Symbiosis

www.symbiosisonlinepublishing.com
ISSN Online: 2474-9257

Research Article Journal of Computer Science Applications and Information Technology Open Access

Gradient boosting machine-based model for predicting hERG K+ channel inhibitory activities

Kazuaki Ishihara[1*], Hideko Sone[1], Kohsuke Hayamizu[1]

[1]Yokohama University of Pharmacy

Received: January 18, 2022; Accepted: February 8, 2022; Published: February 15, 2022

***Corresponding author:** KohsukeHayamizu, Yokohama University of Pharmacy,601 Matano-cho, Totsuka-ku, Yokohama, Kanagawa 254-0066, Japan. E-mail: k.hayamizu@hamayaku.ac.jp*

Abstract

Inhibition of the human-ether-a-go-go-related gene (hERG) channel by small-molecule compounds induces prolongation of the QT interval and an arrhythmia called torsades de pointes, resulting in significant cardiac toxicity. Therefore, it is important to evaluate the effects of hERG in the early stage of the drug discovery process. In recent years, hERG-related bioactivity information has been extracted from public databases, and machine learning has been used to develop predictive models of hERG inhibition. However, many of these models predicting hERG inhibition have the risk of biased accuracy in sensitivity and specificity, and overestimation of accuracy and ROC-AUC due to an imbalance between on inhibitors and non-inhibitors in the data set. In this study, we obtained the hERG assay information from an integrated database consisting of four public databases on cardiotoxicity and constructed the model. To examine the effect of inhibitor and non-inhibitor imbalance in the data set, we performed Borderline SMOTE, which is one of the oversampling methods. The model constructed by XGBoost, a type of gradient boosting machine (GBM), showed the highest predictive accuracy of 0.887, 0.806, and 0.930 for sensitivity, specificity, and ROC-AUC, respectively, among the single algorithms. Furthermore, the ensemble model with the top five ROC-AUC models showed corresponding values of 0.898, 0.832, and 0.939, respectively. To interpret the characteristics of hERG inhibition, the importance of molecular descriptors was calculated using Gain and SHAP values. We propose a model of hERG inhibition with high prediction accuracy using XG-Boost.

Introduction

The human-ether-à-go-go-related gene (hERG) encodes a potassium channel that is one of the most critical components associated with QT interval prolongation and arrhythmia called torsades de pointes [1,2]. Several drugs, including astemizole, terfenadine, vardenafil, cisapride, and sertindole, have been withdrawn from the market because of the inhibition of hERG [3-7]. Therefore, hERG is a key target as part of preclinical toxicity screening in drug discovery. In addition, not only chemical drugs but also natural chemical substances have recently been reported to inhibit hERG. For example, Orvos et al. reported that berberine, chelidonine, and sanguinarine contained in *Chelidonium majus* have a significant hERG potassium channel blocking effect [8]. These extracts and alkaloids also prolong the cardiac action potential in dog ventricular muscle [8]. In addition, the extract of *Evodia rutaecarpa*, a crude drug widely used in traditional Chinese medicine, was found to inhibit hERG channels. The compounds in *Evodia* extract exerting inhibitory activity were identified as dehydroevodiamine (DHE) and hortiamine by patch clamp studies [9]. These cases have indicated the importance of applying the hERG assay not only for chemicals as candidates for drug discovery, but also for any chemicals to which humans are exposed. hERG blocking assays, performed *in vivo* and *in vitro*, are conducted by various methods, such as rubidium-flux assay, fluorescence-based assay, electrophysiological measurement, and radioligand binding assay [10]. Recently, a patch clamp assay and a cardiotoxicity assay using iPS cell-derived cardiomyocytes have been developed to monitor the effects on hERG [11]. While these techniques are tractable when the sample size is small, their use for assaying thousands of compounds is prohibitively time-consuming and expensive. An *in silico* assay has been proposed as one method for resolving these issues. There are two approaches for performing *in silico* assays of hERG blocking. The first involves structure- and ligand-based methods [12,13]. These methods employ models based on the statistically generated 2D or 3D chemical structure poses of small compounds, and structure-based approaches conducting docking simulations using a 3D structure of hERG modeled on a computer. The most important part of this approach is the information on the hERG structure and the electron microscopy structure of hERG, as reported in 2017 [14]. The related information was subsequently reviewed in detail [15]. However, docking simulations with hERG are still difficult due to the high flexibility of this molecule. The second type of approach is based on the quantitative structure–activity relationship (QSAR). QSAR refers to the quantitative relationship between the structure of a chemical substance and its biological (pharmaceutical or toxicological) activity. The purpose of this method is to predict the "bioactivity or toxicity" of structurally similar compounds. The basic concept of QSAR is a multivariate function with the descriptor of a compound of a small-molecule compound as an independent variable and the assay result as a dependent variable. In 2006, Seierstad and Agrafiotis reported a QSAR model for predicting the pIC50 of hERG binding by a multiple linear regression model using a dataset of 439 compounds [16]. In the classified model, Wang et al. proposed a binary classification model in which the compounds were defined as "inhibitor" or "non-inhibitor," for the prediction of hERG blocking by naïve Bayes classifiers using

 ***Corresponding author email:** k.hayamizu@hamayaku.ac.jp*

a dataset of 806 molecules [17], and Shen et al. reported a QSAR model for binary classification using support vector machine (SVM) with a dataset of 1,668 compounds [18]. SVM is a kernel-based supervised machine learning technique that is well suited for the separation of compounds into two classes based on their biological end-point measure [19]. In 2015, Braga reported Pred-hERG, a web-accessible computational tool for predicting hERG inhibitors [20,21]. To construct the classification model, they used a dataset of 5,984 compounds and calculated descriptors by Morgan fingerprints and Chemistry Development Kit (CDK) using RDkit[22]and the PaDEL Descriptor plugin for KNIME [23]. Finally, they built a consensus model using each descriptor and proposed as the Pred-hERG system. In addition, in 2017, Xiao et al. reported consensus models that were based on the results of multiple models with different prediction methods using a dataset of 4,813 compounds [24]. First, they calculated simple models using methods with different descriptor generations, such as k nearest neighbor (kNN), associative neural networks (ASNN), SVM, fast stepwise multivariate linear regression (FSMLR), and partial least squares (PLS). Then, they developed two consensus models using those models. Moreover, in 2019, Ogura et al. integrated the information from multiple public databases on cardiotoxicity to create a dataset of more than 291,000 compounds [25]. Using this big data, they reported improved performance of the predictive model by selecting the descriptors that are explanatory variables for the model using an algorithm called Non-dominated Sorting Genetic Algorithm-II (NSGA-II). They reported the performance of the SVM classification model using 72 descriptors and the ECFR_4 fingerprint in terms of the accuracy and kappa statistics, which were 0.984 and 0.733, respectively. However, the sensitivity of their predicted model was 0.67, so the issue of false negatives remained to be addressed.

Against this background, in recent years, there has been an explosion of bioactivity data on hERG inhibition in public databases, and statistical models using machine learning technology have been improved. In addition, technological developments for machine learning such as various classification models and descriptor selection algorithms are being developed.

The general challenges of prediction models that perform positive/negative classification include the problem of data size and the problem of an imbalanced dataset. The purpose of the prediction model is to reduce the generalization error, that is, to increase the predictive accuracy for unknown data. When less data is available, the training error is smaller and the generalization error is larger. In addition, an imbalance of the data set can lead to overlearning of positive or negative, resulting in an imbalance in sensitivity and specificity. The purpose of the hERG inhibition prediction model is to detect positive substances, but the ultimate goal is to develop a model with precise predictive accuracy in terms of both sensitivity and specificity.

In this study, we developed a new hERG classification model, based on XGBoost, a type of gradient boosting machine (GBM), and an ensemble model with stacking multiple prediction models from an integrated database. To avoid overfitting of the prediction model, the descriptor set was optimized by the Boruta feature selection algorithm [26]. To address the imbalance of the training data, we used Borderline SMOTE, which is one of the oversampling methods, and confirmed the effect of an imbalance of the training data on predictive accuracy. Finally, we calculated the importance of molecular descriptors that contribute to the prediction of hERG inhibition.

Methods

Computed

The software programs used to build the prediction model were as follows: Python 3.7.6 [27], Anaconda 4.9.2 [28], Jupyter Notebook 6.2.0 [29], Boruta_py 0.3 [30], Hyperopt 0.2.5 [31], Keras 2.4.3 [32], Mordred 1.2.0 [33], Py-xgboost 0.90 [34], RDkit 2020.09.1.0 [22], Scikit-learn 0.24.2 [35], and Tensorflow 2.3.0 [36].

Dataset

Data on hERG inhibition were extracted from four databases related to cardiotoxicity: CheMBL, GOSTAR, the NIH Chemical Genomics Center (NCGC), and the hERGCentral integrated dataset [37]. From the hERG assay dataset containing a total of 291,219 compounds, we extracted data that provide clear classification as an inhibitor (IC50 < 10 μM) or a non-inhibitor (IC50 ≥ 10 μM). Consequently, we selected a dataset with a total of 13,944 compounds, which consisted of 8,325 inhibitors and 5,669 non-inhibitors. We used this data set to build a hERG inhibition prediction model. The dataset was randomly split into a training set (85%, 7,076 inhibitors/4,818 non-inhibitors) and a test set (15%, 1,249 inhibitors/851 non-inhibitors) for external validation (Table 1).

Table 1: Statistics of chemicals in the training and test sets

Data sets	Total	Inhibitors	Non-inhibitors
Whole data set	13994	8325	5669
Training set	11894	7076	4818
Test set	2100	1249	851

Calculation of molecular descriptors

To build the hERG inhibition prediction model, we used two types of chemical descriptor generator: Mordred [33] and RDkit[22]. Using Mordred and RDkit, 1,613 and 208 descriptors were calculated, respectively.

Molecular descriptor selection

Feature selection is a very important process in the preprocessing stage of model building.

Reducing the number of features is expected to improve the interpretability of the model, reduce the calculation cost, avoid overfitting, and improve the accuracy [38]. In this study, missing values were imputed using the average value of each feature, and the molecular descriptors corresponding to the following two items were excluded:

1) Molecular descriptor whose data type is a nominal variable.

2) Molecular descriptor with variance less than 0.1.

Furthermore, the descriptor selection base on the Boruta algorithm [26] was applied to maximize the predictive performance with the minimal number of descriptors. The Boruta algorithm runs Random Forest multiple times to collect Z-scores and compare the importance of actual features with meaningless fake shadow features. Each run copies the actual feature and shuffles it against the feature to create a shadow feature with the non-correlation with the objective variable. The importance of each feature is calculated by Random Forest upon combining the actual feature and the shadow feature. For each feature, the true important feature is selected by performing a statistical test between the importance of the actual feature and the shadow feature with the highest importance. Upon optimizing the number of molecular descriptors using Boruta, 294 and 110 descriptors were adopted from Mordred and RDkit, respectively. These molecular descriptors were standardized and used for model construction.

QSAR model building

In this study, we used five machine learning algorithms, Naïve Bayes (NB), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and single-layer Artificial Neural Network (ANN). SVM, NB, and RF were performed using Scikit-learn [35], XGBoost was performed using Py-xgboost[34], and ANN was performed using Tensorflow[36]and Keras[32]. After building discriminant models with each machine learning algorithm, an ensemble model by stacking methods was constructed. Stacking is one of the ensemble methods for constructing one meta-classifier by combining multiple weak classifiers. The method is performed using each prediction of the model constructed by the machine learning method as a feature quantity, and an ensemble predicted value is calculated. In this study, we used the predicted values of the top five models assessed using ROC-AUC as features, performed XGBoost, and output the ensemble predicted values for the prediction of hERG inhibition. For robustness of the developed models, optimization of the hyperparameters was conducted using Hyperopt[31], with 70 iterations and 5-fold cross-validation.

Evaluation of predictive performance

To evaluate the model performance, we used five indexes: correct answer rate (ACC), sensitivity (SEN), specificity (SPE), precision rate (PRE), and Matthews correlation coefficient (MCC) [39]. These evaluation indexes were calculated using four terms: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP and FN denote the numbers of known inhibitors predicted to be active and inactive, while TN and FP are the numbers of known non-inhibitors predicted to be inactive and active, respectively. The indexes of model performance were calculated as follows.

$$ACC = (TP+TN)/(TP+FP+FN+TN) \quad (1)$$

$$SEN = TP/(TP+FN) \quad (2)$$

$$SPE = TN/(TN+FP) \quad (3)$$

$$PRE = TP/(TP+FP) \quad (4)$$

$$MCC = (TP\times TN-FP\times FN)/\sqrt{((TPgtt+FP)(TP+FN)(TN+FN)(TN+FP))} \quad (5)$$

The performance of each model was also measured by the ROC score, defined as the area under the receiver operating curve (ROC-AUC), which plots the ratio of true positives on the axis of false positive fractions and ranges from 0 to 1. The closer the value is to 1, the better the model accuracy, while 0.5 represents random prediction.

Borderline SMOTE (Synthetic Minority Oversampling Technique)

Since the ratio of inhibitor to non-inhibitor of the data set used for model construction is approximately 1.5:1, the prediction model may overestimate the inhibitory compounds. Therefore, using Borderline SMOTE, one of the oversampling methods [40], we verified the data imbalance by equalizing the ratio of inhibitors to non-inhibitors.

Feature importance

SHAP (SHapley Additive exPlanations) and Gain were used to calculate the importance of the explanatory variables [41-44]. SHAP is a calculation method based on cooperative game theory of how much each player contributed in a game. In machine learning, the player is replaced with an explanatory variable to show how much it contributed to the prediction of the model. Gain is an indicator of the importance of a feature in XGBoost.

Results

Chemical diversity analysis

Prior to comparing the performances of different models, we verified the diversity of the chemical space of the data sets. A large chemical space, to a certain extent, reflects the effectiveness of the application domain of the models [45]. In this study, the chemical space was analyzed using principal component analysis (PCA) [46] with descriptors generated by Modred (294 descriptors) and RDkit (110 descriptors) (Figure 1). From PCA,

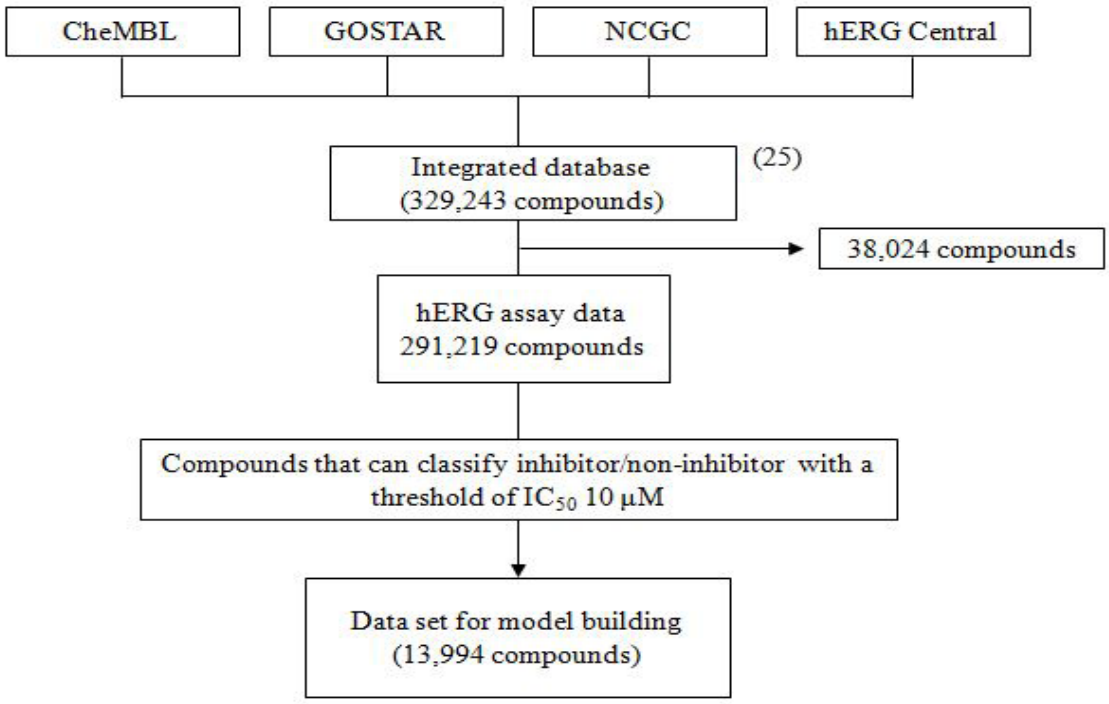

Figure 1 : Flowchart of dataset selection

the chemical space is demonstrated to be sufficiently diverse, and it is considered that the training data and the test data have similarchemical spaces.

Predictive performance of constructed models

In this study, two molecular descriptor calculation modules, Mordred and RDkit, and five machine learning methods, SVM, NB, RF, XGBoost, and ANN, were used to construct a hERG inhibition prediction model. The performance of each model was assessed using stratified split test data, consisting of 1,249 inhibitors and 851 non-inhibitors (total of 2,100 compounds) (Table 2). The reference Models 1 (NB -Mordred) and 2 (NB -RDkit) showed ACC: 0.671/0.699, SEN: 0.657/0.756, SPE: 0.615/0.691, PRE: 0.742/0.757, MCC: 0.342/0.372, and ROC-AUC: 0.730/0.751, respectively. Model 6 (XGBoost -RDkit) showed the best fitted performance (ACC: 0.854, SEN: 0.887, SPE: 0.806, PRE: 0.870, MCC: 0.697, ROC-AUC: 0.930). Subsequently, we contracted an ensemble model by stacking methods (Model 11) using Models 3, 4, 5, 6, and 9, which are the top five models in terms of ROC-AUC. The predictive performance of Model 11 was as follows: ACC: 0.871, SEN: 0.898, SPE: 0.832, PRE: 0.887, MCC: 0.732, and ROC-AUC: 0.939. These results showed that the predictive accuracy of the ensemble model was higher than that of the model constructed by a single algorithm for all of the evaluation indexes.

To confirm the effect of data set imbalance on the predictive accuracy, the training data were verified by Borderline SMOTE. Model 6 showed the largest change in ROC-AUC, but the change was 0.011, which was very small (Table 3). This was interpreted to indicate that there was no overfitting of inhibitory compounds due to dataset imbalance.

Importance of molecular descriptors

To understand the relationship between compound structure and hERG inhibition, we evaluated the importance of the molecular descriptors selected for model building. The importance of the molecular descriptors was calculated using Gain and SHAP values for Model 6, which is the most applicable model (Figure 4). The top five molecular descriptors based on Gain value were MinEStateIndex, MaxEStateIndex, BalabanJ, MolLogP, and PEOE_VSA8, while those based on SHAP value were VSA_EState2, MolLogP, TPSA, fr_C_O, and BalabanJ. MinEStateIndex, MaxEStateIndex, and VSA_Estate2, which are Estate Index (Electro topological-state Index) types, represent both the electrical and the topological properties of each skeletal atom in a molecule. The Estate Index is formulated as the intrinsic value I and the perturbation term Al arising from the electronic interaction of each atom in the molecule in the molecular topological environment [47]. SlogP_VSA represents the hydrophilic and hydrophobic interactions with respect to the receptor, SMR_VSA the polarizability of the

Table 2: Performance of hERG inhibitor/non-inhibitor classification models									
Model	Algorithm	Descriptor	ACC	SEN	SPE	PRE	MCC	Kappa	ROC-AUC
1	Naïve Bayes	Mordred	0.671	0.657	0.691	0.757	0.342	0.338	0.73
2	Naïve Bayes	RDkit	0.699	0.756	0.615	0.742	0.372	0.372	0.751
3	SVM	Mordred	0.857	0.874	0.831	0.883	0.703	0.703	0.921
4	SVM	RDkit	0.861	0.883	0.828	0.883	0.712	0.712	0.924
5	XGBoost	Mordred	0.839	0.878	0.781	0.855	0.663	0.663	0.918
6	XGBoost	RDkit	0.854	0.887	0.806	0.87	0.697	0.696	0.93
7	RF	Mordred	0.792	0.869	0.678	0.799	0.563	0.559	0.862
8	RF	RDkit	0.82	0.893	0.714	0.821	0.624	0.62	0.888
9	ANN	Mordred	0.847	0.878	0.803	0.867	0.682	0.682	0.92
10	ANN	RDkit	0.842	0.834	0.853	0.893	0.679	0.677	0.915
11	Stacking	-	0.871	0.898	0.832	0.887	0.732	0.732	0.939

Table 3: Performance of hERG inhibitor/non-inhibitor classification models with Borderline SMOTE									
Model	Algorithm	Descriptor	ACC	SEN	SPE	PRE	MCC	Kappa	ROC-AUC
1’	Naïve Bayes	Mordred	0.65	0.58	0.751	0.774	0.327	0.313	0.735
2’	Naïve Bayes	RDkit	0.694	0.703	0.682	0.764	0.379	0.377	0.754
3’	SVM	Mordred	0.849	0.866	0.824	0.878	0.688	0.688	0.919
4’	SVM	RDkit	0.856	0.869	0.837	0.886	0.702	0.702	0.919
5’	XGBoost	Mordred	0.833	0.84	0.823	0.874	0.657	0.656	0.923
6’	XGBoost	RDkit	0.833	0.837	0.826	0.876	0.658	0.657	0.919
7’	RF	Mordred	0.773	0.744	0.817	0.856	0.551	0.544	0.867
8’	RF	RDkit	0.8	0.778	0.831	0.871	0.599	0.594	0.89
9’	ANN	Mordred	0.838	0.833	0.845	0.887	0.67	0.668	0.92
10’	ANN	RDkit	0.834	0.832	0.838	0.883	0.662	0.661	0.907

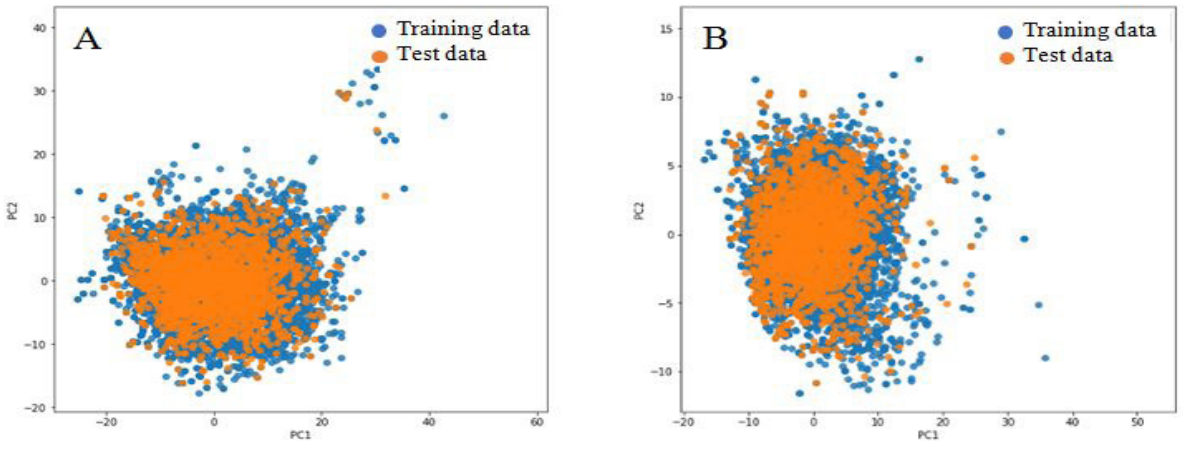

Figure 2: Chemical space defined by molecular descriptors calculated with Mordred and RDkit in the dataset.
(A) Mordred descriptor, (B) RDkit descriptor

molecule, and PEOE_VSA the direct electrostatic interaction [48]. These descriptors are the sum of the contributions of the atomic van der Waals surface area (VSA) with atomic components for a certain range of properties (SlogP, SMR, and PEOE). TPSA is the sum of the polar areas of the molecular surface area inferred from intramolecular binding patterns. TPSA has been shown to correlate well with passive molecular transport through a membrane [49]. BalabanJ represents a topological index based on the sum of distances [50] and fr_C_O represents the number of carbonyl groups. Figure 5 shows violin plots of the top five inhibitory/non-inhibitory molecular descriptors calculated based on Gain and SHAP values.

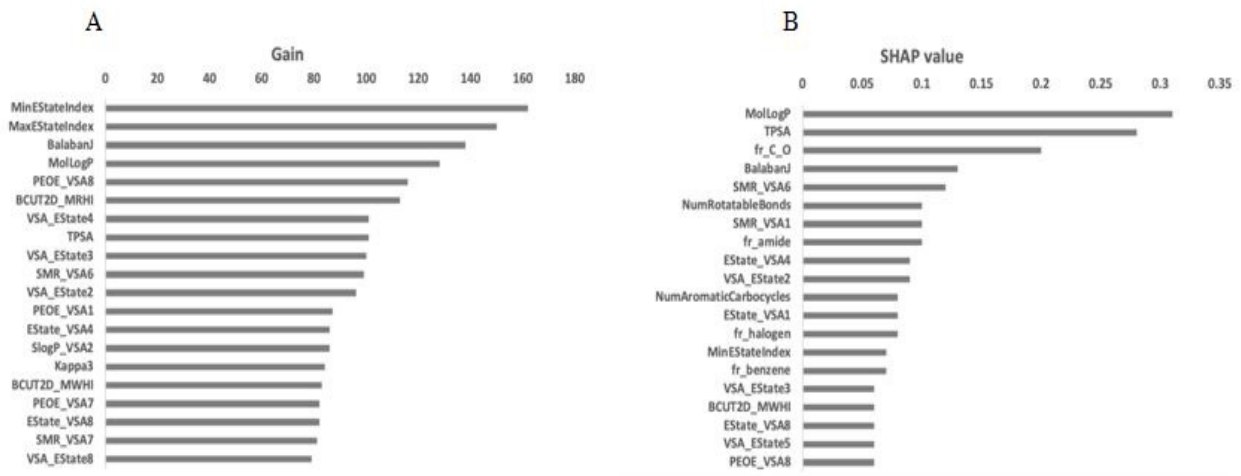

Figure 3: Importance of molecular descriptors (A) Gain, (B) SHAP

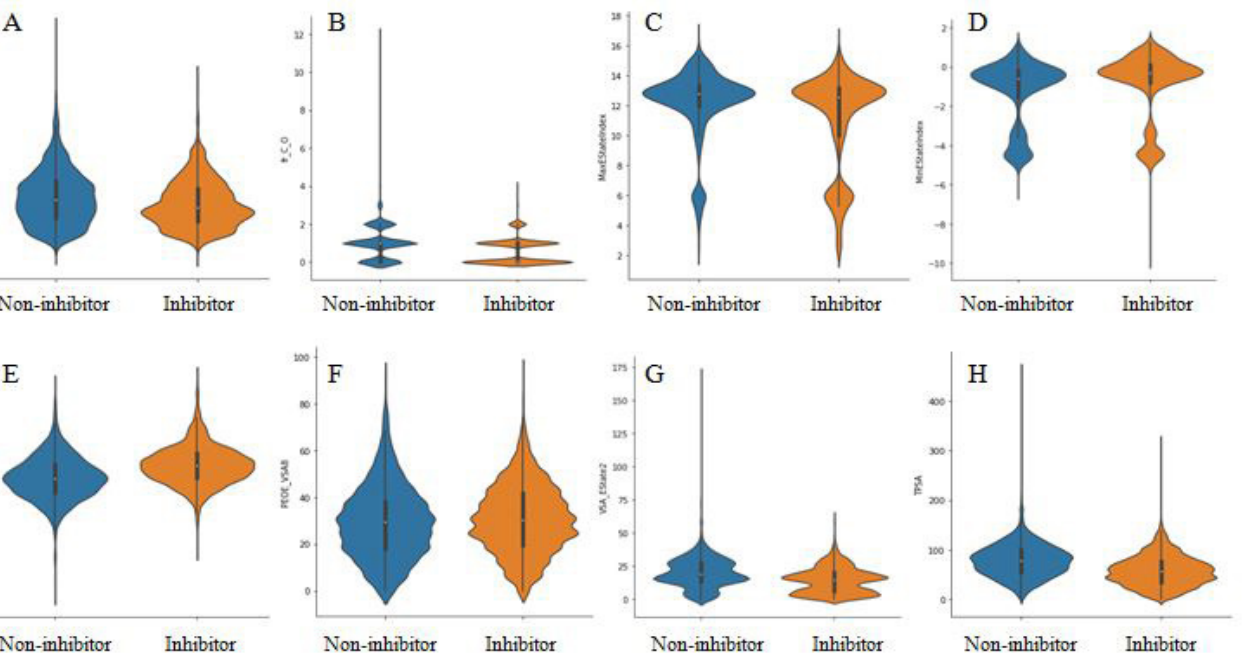

Figure 4: Violin plot of molecular descriptors in the top five for feature importance. (A) BalabanJ, (B) fr_C_O, (C) MaxEStateIndex, (D) MinEStateIndex, (E) MolLogP, (F) PEOE_VSA8, (G) VSA_EState2, (H) TPSA

Results

In this study, we constructed 10 models for predicting hERG inhibition using two molecular descriptor calculation modules (Mordred and RDkit) and five machine learning methods (SVM, NB, RF, XGBoost, and ANN). We proposed a prediction model using XGBoost as the optimal model (Model 6) with the chemical descriptor generated by RDkit. Here, let us consider the effects of two types of molecular descriptor calculation module selection on the predictive performance of the model. Comparing values of ROC-AUC as an index of model performance, when the model was built with NB, SVM, and XGBoost, the RDkit-based model gave ROC-AUC values 0.021, 0.003, 0.012, and 0.026 higher than those of the Mordred-based model, respectively. On the other hand, when built with ANN, the Mordred-based model (Model 9) gave an ROC-AUC value 0.005 higher than that of the RDkit-based model (Model 10). Although they were small, there were differences in performance due to the difference in the generator of the chemical descriptor. Regarding the reason why RDkit-based models had higher ROC-AUC values in NB, SVM, XGBoost, and RF, it is speculated that Mordred generated a much larger number of molecular descriptors than RDkit, so it could not adequately address multicollinearity between independent variables. On the other hand, in ANN, it is speculated that the reason why Mordred had a higher ROC-AUC value is that, although the number of molecular descriptors was large, the independent variables that were truly correlated with the objective variable were extracted as weighting and hyperparameters. In this way, it is important to deal with multicollinearity in the preprocessing of model construction, and it can be understood that appropriate preprocessing affects the performance of the constructed model. In this study, we ran the Boruta algorithm for preprocessing, but in the future, we plan to try additional algorithms, such as Non-dominated Sorting Genetic algorithm-II (NSGA-II)[51].

In addition, the stacking ensemble model was able to achieve more accurate predictions. To interpret the prediction model with chemical theory, we assessed the importance of selected molecular descriptors in Model 6 by ranking them using SHAP and Gain values.

The hERG-inhibitory pharmacophore model of the charged ligand shows that there are at least three aromatic/hydrophobic sites and charged moieties. In the case of the hERG-inhibitory pharmacophore model of the neutral ligand, the intramolecular aromatic (and/or hydrophore) contains at least one hydrogen bond acceptor separated from the group at a distance of 3.5 Å [52].Thus, the presence of hydrophobic groups plays an important role in hERG inhibition. In this study, MolLogP and TPSA were selected as molecular descriptors related to hydrophobicity. In our study, MolLogP tended to be larger and TPSA tended to be smaller in the inhibitor, which is consistent with previously reported results. Ogura et al. reported hERG inhibitory activity and the involvement of atomic charge [25].In our model, PEOE_VSA8 was also selected as being associated with atomic charge and results similar to those in the above-mentioned report were obtained.

In this study, we used Gain and SHAP values to interpret the model, but other approaches such as using Partial Dependence and ICE plot may also be effective [44,53].

In addition, in the calculation of the importance of features, the multicollinearity of the explanatory variables is not considered in this study. These issues should be addressed in future work.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Redfern WS, Carlsson L, Davis AS, Lynch WG, MacKenzie I, Palethorpe S, et al. Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development. Cardiovasc Res. 2003;58(1):32-45.doi: 10.1016/s0008-6363(02)00846-5.
2. Sanguinetti MC and Tristani-Firouzi M. hERG potassium channels and cardiac arrhythmia. Nature.2006;440(7083):463-469.doi:10.1038/nature 04710.
3. Taglialatela M, Castaldo P, Pannaccione A, Giorgio G, Genovese A, Marone G, et al. Cardiac ion channels and antihistamines: possible mechanisms of cardiotoxicity. Clin Exp Allergy. 1999;29 Suppl 3:182-189.doi: 10.1046/j.1365-2222.1999.0290s3182.x.
4. Woosley RL. Cardiac actions of antihistamines. Annu Rev Pharmacol Toxicol. 1996;36:233-252.doi: 10.1146/annurev.pa.36.040196.001313.
5. Rampe D, Roy ML, Dennis A, Brown AM. A mechanism for the proarrhythmic effects of cisapride (Propulsid): high affinity blockade of the human cardiac potassium channel HERG. FEBS Lett. 1997;417(1):28-32.doi: 10.1016/s0014-5793(97)01249-0.
6. Rampe D, Murawsky MK, Grau J, Lewis EW. The antipsychotic agent sertindole is a high affinity antagonist of the human cardiac potassium channel HERG. J Pharmacol Exp Ther. 1998;286(2):788-793.
7. Alvarez PA and Pahissa J. QT alterations in psychopharmacology: proven candidates and suspects. Curr Drug Saf. 2010;5(1):97-104. doi: 10.2174/157488610789869265.
8. Orvos P, Virág L, Tálosi L, Hajdú Z, Csupor D, Jedlinszki N, et al. Effects of Chelidonium majus extracts and major alkaloids on hERG potassium channels and on dog cardiac action potential - a safety approach. Fitoterapia. 2015;100:156-165. doi: 10.1016/j.fitote.2014.11.023.
9. Baburin I, Varkevisser R, Schramm A, Saxena P, Beyl S, Szkokan P, et al. Dehydroevodiamine and hortiamine, alkaloids from the traditional Chinese herbal drug Evodia rutaecarpa, are IKr blockers with proarrhythmic effects in vitro and in vivo. Pharmacol Res. 2018;131:150-163.doi: 10.1016/j.phrs.2018.02.024.
10. Polak S, Wiśniowska B, Brandys J. Collation, assessment and analysis of literature in vitro data on hERG receptor blocking potency for subsequent modeling of drugs' cardiotoxic properties. J Appl Toxicol. 2009;29(3):183-206.doi: 10.1002/jat.1395.
11. Gintant G, Sager PT, Stockbridge N. Evolution of strategies to

improve preclinical cardiac safety testing. Nat Rev Drug Discov. 2016;15(7):457-471. doi: 10.1038/nrd.2015.34.

12. Wang S, Li Y, Xu L, Li D, Hou T. Recent developments in computational prediction of HERG blockage. Curr Top Med Chem. 2013;13(11):1317-1326.doi: 10.2174/15680266113139990036.
13. Villoutreix BO and Taboureau O. Computational investigations of hERG channel blockers: New insights and current predictive models. Adv Drug Deliv Rev. 2015;86:72-82. doi: 10.1016/j.addr.2015.03.003.
14. Wang W and MacKinnon R. Cryo-EM Structure of the Open Human Ether-à-go-go-related K+Channel hERG. Cell. 2017;169(3):422-430. e10.doi: 10.1016/j.cell.2017.03.048.
15. Butler A, Helliwell MV, Zhang Y, Hancox JC, Dempsey CE. An update on the structure of hERG. Front Pharmacol 2020;10:1572.doi: 10.3389/fphar.2019.01572.
16. Seierstad M and Agrafiotis DK. A QSAR Model of hERG Binding Using a Large, Diverse, and Internally Consistent Training Set. Chem Biol Drug Des.2006;67(4):284-296. doi: 10.1111/j.1747-0285.2006.00379.x.
17. Wang S, Li Y, Wang J, Chen L, Zhang L, Yu H, et al. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. Mol Pharm. 2012;9(4):996-1010.doi: 10.1021/mp300023x.
18. Shen MY, Su BH, Esposito EX, Hopfinger AJ, Tseng YJ. A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets. Chem Res Toxicol 2011; 24(6): 934-949.doi:10.1021/tx200099j.
19. Van Gestel T, Suykens JA, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J. Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel Fisher discriminant analysis. Neural Comput.2002;14(5):1115-1147. doi:10.1162/089976602 753633411.
20. Braga RC, Alves VM, Silva MF, Muratov E, Fourches D, Lião LM, et al. Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. Mol Inform. 2015;34(10):698-701.doi: 10.1002/minf.201500040.
21. Pred-hERG 4.2
22. RDkit: Open-Souce Cheminformatics Software
23. Mazanetz MP, Marmon RJ, Reisser CB, Morao I. Drug discovery applications for KNIME: an open source data mining platform. Curr Top Med Chem. 2012;12(18):1965-1979.doi: 10.2174/156802612804910331.
24. Li X, Zhang Y, Li H, Zhao Y. Modeling of the hERG K+ Channel Blockage Using Online Chemical Database and Modeling Environment (OCHEM). Mol Inform. 2017;36(12).doi: 10.1002/minf.201700074.
25. Ogura K, Sato T, Yuki H, Honma T. Support Vector Machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II. Sci Rep 2019; 9(1): 12220. doi:10.1038/s41598-019-47536-3.
26. Kursa M and Rudnicki W. Feature Selection with the Boruta Package. Journal of Statistical Software.2010;36(11):1-13.doi:10.18637/jss.v036.i11.
27. Python
28. https://www.anaconda.com
29. https://jupyter.org
30. https://github.com/scikit-learn-contrib/boruta_py/blob/master/boruta/boruta_py.py
31. https://github.com/hyperopt/hyperopt
32. https://keras.io/
33. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform 2018; 10(1): 4.doi:10.1186/s13321-018-0258-y.
34. https://xgboost.readthedocs.io/en/latest/
35. http://scikit-learn.org/stable/
36. www.tensorflow.org
37. Sato T, Yuki H, Ogura K, Honma T. Construction of an integrated database for hERG blocking small molecules. PLoS One. 2018;13(7): e0199348.doi: 10.1371/journal.pone.0199348.
38. Kohavi R and John GH. Wrappers for Feature Subset Selection. Artificial Intelligence 1997; 97(1-2): 273-324.
39. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405(2): 442-451.doi:10.1016/0005-2795(75)90109-9.
40. Han H, Wang WY, Mao BH, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Proceedings of the 2005 international conference on Advances in Intelligent Computing 2005; 1: 878-887.
41. Surdararajan M and Arir N. The many Shapley values for model explanation. Proceedings of Machine Learning 2020;119.
42. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. NIPS 2017; 17: 4768-4777.
43. Friedman J, Hastie T, Tibshirani R. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics 2000; 28(2): 337–407.doi:10.1214/aos/1016218223.
44. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann. Statist.2001;29(5): 1189-1232.doi: 10.1214/aos/1013203451.
45. Cai C, Wu Q, Luo Y, Ma H, Shen J, Zhang Y, et al. In silico prediction of ROCK II inhibitors by different classification approaches. Mol Divers. 2017;21(4):791-807.doi: 10.1007/s11030-017-9772-5.
46. Ma S and Dai Y. Principal component analysis based methods in bioinformatics studies. Brief Bioinform. 2011;12(6):714-22. doi:10.1093/bi b/bbq090.
47. Hall LH, Mohney B, Kier LB. The electrotopological state: an atom index for QSAR 1991; 10(1): 43-51.doi:10.1002/qsar.19910100108.
48. Labute P. A widely applicable set of descriptors. J Mol Graph Model.2000;18(4-5):464-77.doi: 10.1016/s1093-3263(00)00068-1.
49. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 2000; 43(20): 3714-3717.doi:10.1021/jm000942e.
50. Balaban AT. Highly discriminating distance-based topological index 1982; 89(5): 399-404.doi:10.1016/0009-2614(82)80009-2.
51. Wang NN, Dong J, Deng YH, Zhu MF, Wen M, Yao ZJ, et al. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. J Chem

Inf Model. 2016;56(4):763-773.doi: 10.1021/acs.jcim.5b00642.

52. Kalyaanamoorthy S and Barakat KH. Development of Safe Drugs: The hERG ChallengeMed Res Rev. 2018;38(2):525-555.doi:10.1002/med.21445.

53. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 2015; 24(1): 44-65.doi: 10.1080/10618600.2014.907095.