Iranian Journal of Analytical Chemistry

مجله ایرانی شیمی تجزیه

Volume 11, Issue 1, March 2024 (44-53) Original Research Article دوره ۱۱، شماره ۱، فروردین۱۴۰۳ (۵۳-۴۴)

Examining the Impact of Variable Selection Methods on Classification Outcomes of BCL-2 and BCL-XL Isoform-Selective Ligands

Marzieh Sadat Neiband

Department of Chemistry, Payam Noor University, 19395-4697, Tehran, Iran

Received: 12 July 2024 **Accepted: 15** August 2024
DOI: 10.30473/ijac.2024.70820.1290

Abstract

Feature selection is crucial in Quantitative Structure-Activity Relationship (QSAR) studies, enhancing learning algorithms' performance and reducing computational costs. This study evaluates the impact of eight variable selection methods on the classification of isoform-selective ligands for Bcl-2 and Bcl-xL targets using three machine learning techniques: Supervised Kohonen Network (SKN), Support Vector Machine (SVM), and Partial Least Squares Discriminant Analysis (PLS-DA). Classification models were assessed using confusion matrix parameters, 10-fold Venetian blind cross-validation, and test sets.

The results show that PLS-DA and SVM have comparable classification capabilities, outperforming SKN. However, PLS-DA occasionally leaves some ligands unassigned, making SVM a more robust and efficient choice. Despite using different variable selection methods, no clear advantage was found for any specific method, with all achieving around 70% classification accuracy in validation and test series. This suggests that the choice of variable selection method does not consistently affect outcomes across all techniques.

Ensuring the reliability of selected variables involves meticulous data quality assessments, literature review, and robust cross-validation. Eliminating redundant features is essential for accurate classification models, as many physicochemical properties may be irrelevant to target bioactivity. While no single method guarantees superior models, selecting important variables is vital for extracting relevant features. This study highlights the importance of careful variable selection in QSAR studies, emphasizing its role in reducing dimensionality and improving model interpretability. Ultimately, this enhances drug discovery efficiency by identifying safer and more effective compounds, reducing time and cost.

Keywords

Variable Selection Methods; QSAR; Drug Design; Bcl-2; Bcl-$x_{L.}$

1.INTRODUCTION

In the QSAR studies, variable selection methods are crucial for identifying the most relevant features or descriptors that contribute significantly to the classification model's statistical performance. Feature selection refers to the process of selecting a subset of features suitable for used in quantitative models. The main purpose of using feature selection methods is to eliminate irrelevant and redundant features from the feature set. Irrelevant features are those that do not provide useful information, and therefore eliminating them has no impact on modeling efficiency [1]. Redundant features are a set of irrelevant features that already provided by another feature [2]. Since the redundant features do not provide any additional information, eliminating them and keeping one of them does not affect the efficiency of the training and the performance of the model. In fact, the elimination of redundant and irrelevant features does not cause any problem in terms of the information obtained, but their existence increases the computational cost of building mathematical models. Nowadays, variable selection holds significant importance in numerous studies due to the varying levels of in informativeness among variables within collected datasets (such as data from data mining, virtual screening, and genomic studies). This is especially true for for drug design studies, with the availability of a wide range of molecular descriptor computing tools (such as Dragon software). Consequently, the demand for employing feature selection techniques has substantially heightened in this field.

When the most relevant and informative features have been selected in drug design studies, the

* Corresponding author:
M. Sadat Neiband; E-mail: m.neiband@pnu.ac.ir & neiband.mrs@gmail.com

model becomes interpretable and leads to better performance in predicting the target property or activity. Moreover, using variable selection techniques offer**s** additional benefits such as increasing transferability and generalization of models for their wider applicability and minimizing the risk of multicollinearity. Consequently, the insights derived from these models are valuable in understanding the underlying mechanisms and in designing new compounds with the desired properties or activities. They can aid pharmacists in the development of new drugs or the modification of existing ones, leading to improved therapeutic outcomes.

Considering the above explanations, in this work, eight commonly used variable selection techniques in QSAR studies were investigated. These techniques were Variable Importance in Projection (VIP), Feature Selection by concave minimisation (FSV), ReliefF, B2 and B4 algorithms, Non-iterative B2, Particle Swarm Optimisation (PSO) and Ant Colony Optimisation (ACO). The impact of these methods was examined on the statistical results of the classification of Bcl-2 and Bcl-x_L isoform selective ligands. ~~In~~ our previous research emphasized the importance of identifying key structural features of these inhibitors due to the toxicity associated with dual inhibition of these proteins [3]. The evaluation of the constructed classification models was done based on statistical measures derived from the confusion matrix. These measures included Sensitivity, Specificity, Precision and Non-error rate percentage, which were calculated for the training, validation, and test sets. Accuracy and Matthews Correlation Coefficients (MCC) **were** also calculated to measure overall classification performance. Moreover, the efficacy of the developed models was investigated by evaluating their predictive power using the test set.

The results obtained in this work demonstrate an approximate accuracy of 70% in the classification models developed using three different machine learning techniques and eight variable selection methods in the test datasets. These findings reveal that the sole utilization of a variable selection method does not significantly influence ~~on~~ the statistical outcomes of the classification models created. These observations have been previously documented in several articles [4-13]. For example, in a study conducted by H. Kaneko, it was reported that even when variables unrelated to the target were selected, regression models with good accuracy could still be constructed. This suggests that variable selection methods might not always have a significant impact on predictive performance [10]. R. Davronov and her colleague S. Kushmuratov compared several feature selection methods, including Chi-square, Mutual Information, and Recursive Feature Elimination [11]. Their analysis showed that different methods could yield varying results, but no single method consistently outperformed others across all datasets [11]. Priyanka De et al. reviewed various validation tools for QSAR models, emphasizing the importance of descriptor selection. Their study highlighted that while variable selection is crucial for model reliability, the choice of validation tools and dataset characteristics also play significant roles in overall performance [12]. S. Kausar and A. O. Falcao developed an automated framework for QSAR model building, focusing on data curation, variable selection, and validation. Their results indicated that feature selection significantly reduced prediction errors and increased the percentage of variance explained (PVE) by about 49% compared to models without feature selection [13]. These studies indicate that while variable selection methods can enhance QSAR model performance, their impact may vary depending on the specific context and dataset. Some research suggests that the choice of variable selection method might not always lead to significant differences in statistical results, highlighting the importance of considering other factors such as model interpretability and biological relevance.

This study presents the first comprehensive comparison of the impact of various variable selection methods on the statistical outcomes of classification models for selective inhibitors of Bcl-2 and Bcl-x_L proteins. The findings offer valuable insights for researchers engaged in computational drug design. Both the variable selection methods and machine learning techniques employed in this study are recognized as some of the most reliable and widely used in the field of drug design. Additionally, our previous work has underscored the significance of exploring the structure-activity relationship of selective Bcl-2 and Bcl-x_L inhibitors. This research aims to further enhance the understanding and application of these methodologies in drug design.

In this study, we also concluded that the choice of variable selection method does not directly influence the statistical outcomes. Notably, satisfactory statistical results can be achieved even when using descriptors that are not relevant to the activity-structure relationship of the investigated models. However, this finding is misleading and incorrect, as it fails to offer researchers reliable guidance for drug design and synthesis. Numerous parameters play a crucial role in the creation of a precise and trustworthy QSAR model.

While comparing statistical parameters is a good initial step, it is imperative to consider multiple factors, including performance metrics, stability, computational cost, expert knowledge,

DOI: 10.30473/ijac.2024.70820.1290

reproducibility, and validation. These factors are essential for selecting the most suitable variable selection method for your QSAR investigation.

2.EXPERIMENTAL

2.1. Datasets

This study involved the acquisition of two distinct datasets, consisting of 485 and 235 ligands of active Bcl-2 and Bcl-x$_L$ inhibitors, downloaded from the Binding DataBase [14] as 3D-SD files. This database includes a comprehensive listing of ligands alongside their corresponding biological activity, typically quantified in terms of IC$_{50}$ values expressed in nanomolar units. The Open Babel software version 2.4.1 [15] was utilized to generate the 3D structures with optimal structural energy from SDF files. The process of ligand preparation consisted of the following steps:

1- Hydrogen atoms were added to the downloaded SD files.

2-3D molecular structures were generated.

3-The partial charges of the atoms were assigned by the Merck molecular force field (MMFF94).

4-Duplicate conformers were removed.

5-The 3D structures of the molecules were subjected to geometry optimization using the MMFF94 force field.

The energy optimization parameters, including a maximum number of steps of 2500 and the utilization of the steepest descent algorithm, were kept at their default values (**c**onvergence criteria were set at 10^{-6} kcal mol^{-1}).

6-The SD files were converted to sequentially numbered output files and .hin format.

Subsequently, the optimized 3D structures of the molecules in .hin format were imported into Dragon software (version 5.5) [16] to calculate molecular descriptors. For each compound 3224 molecular descriptors including 0D (atomic and molecular counts, molecular weight, and sum of atomic properties), 1D (fragment counts), 2D (topological descriptors) and 3D (geometric, atomic coordinates) descriptors were calculated.

In order to determine isoform-selective inhibitors for Bcl-2 and Bcl-x$_L$ targets, the IC$_{50}$ threshold value was employed. In this manner, molecules demonstrating an IC$_{50}$ value below 200 nM were considered as active inhibitors for each respective target. Subsequently, if a molecule showed activity against both targets, we used the selectivity factor parameter to identify it~~’s~~ as a highly selective inhibitor for one target. The relevant relationship of this selection is presented below.

$$\text{Selectivity Factor to } Bcl-2 = \frac{\mathrm{IC}_{50}\left(Bcl-2\right)}{\mathrm{IC}_{50}\left(Bcl-x_{L}\right)}$$

For instance, a compound exhibiting an IC$_{50}$ value that is tenfold greater against Bcl-2 than its IC$_{50}$ against Bcl-x$_L$ can be designated as a Bcl-2 specific inhibitor. Conversely, a compound displaying an IC$_{50}$ value that is ten times higher against Bcl-x$_L$ compared to its IC$_{50}$ against Bcl-2 can be classified as a Bcl-x$_L$ specific inhibitor.

According to the aforementioned criteria, two datasets were generated. The molecules were categorized into two distinct classes: class “1” denoted molecules that exhibited selectivity in inhibiting Bcl-2, whereas class “2” represented molecules that displayed selectivity in inhibiting Bcl-x$_L$. Consequently, a response matrix (denoted as y) was constructed, representing the biological activity (IC$_{50}$) of the molecules, in a 'binary' format ('1' indicating Bcl-2 selective and '2' indicating Bcl-x$_L$ selective molecules). This response matrix was then correlated with the molecular descriptors of the molecules. The datasets were created using Microsoft Excel and subsequently utilized as input in MATLAB software for the development of isoform-selective classification models.

In order to enhance the predictive power and interpretability of the models, a systematic approach was followed. Firstly, from the total calculated descriptors, a set of 450 descriptors was selected based on their simplicity of interpretation. Subsequently, descriptors exhibiting 90% similarity or constant values were eliminated. Moreover, in cases where two descriptors displayed a correlation exceeding 0.9, the descriptor with the highest correlation with all other descriptors was excluded from the data matrix.

Considering that molecular descriptors encompass a wide spectrum of numerical values, it was crucial to mitigate model biases and ensure equal contribution of variables in the data analysis. To achieve this, the remaining variables underwent preprocessing using the auto-scale method, which involved centering and variance scaling. This normalization technique served to level the playing field, reducing the influence of variable magnitude on model outcomes.

Ultimately, an input matrix with the dimensions of (485 + 235) × 211 was created for models development.

2.2. Model Validtion

The statistical evaluation of all models was assessed using several metrics derived from the confusion matrix. These include: sensitivity, specificity, precision, and no-nerror rate. Accuracy and MCC values were calculated to measure the overall performance of the classifiers. For more in-depth information regarding these statistical parameters, please refer to [3]. To ensure the stability and predictive capability of the SVM, SKN, and PLS-DA classification models generated, a tenfold Venetian blind cross-validation technique and test set were used. The data sets were randomly divided, with 70% allocated to calibration (training and validation)

DOI: 10.30473/ijac.2024.70820.1290

and 30% to the test sets. It is important to note that the molecules within the test set had no involvement in the selection of significant variables nor in the development of the models, including the cross-validation process.

All calculations were performed using MATLAB R2017b, FSlib-v5.2-2017 and Classification Toolbox.

2.3. Methods

This work is a comparative study of the results obtained from models constructed by variables selected from eight filter feature selection methods, namely the PSO and ACO algorithms based on collective intelligence optimization, B2, B2 without repetition, and B4 based on the score vector resulted from the PCA, ReliefF based on Euclidean distance, and the VIP based on weights of the variables in the PLS-DA method, and a wrapper feature selection method known as minimizing a concave function (FSV) by training a SVM classifier.

The objective of this study was to identify the most suitable subset of variables that yielded the highest score in accurately modeling the structure-target selection relationships. To determine the most optimal number of variables, a series of models consisting of 7 to 12 variables were constructed. Upon comparing the results, it was observed that the models with 10 variables in the validation series yielded superior results. This resulted in an input matrix of (720×10) dimensions. Then, the classification models were built using three machine learning techniques: PLS-DA, SKN and SVM.

A brief description of these techniques and eight variable selection methods can be found in the supplementary material.

3.RESULTS AND DISCUSSION

After removing the constant and redundant variables, the auto-scaled data was utilized in the variable selection algorithms mentioned earlier to determine the most optimal variables. The abbreviations for the selected descriptors in each method can be found in Table 1. Detailed information regarding the type and definition of the selected descriptors is provided in Tables S1-8, which are included in the supplementary material. Furthermore, more detailed explanations of the selected descriptors can be found at the bottom of their respective tables in the supplementary material.

The results in Table 1 indicate that the selected variables from different variable selection methods are often not the same. Variable selection methods use different algorithms and criteria to determine the most relevant and important variables for a given analysis. Therefore, these methods may prioritize different variables based on their individual characteristics, such as algorithmic differences, statistical considerations, model complexity, search patterns and the nature of the data. These variations in approaches and considerations make it important to carefully evaluate and compare different methods to understand their implications for model performance and interpretation. In this regard, to construct classification models that yield accurate results, the parameters of SVM, SKN and PLS-DA methods were optimized using the 10-fold venetian blind cross-validation technique.

Table 1. Abbreviated the selected descriptors using eight variable selection methods.

Methods	Abbreviation of descriptors
ACO	nHBonds, nArCONR2, S-110, HIC, nC, RBN, nCrq, O-059, RNCG, TPC
PSO	N-۰۷۵, FDI, T(O..S), qnmax, nArNHR, N-۰۷۱, STN, nS, Hy, T(N..Cl)
B2	nN, nHAcc, C-۰۰۲, nH, TPSA(Tot), nBM, nArX, H-047, T(O..O), S-110
NIB2	nPyrroles, C-012, ECC, T (N...O), RBF, nH, nCrq, TPC, nROH, nN$^+$
B4	ZM1, O-056, nAT, PCR, H-052, nArNR2, nN, C-006, nArOH, nArC=N
ReliefF	qnmax, nOHp, ASP, nR07, nCconj, L/Bw, nCt, C-007, ARR, C-003
FSV	GNar, C-011, T(O..O), Qpos, ECC, Cl-086, Qtot, C-015, nR=CRX, nCp
VIP	Qmean, nBnz, DDI, nArCOOH, nRCONR$_2$, nCar, nO, H$_3$D, RBF, nCL

The classification models were constructed using the SVM method with a Radial Basis Function (RBF) kernel. The values of the C and Gamma (λ) parameters were optimized using cross-validation and heat maps. Various **C** parameter values (0.1, 1, 10, 100, and 1000) were tested. Additionally, Gamma **(λ)** parameters in the range of 0.05, 0.07, 0.10, 0.14, 0.20, 0.28, 0.34, 0.40, 0.57, 0.80, 1.13, 1.60, 2.26, 3.20, 4.53, 6.40, and 9 were also considered.

DOI: 10.30473/ijac.2024.70820.1290

For a detailed explanation of the SVM method, please refer to the Supplementary m4Zaterial file. The optimal SVM parameters for the eight variable selection methods in the validation sets are provided in Table 2.

The outcomes achieved for the 8 variable selection techniques applied in the SVM-based models indicate that, although there is only a negligible numerical difference, the FSV method emerges as the most effective with a prediction accuracy of 0.737 in the test series. After that, the PSO, NIB2, B2, VIP and ReliefF techniques demonstrated promising outcomes, achieving classification accuracies exceeding 70%. Hence, it can be inferred that in the SVM approach, the aforementioned variable selection methods are likely to yield the most optimal results for classification purposes. Due to the fact that the statistical differences observed in the test dataset among these methods were insignificant.

Table 2. The statistical results of different variable selection methods in the SVM classification model for the training, Validation and test sets.

Variable selection method	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC	λ	C
Training								
ACO	0.849	0.862	0.926	0.855	0.853	0.702		
PSO	0.879	0.809	0.903	0.844	0.856	0.703		
B2	0.849	0.756	0.876	0.802	0.818	0.663		
NIB2	0.855	0.738	0.869	0.796	0.816	0.652		
B4	0.868	0.782	0.890	0.825	0.840	0.704		
ReliefF	0.884	0.822	0.910	0.853	0.863	0.721		
FSV	0.864	0.702	0.855	0.783	0.811	0.658		
VIP	0.781	0.916	0.949	0.848	0.825	0.698		
Validation								
ACO	0.781	0.596	0.796	0.688	0.719	0.542	1.6	10
PSO	0.833	0.689	0.844	0.761	0.786	0.563	3.2	10
B2	0.803	0.716	0.851	0.759	0.774	0.533	1.13	1
NIB2	0.816	0.711	0.851	0.763	0.781	0.612	1.6	1
B4	0.770	0.596	0.794	0.683	0.712	0.498	2.26	10
ReliefF	0.807	0.591	0.800	0.699	0.736	0.534	1.6	10
FSV	0.835	0.671	0.837	0.753	0.781	0.529	0.4	1
VIP	0.748	0.796	0.881	0.772	0.764	0.517	2.26	10
Test								
ACO	0.727	0.524	0.732	0.642	0.678	0.496		
PSO	0.782	0.567	0.803	0.726	0.735	0.532		
B2	0.754	0.663	0.825	0.714	0.724	0.471		
NIB2	0.759	0.678	0.833	0.727	0.729	0.479		
B4	0.694	0.449	0.691	0.639	0.665	0.419		
ReliefF	0.738	0.490	0.706	0.642	0.701	0.482		
FSV	0.798	0.631	0.814	0.719	0.737	0.522		
VIP	0.663	0.726	0.832	0.731	0.716	0.603		

DOI: 10.30473/ijac.2024.70820.1290

In the PLS-DA method, the optimal number of latent variables was selected by utilizing the statistical outcomes obtained from the validation series. Table 3 presents the optimal number of latent variables for the classification models constructed using the PLS-DA method with eight different variable selection techniques. For more detailed information about the PLS-DA method, please refer to the Supplementary file section.

The PLS-DA method demonstrated superior results when models were constructed using variables selected by the VIP method. Furthermore, alternative variable selection techniques, namely NIB2, ACO, PSO, B2, ReliefF, and B4, exhibited comparable performance to the VIP method regarding classification statistical results of the scrutinized ligands within the test datasets.

In Table 4, the SKN method involved the development of multiple models, each utilizing varying numbers of neurons (12, 14, 16, 18, 20, 22, and 24) and different training frequencies referred to as epochs (20, 30, 40, 50, 80, 100, 150, 200, 250, and 300). The analysis of the validation sets indicated that the most optimal outcomes were obtained by employing maps consisting of 16 × 16 neurons and conducting 50 training cycles. For additional details about the SKN method, please consult the Supplementary file section.

In the SKN method, similar to the PLS-DA approach, optimal outcomes were achieved by constructing models utilizing variables chosen via the VIP method. Furthermore, within the SKN model, the ACO, ReliefF, PSO, NIB2, and B2 techniques exhibited noteworthy classification accuracy, exhibiting marginal variances compared to the VIP method.

Table 3. The statistical results of different variable selection methods in the PLS-DA classification model for the training, Validation and test sets.

Variable selection method	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC	Not assigned	Number of latent variables
Training								
ACO	0.706	0.916	0.944	0.811	0.775	0.627	0	
PSO	0.754	0.800	0.884	0.777	0.769	0.614	0.01	
B2	0.715	0.876	0.921	0.795	0.768	0.609	0.01	
NIB2	0.761	0.840	0.906	0.800	0.787	0.631	0.03	
B4	0.660	0.920	0.944	0.790	0.746	0.596	0	
ReliefF	0.767	0.782	0.877	0.775	0.772	0.643	0.01	
FSV	0.664	0.849	0.899	0.757	0.725	0.613	0.01	
VIP	0.727	0.902	0.947	0.810	0.788	0.635	0	
Validation								
ACO	0.704	0.911	0.941	0.807	0.772	0.545	0	5
PSO	0.755	0.806	0.887	0.781	0.772	0.534	0.01	5
B2	0.713	0.887	0.928	0.800	0.770	0.528	0.01	3
NIB2	0.770	0.840	0.907	0.805	0.793	0.552	0.02	5
B4	0.666	0.910	0.938	0.788	0.746	0.527	0.01	2
ReliefF	0.759	0.784	0.878	0.772	0.767	0.524	0.01	4
FSV	0.678	0.821	0.885	0.749	0.725	0.519	0.02	2
VIP	0.726	0.898	0.943	0.807	0.796	0.552	0	5
Test								
ACO	0.685	0.883	0.925	0.776	0.744	0.512	0	
PSO	0.746	0.793	0.825	0.761	0.732	0.506	0.01	
B2	0.693	0.871	0.912	0.784	0.725	0.511	0.01	
NIB2	0.752	0.829	0.894	0.788	0.761	0.526	0.02	
B4	0.650	0.894	0.926	0.763	0.712	0.507	0	
ReliefF	0.738	0.766	0.865	0.742	0.721	0.503	0.01	
FSV	0.652	0.804	0.835	0.711	0.690	0.445	0.2	
VIP	0.708	0.863	0.929	0.775	0.762	0.537	0	

Table 4. The statistical results of different variable selection methods in the SKN classification model for the training, Validation and test sets.

Variable selection method	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC
Training						
ACO	0.901	0.680	0.851	0.791	0.828	0.703
PSO	0.910	0.690	0.856	0.799	0.837	0.716
B2	0.827	0.782	0.885	0.804	0.812	0.669
NIB2	0.842	0.769	0.880	0.805	0.818	0.671
B4	0.862	0.711	0.858	0.786	0.812	0.664
ReliefF	0.886	0.800	0.900	0.843	0.858	0.732
FSV	0.901	0.609	0.824	0.755	0.805	0.656
VIP	0.805	0.858	0.920	0.831	0.822	0.693
Validation						
ACO	0.805	0.640	0.819	0.722	0.750	0.585
PSO	0.818	0.578	0.797	0.698	0.734	0.545
B2	0.785	0.613	0.805	0.699	0.728	0.529
NIB2	0.805	0.578	0.794	0.691	0.730	0.537
B4	0.798	0.582	0.795	0.690	0.727	0.519
ReliefF	0.825	0.569	0.795	0.697	0.740	0.571
FSV	0.805	0.538	0.779	0.671	0.717	0.481
VIP	0.818	0.640	0.822	0.729	0.759	0.592
Test						
ACO	0.753	0.604	0.762	0.705	0.724	0. 687
PSO	0.741	0.551	0.737	0.672	0.712	0.659
B2	0.722	0.561	0.746	0.676	0.701	0.645
NIB2	0.755	0.533	0.731	0.667	0.703	0.632
B4	0.783	0.549	0.734	0.671	0.699	0.489
ReliefF	0.731	0.524	0.723	0.666	0.718	0.511
FSV	0.749	0.514	0.710	0.653	0.686	0.467
VIP	0.756	0.611	0.764	0.700	0.726	0.507

4.CONCLUSION

Upon analysis of the findings presented in Tables 2-4, it is evident that both the PLS-DA and SVM techniques possess comparable capabilities in accurately classifying the selective inhibitors of Bcl-2 and Bcl-x$_L$, surpassing the performance of the SKN method. However, it is crucial to note that the PLS-DA models have a small proportion of ligands that remain not-assigned for some variable selection methods. Consequently, the SVM method emerges as a more robust and efficient approach for constructing classification models.

Based on the results presented in Tables 2-4, there seems to be no clear advantage associated with using a specific variable selection method. All three learning techniques, along with the eight variable selection methods, achieved classification accuracies of approximately 70% in both the validation and test series. These results do not show any statistically significant differences,

DOI: 10.30473/ijac.2024.70820.1290

suggesting that the choice of a particular variable selection method does not yield consistent outcomes for all three machine learning techniques. Thus, merely considering the statistical results may not suffice in determining the optimal variable selection method.

In order to ensure the reliability of the selected variables and to gain a comprehensive understanding of their logical relationship with the biological activity of molecules, it is essential to consider other important parameters. Firstly, data quality assessment should be conducted meticulously, beginning with a thorough evaluation of the dataset's quality. This evaluation involves checking for errors, outliers, missing values, and inconsistencies. It is crucial to ensure that the dataset encompasses a diverse range of molecules and that the biological activities are accurately measured. Moreover, reviewing relevant literature becomes an important parameter as it allows researchers to acquire knowledge about significant structural features and biological functions of the related compounds. Robust cross-validation techniques should also be used to assess the performance of the QSAR model. It is important to test model performance on new datasets not initially considered. Applying the model to new compounds can verify that the selected variables remain consistently relevant for predicting activity. Ultimately, the performance of feature selection methods varies depending on the dataset and evaluation metric, emphasizing the need for careful consideration during the selection process.

Acknowledgembts

The results of this study indicate that the importance of choosing the right variable selection methods in QSAR studies cannot be overstated. The elimination of redundant features is crucial in constructing accurate classification models, as thousands of physicochemical properties may be related or irrelevant to the target bioactivity. This work is a comparative study of variable selection techniques on Bcl-2 and Bcl-x_L selective inhibitors, reveals that no single method alone can ensure superiority in classification models. Instead, the selection of the most important variables leads to the extraction of relevant features essential for biological responses, ensuring the acquisition of correct information and knowledge from the QSAR model. However, there is no general consensus on the use of a variable selection method, and the development of a suitable algorithm remains a subject of interest among researchers. By reducing dimensionality and improving model interpretability, feature selection plays a crucial role in building robust QSAR models. This in turn has significant societal value as it enhances the efficiency of drug discovery processes, by identifying potentially safer and more effective compounds while reducing the time and cost. Ultimately, the choice of variable selection method should be driven by a balance between the accuracy, interpretability and appropriateness of the method given the specific subject and requirements of the analysis.

REFERENCES

[1]M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, Choosing feature selection and learning algorithms in QSAR, *J. Chem. Inf. Model 54* (2014) 837-843.

[2]M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, Benchmarking Variable Selection in QSAR, *Mol. Inform. 31* (2012) 173–179.

[3]N. Georges, I. Mhiri, and I. Rekik, Alzheimer's disease Neuroimaging Initiative Identifying the best datadriven feature selection method for boosting reproducibility in classification tasks, *Pattern Recognition* 101 (2020) 1- 14.

[4]M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, *Nucleic Acids Res.* 44(D1) (2016) D1045–D1053.

[5]S. Goswami, and A. Chakraborty, An efficient feature selection technique for clustering based on a new measure of feature importance, *J. Intell. Fuzzy Syst.* 32(6) (2017) 3847–3858.

[6]A. Mani-Varnosfaderani, M. S. Neiband, and A. Benvidi, Identification of molecular features necessary for selective inhibition of B cell lymphoma proteins using machine learning techniques, *Mol. Divers.* 23 (2019) 55–73.

[7]A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, Dragon software: An easy approach to molecular descriptor calculations, *Match*, *56*(2) (2006) 237-248.

[8]M. W. Mwadulo, A Review on Feature Selection Methods For Classification Tasks, *Int. J. Comput. Appl. Technol. 5* (2016) 395-402.

[9]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Chem. inf. 3* (2011) 33.

[10]H. Kaneko, Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables, *Heliyon 7* (2021).

[11]R. Davronov, and S. Kushmuratov, Comparative analysis of QSAR feature selection methods, In *AIP Conference Proceedings* 3004 (2024).

DOI: 10.30473/ijac.2024.70820.1290

[12] P. De, S. Kar, P. Ambure, and K. Roy, Prediction reliability of QSAR models: an overview of various validation tools, *Arch. Toxicol.* 96 (2022) 1279-1295.

[13] S. Kausar, and A. O. Falcao, An automated framework for QSAR model building, *J. Chem. Inf. Comput. Sci.* 10 (2018) 1.

[14] I. Ponzoni, V. Sebastián-Pérez, C. Requena-Triguero, C. Roca, M. J. Martínez, F. Cravero, M. F. Díaz, J. A. Páez, R. G. Arrayás, J. Adrio, and N. E. Campillo, Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery, *Sci. Rep.* 7 (2017) 2403.

[15] J. Tang, S. Alelyani, and H. Liu, Feature selection for classification: A review, *Data Classification: Algorithms and Applications book* (2014) 37-64.

[16] L. Yu, and H. Liu, Efficient feature selection via analysis of relevance and redundancy, *J. Mach. Learn. Res. 5* (2004) 1205-1224.

DOI: 10.30473/ijac.2024.70820.1290

بررسی تاثیر روش های انتخاب متغیر بر نتایج طبقه بندی لیگاندهای ایزوفرم انتخابی Bcl-2 و Bcl-x$_L$

مرضیه سادات نی بند

بخش شیمی، دانشگاه پیام نور، تهران، ایران
* E-mail: m.neiband@pnu.ac.ir & neiband.mrs@gmail.com

تاریخ دریافت: ۲۲ تیر ۱۴۰۳ **تاریخ پذیرش:** ۲۵ مرداد ماه ۱۴۰۳

چکیده

انتخاب ویژگی‌ها در مطالعات رابطه کمّی ساختار–فعالیت (QSAR) بسیار مهم است، زیرا عملکرد الگوریتم‌های یادگیری را بهبود می‌بخشد و هزینه‌های محاسباتی را کاهش می‌دهد. این مطالعه تأثیر هشت روش انتخاب متغیر را بر طبقه‌بندی لیگاندهای ایزوفورم–انتخابی برای اهداف Bcl-2 و Bcl-xL با استفاده از سه تکنیک یادگیری ماشین: شبکه کوهونن نظارت‌شده(SKN) ، ماشین بردار پشتیبان (SVM) و تحلیل تفکیکی حداقل مربعات جزئی (PLS-DA) ارزیابی می‌کند. مدل‌های طبقه‌بندی با استفاده از پارامترهای ماتریس سردرگمی، اعتبارسنجی متقاطع ۱۰–تایی و مجموعه‌های آزمون ارزیابی شدند.

نتایج نشان می‌دهد که PLS-DA و SVM قابلیت‌های طبقه‌بندی مشابهی دارند و از SKN بهتر عمل می‌کنند. با این حال، PLS-DAگاهی برخی لیگاندها را بدون تخصیص باقی می‌گذارد، که SVM را به یک انتخاب قوی‌تر و کارآمدتر تبدیل می‌کند. با وجود استفاده از روش‌های مختلف انتخاب متغیر، هیچ مزیت واضحی برای هیچ روش خاصی یافت نشد و همه حدود ۷۰٪ دقت طبقه‌بندی را در سری‌های اعتبارسنجی و آزمون به دست آوردند. این نشان می‌دهد که انتخاب روش انتخاب متغیر به طور مداوم بر نتایج در تمام تکنیک‌ها تأثیر نمی‌گذارد.

اطمینان از قابلیت اطمینان متغیرهای انتخاب‌شده شامل ارزیابی دقیق کیفیت داده‌ها، مرور ادبیات و اعتبارسنجی متقاطع قوی است. حذف ویژگی‌های زائد برای مدل‌های طبقه‌بندی دقیق ضروری است، زیرا بسیاری از خواص فیزیکوشیمیایی ممکن است به فعالیت زیستی هدف مرتبط نباشند. در حالی که هیچ روش واحدی مدل‌های برتر را تضمین نمی‌کند، انتخاب متغیرهای مهم برای استخراج ویژگی‌های مرتبط حیاتی است. این مطالعه اهمیت انتخاب دقیق متغیرها در مطالعات QSARرا برجسته می‌کند و نقش آن را در کاهش ابعاد و بهبود تفسیر مدل‌ها تأکید می‌کند. در نهایت، این کارایی کشف دارو را با شناسایی ترکیبات ایمن‌تر و مؤثرتر افزایش می‌دهد و زمان و هزینه را کاهش می‌دهد.

کلید واژه ها

روش انتخاب متغیر، QSAR، طراحی دارو، Bcl-2، Bcl-x$_L$.

DOI: 10.30473/ijac.2024.70820.1290