. 2016.

[45] L. Zhang, Y. He, H. Song, X. Wang, N. Lu, L. Sun, and H. Liu, "Elastic net regularized softmax regression methods for multi-subtype classification in cancer," *Current Bioinf.*, vol. 15, no. 3, pp. 212–224, May 2020, doi: 10.2174/1574893613666181112141724.

[46] H. L. Harney, "Bayes' theorem," in *Bayesian Inference: Data Evaluation and Decisions*. Cham, Switzerland: Springer, 2016, pp. 11–25.

[47] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.

[48] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[49] Y.-M. Dong, J.-H. Bi, Q.-E. He, and K. Song, "ESDA: An improved approach to accurately identify human snoRNAs for precision cancer therapy," *Current Bioinf.*, vol. 15, no. 1, pp. 34–40, Feb. 2020, doi: 10.2174/1574893614666190424162230.

[50] W. Wang, Q. Dai, F. Li, Y. Xiong, and D.-Q. Wei, "MLCDForest: Multi-label classification with deep forest in disease prediction for long non-coding RNAs," *Briefings Bioinf.*, Jun. 2020, doi: 10.1093/bib/bbaa104.

[51] Y. Chu, A. C. Kaushik, X. Wang, W. Wang, Y. Zhang, X. Shan, D. R. Salahub, Y. Xiong, and D.-Q. Wei, "DTI-CDF: A cascade deep forest model towards the prediction of drug-target interactions based on hybrid features," *Briefings Bioinf.*, Dec. 2019, doi: 10.1093/bib/bbz152.

[52] F. Shi, Y. Yao, Y. Bin, C.-H. Zheng, and J. Xia, "Computational identification of deleterious synonymous variants in human genomes using a feature-based approach," *BMC Med. Genomics*, vol. 12, no. S1, p. 12, Jan. 2019, doi: 10.1186/s12920-018-0455-6.

[53] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, "AIPpred: Sequence-based prediction of anti-inflammatory peptides using random forest," *Frontiers Pharmacol.*, vol. 9, p. 276, Mar. 2018, doi: 10.3389/fphar.2018.00276.

[54] Z. Yue, X. Chu, and J. Xia, "PredCID: Prediction of driver frameshift indels in human cancer," *Briefings Bioinf.*, Jun. 2020, doi: 10.1093/bib/bbaa119.

[55] K. Li, S. Zhang, D. Yan, Y. Bin, and J. Xia, "Prediction of hot spots in protein–DNA binding interfaces based on supervised isometric feature mapping and extreme gradient boosting," *BMC Bioinf.*, vol. 21, no. S13, p. 381, Sep. 2020, doi: 10.1186/s12859-020-03683-3.

[56] T. Chen, X. Wang, Y. Chu, Y. Wang, M. Jiang, D.-Q. Wei, and Y. Xiong, "T4SE-XGB: Interpretable sequence-based prediction of type IV secreted effectors using eXtreme gradient boosting algorithm," *Frontiers Microbiol.*, vol. 11, Sep. 2020, Art. no. 580382, doi: 10.3389/fmicb.2020.580382.

[57] J. Y. Hesterman, L. Caucci, M. A. Kupinski, H. H. Barrett, and L. R. Furenlid, "Maximum-likelihood estimation with a contracting-grid search algorithm," *IEEE Trans. Nucl. Sci.*, vol. 57, no. 3, pp. 1077–1084, Jun. 2010.

[58] B. Gandek, J. E. Ware, N. K. Aaronson, G. Apolone, J. B. Bjorner, J. E. Brazier, M. Bullinger, S. Kaasa, A. Leplege, L. Prieto, and M. Sullivan, "Cross-validation of item selection and scoring for the SF-12 health survey in nine countries," *J. Clin. Epidemiol.*, vol. 51, no. 11, pp. 1171–1178, Nov. 1998.

[59] L. Kuang, H. Zhao, L. Wang, Z. Xuan, and T. Pei, "A novel approach based on point cut set to predict associations of diseases and LncRNAs," *Current Bioinf.*, vol. 14, no. 4, pp. 333–343, Apr. 2019, doi: 10.2174/1574893613666181026122045.

[60] X. Shan, X. Wang, C.-D. Li, Y. Chu, Y. Zhang, Y. Xiong, and D.-Q. Wei, "Prediction of CYP450 enzyme–substrate selectivity based on the network-based label space division method," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4577–4586, Nov. 2019, doi: 10.1021/acs.jcim.9b00749.

[61] S. Zhang, L. Zhao, C.-H. Zheng, and J. Xia, "A feature-based approach to predict hot spots in protein–DNA binding interfaces," *Briefings Bioinf.*, vol. 21, no. 3, pp. 1038–1046, May 2020, doi: 10.1093/bib/bbz037.

[62] B. Manavalan, S. Subramaniyam, T. H. Shin, M. O. Kim, and G. Lee, "Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy," *J. Proteome Res.*, vol. 17, no. 8, pp. 2715–2726, Aug. 2018, doi: 10.1021/acs.jproteome.8b00148.

[63] M. M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, and B. Manavalan, "HLPpred-fuse: Improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation," *Bioinformatics*, vol. 36, no. 11, pp. 3350–3356, Jun. 2020, doi: 10.1093/bioinformatics/btaa160.

[64] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, "A multimodal deep learning framework for predicting drug–drug interaction events," *Bioinformatics*, vol. 36, no. 15, pp. 4316–4322, Aug. 2020, doi: 10.1093/bioinformatics/btaa501.

[65] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, and X. Zhang, "SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions," *PLOS Comput. Biol.*, vol. 14, no. 12, Dec. 2018, Art. no. e1006616, doi: 10.1371/journal.pcbi.1006616.

[66] J. Kang, Y. Fang, P. Yao, N. Li, Q. Tang, and J. Huang, "NeuroPP: A tool for the prediction of neuropeptide precursors based on optimal sequence composition," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 11, no. 1, pp. 108–114, Mar. 2019, doi: 10.1007/s12539-018-0287-2.
[67] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.-Q. Wei, "PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers Microbiol.*, vol. 9, p. 2571, Oct. 2018, doi: 10.3389/fmicb.2018.02571.
[68] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA sequence information," *BMC Bioinf.*, vol. 19, no. 1, p. 306, Aug. 2018, doi: 10.1186/s12859-018-2321-0.
[69] T.-H. Zhang and S.-W. Zhang, "Advances in the prediction of protein subcellular locations with machine learning," *Current Bioinf.*, vol. 14, no. 5, pp. 406–421, Jun. 2019, doi: 10.2174/1574893614666181217145156.
[70] P. Zhang, J. Meng, Y. Luan, and C. Liu, "Plant miRNA–lncRNA interaction prediction with the ensemble of CNN and IndRNN," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 12, no. 1, pp. 82–89, Mar. 2020, doi: 10.1007/s12539-019-00351-w.
[71] J. F. B. Lissabet, L. H. Belén, and J. G. Farias, "PPLK+C: A bioinformatics tool for predicting peptide ligands of potassium channels based on primary structure information," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 12, no. 3, pp. 258–263, Sep. 2020, doi: 10.1007/s12539-019-00356-5.
[72] X. Du, X. Li, W. Li, Y. Yan, and Y. Zhang, "Identification and analysis of cancer diagnosis using probabilistic classification vector machines with feature selection," *Current Bioinf.*, vol. 13, no. 6, pp. 625–632, Nov. 2018, doi: 10.2174/1574893612666170405125637.
[73] N. Cheng, M. Li, L. Zhao, B. Zhang, Y. Yang, C.-H. Zheng, and J. Xia, "Comparison and integration of computational methods for deleterious synonymous mutation prediction," *Briefings Bioinf.*, vol. 21, no. 3, pp. 970–981, May 2020, doi: 10.1093/bib/bbz047.
[74] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "MAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation," *Bioinformatics*, vol. 35, no. 16, pp. 2757–2765, Aug. 2019, doi: 10.1093/bioinformatics/bty1047.
[75] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "AtbPpred: A robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees," *Comput. Struct. Biotechnol. J.*, vol. 17, pp. 972–981, Jan. 2019, doi: 10.1016/j.csbj.2019.06.024.

**ZHIWEN SHI** is currently pursuing the master's degree with the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His research interests include bioinformatics and computer-aided drug design.

**YANYI CHU** is currently pursuing the Ph.D. degree with the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. She works on drug discovery through machine learning methods.

**YONGHONG ZHANG** received the M.S. degree in analytical chemistry and the Ph.D. degree in biomedical engineering from Chongqing University, China. She has studied, for one year, as an Academic Scholar with the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, and with the Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, USA. She is currently an Associate Professor with Chongqing Medical University, China. Her research interest includes computational-aided drug analysis and evaluation.

**YANJING WANG** received the Ph.D. degree in biology from Shanghai Jiao Tong University, Shanghai, China, in 2019. She is currently a Postdoctoral Fellow with the Department of Bioinformatics and Biological Statistics, Shanghai Jiao Tong University. Her research interests include bioinformatics analysis of cancer genomics, computer-aided drug design, computational structural biology, and machine learning.

**DONG-QING WEI** received the Ph.D. degree in chemical physics from the University of Puerto Rico, Rio Piedras, Puerto Rico, USA, in 1987. He is currently a Distinguished Professor with the Department of Bioinformatics and Biostatistics, College of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. He does research in bioinformatics and biostatistics, protein-protein interactions and networks, molecular machines, proteins-drug interactions and drug designs, membrane protein dynamics, deep learning and precise medicine, CYP450, and personalized medicine. He is the Editor-in-Chief of *Interdisciplinary Sciences: Computational Life Sciences*.