A machine learning model for the prediction of drug permeability across the Blood-Brain Barrier: a comparative approach

Ralph Saber

Holy Spirit University of Kaslik <https://orcid.org/0000-0003-4770-8631>

Rami Mhanna

American University of Beirut

Sandy Rihana (✉ sandyrihana@usek.edu.lb)

<https://orcid.org/0000-0003-1191-1603>

Research article

Keywords: blood-brain barrier, classification, machine learning, genetic algorithm, feature selection, artificial intelligence, in silico modeling

Posted Date: May 28th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-29117/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Abstract

Background: Drug permeability across the blood-brain barrier (BBB) is a critical challenge for successful drug discovery which has led to multiple efforts to develop *in silico* predictive models. Most of the *in silico* models are based on the molecular descriptors of the drugs. In this work, we compare the ability of sequential feature selection and genetic algorithms in selecting the most relevant descriptors and hence enhancing the permeability prediction accuracy.

Methods: Five different classifiers were initially trained on a dataset using eight molecular descriptors. Then, sequential feature selection and genetic algorithms were performed separately and the same classifiers were trained using the descriptors chosen by each algorithm.

Results: The highest overall accuracy obtained without feature selection was 94.98%. This accuracy increased with sequential feature selection and genetic algorithms on multiple classifiers. However, the highest accuracy (96.23%) was obtained after performing genetic algorithm on the feature vector. Moreover, genetic algorithm with a fitness function based on the performance of a support vector machine led to an increase in the accuracy of all the tested classifiers unlike sequential feature selection.

Conclusions: The findings show that genetic algorithm is a more robust approach than sequential feature selection in choosing the most relevant molecular descriptors involved in the permeability across the blood-brain barrier. The results also highlight the importance of the polar surface area of drugs in crossing the BBB.

Background

The Blood-Brain Barrier (BBB) is a physiological barrier that maintains brain homeostasis by controlling the exchange of molecules between the blood and the brain [1]. Consequently, the BBB blocks the passage of multiple molecules towards the brain, including administered drugs. This is beneficial when the target of the drug resides outside the brain since it prevents undesirable drug interactions and the ensuing phenotypic side effects. However, in the case of drugs targeting central nervous system (CNS) diseases, transport across the BBB is mandatory [2]. Therefore, the ability of drug candidates to cross the BBB has to be studied by all pharmaceutical companies during drug discovery. In this context, numerous *in silico* BBB models have been implemented by researchers in order to predict the behavior of drugs across the barrier [3]. These predictive models can be used during the early phases of drug discovery, and hence allow companies to save time and money resulting from failed drug investigations. Two different types of *in silico* BBB models exist in the literature: binary models which aim at qualitatively predicting whether drugs cross the BBB (BBB+) or not (BBB-), and quantitative models which attempt to qualify the permeability of the barrier to a given drug by computing the logarithm of the ratio of the concentration of the drug in the brain to that in blood (logBB) or its penetration rate (PR) [3]. In this context, K. Raja et al. [4] proposed two different stepwise regression models, one for the prediction of logBB values and the other for PR values. Other quantitative models are reviewed in [3]. While such models assign specific logBB/PR

values for each drug, binary models have so far reached a higher prediction accuracy and provide a preliminary insight regarding the behavior of candidate drugs which is sufficient in early drug discovery stages. Predominantly, binarization of drug permeability across the BBB is performed by setting empirical thresholds to logBB values [5–9]. However, S. Kunwittaya et al. [6] have shown that varying logBB thresholds lead to a difference in the prediction accuracy. Therefore, binary BBB models based on logBB values are prone to biases introduced by the thresholds setting. On the other hand, Adenot and Lahana [10] introduced a dataset based on the activity of the drug in the CNS: if a drug is CNS active, then it is necessarily BBB+. However, some drugs can cross the BBB but still show no activity in the CNS. Even though finding BBB- drugs based on CNS activity is consequently a challenging task, CNS activity-based datasets require no threshold setting and hence do not introduce the previously mentioned biases.

Machine learning is ubiquitously applied in the case of binary BBB models. In this context, different types of classifiers were trained in the literature including Support Vector Machines (SVM) [6, 8, 11, 12], Linear Discriminant Analysis (LDA) [13], Artificial Neural Networks (ANN) [6] and Multi-Layer Perceptron (MLP) [8, 9], k-Nearest Neighbors (k-NN) [8], Decision Trees (DT) [6, 7] and Random Forests (RF) [5, 8, 9]. Other studies apply consensus models, by training and combining multiple classifiers [8, 9]. While consensus models mitigate the overfitting problem of single classifiers, they naturally require high computational power, especially when dealing with high dimensional data. The features used to train these classifiers are often molecular descriptors which are chemical properties describing the drugs [3]. Some studies also add the fingerprints of the molecules in order to reach better prediction [8, 9, 12]. On the other hand, novel approaches apply the drug side effects and indications for BBB penetration prediction [14]. The model achieved excellent prediction performance but relies on high-level phenotypes which prevent extraction of significant biological explanations concerning drug interaction with the BBB.

Molecular descriptors remain the staple of classification-based BBB models. However, until today, the high dimensionality of the data based on molecular descriptors is still challenging. The selection of the most relevant features is crucial since it guarantees an improved prediction performance on one hand, and a faster computation on the other; by reducing the size of feature vectors. In order to study the effect of the chosen features on the classification performance, Y. Yuan et al. [12] compared the performance of SVM models trained by feature vectors containing different molecular descriptors, fingerprints or a combination of both. Since trying all possible combinations of feature vectors dramatically increases the required computational time and power, an effective feature selection algorithm is needed. In this context, D. Zhang et al. [9] applied genetic algorithm (GA) for the selection of the appropriate features and optimization of SVM parameters. Nevertheless, choosing the most suitable algorithm for a given application is an important step since different algorithms may lead to convergence to different feature subsets and consequently affect the prediction results. This study hence, compares the effect of GA to that of the sequential feature selection (SFS) algorithm on different classifiers applied in the reported *in silico* BBB models.

Methods

The workflow including the use of feature selection is summarized in Fig. 1. In this study, we began by collecting the drug dataset. Then, in order to compare the performance of GA to that of SFS, we started by training and evaluating multiple classifiers without the application of any feature selection algorithm. Subsequently, the same classifiers were implemented while applying each algorithm separately. Finally, the performance of each classifier was evaluated individually.

Dataset preparation

We built and compared the models using a drug permeability dataset made publically available by Zhao et al. [15]. The dataset is composed of 1593 drugs: 1283 that cross the BBB (BBB+) and 310 that don't (BBB-). The authors used the previously described dataset of Adenot and Lahana [10]. For each drug, Zhao et al. [15] calculated a set of molecular descriptors that are also listed in the dataset.

Molecular descriptors as feature extraction

Based on the correlation study performed by Zhao et al. [15], 8 molecular descriptors were chosen in our work among the 19 descriptors reported in the dataset:

- The molecular weight (MW)
- The polar surface area (PSA)
- The octanol/water partition (logP)
- The number of hydrogen bond acceptors (HA)
- The number of hydrogen bond donors (HD)
- pKa (strongest acid)
- pKa (strongest base)
- The number of rotatable bonds (NRB)

Feature selection

In order to obtain a highly predictive model, feature engineering should be followed by feature selection as a means to choose the most relevant features. In fact, while some features hold exclusive information regarding the permeability of the drugs across the BBB, others are simply irrelevant or hold misleading information. In this work, GA and SFS algorithms were evaluated and compared separately on the same dataset.

Part 1: sequential feature selection

This is an iterative algorithm that aims at finding the optimal combination of predictors that lead to the best prediction capacity of a specific classifier. The algorithm may run in two opposite directions. On one hand, it can start with the entire input features set and iteratively remove features that mislead a predefined classifier, until reaching the predictors' subset leading to the best classification performance;

in this case, the algorithm is running in the backward direction. On the other hand, it may run in the forward direction by starting with an empty predictors' subset and successively adding features that would improve the classifier's predictive performance until reaching the optimal predictors' subset. The steps of the forward algorithm are summarized in the flowchart of Fig. 2. The algorithm starts by creating an empty feature subset. Then it randomly adds one feature to the subset and performs 10-fold cross-validation, which returns a criterion value expressing the loss of the classifier. In this work, the criterion used by the algorithm for each combination of features is the number of misclassified observations in the test set. Then, the previously selected feature is removed and a new feature is randomly added to the subset to find a new criterion value. Once all the features have been tried, the algorithm chooses the feature with the least criterion value as a permanent feature in the subset. Then, having this feature permanently present, the algorithm randomly adds a second feature to the subset and a new criterion value is found. This is repeated until all the features have been tried. If the lowest criterion value of the feature subset (with two features) is smaller than the originally chosen subset (with one feature), the algorithm repeats the same steps by testing the addition of a third feature to the subset. Otherwise, the algorithm stops and the previous feature subset is deemed the optimal one. For example, if one needs to choose the optimal features from three initial ones, the algorithm successively calculates the criterion value obtained with each. If feature 1 leads to 0.056 as criterion value, feature 2 0.065 and feature 3 0.078, the algorithm permanently chooses feature 1. Then it tests the addition of feature 2 (classifier built with features 1 and 2) and feature 3 (classifier built with 1 and 3). If at least one the two criterion values is lower than 0.056 (initially obtained with feature 1), it tests the addition of the third feature to the feature subset. Otherwise, feature 1 is selected as optimal feature.

Part 2: Genetic Algorithm

Under the umbrella of feature selection, GA was also applied with its label "the fittest survives" [16]. In fact, GA mimics the genetic evolution by setting an initial population of binary chromosomes. Each gene is hence a binary digit in the chromosome. Afterwards, at each new generation, the chromosomes undergo three different phenomena:

- Selection: the fittest chromosomes of the initial population are preserved for the next generation
- Cross-over: new chromosomes are created in the new generation by mixing gene subsets of one chromosome with those of another
- Mutation: A certain gene from a given chromosome is randomly inverted (0 to 1 or vice-versa). This allows the algorithm to evaluate new options instead of getting stuck on local minima.

These three aforementioned phenomena are repeated during each transition from generation to another in order to progressively decrease the fitness value, until the predefined number of generations is reached. In this work, the fitness value was calculated using two different fitness functions and was taken as the classification loss of a SVM or a k-NN classifier. The population size is initially 5 chromosomes in which each gene is a feature that might be included (1) or rejected (0). The probability that it mutates is 10% and that of a cross-over is 80%. The selection probability is hence 10%.

Classification

In this study, the dataset was divided into 2 different subsets:

- 80% of the dataset was used as a training set: A set of feature vectors with known output was employed to build the classifier.
- 20% was used as testing set: A classifier is tested by predicting the outputs of a test set and comparing the predicted results to the actual ones. This step is important to evaluate the performance of any classifier used.

Once both sets were ready, the following types of classifiers were applied for performance comparison on the different classifiers: SVM [17] (linear SVM and using polynomial and Radial Basis Function (RBF) kernels), LDA [18] and Quadratic Discriminant Analysis (QDA) [19], and k-NN.

Performance evaluation

The performance of each classifier was individually evaluated by using the confusion matrix technique [20] which allows to compute the following parameters:

- The sensitivity (SE) which reflects the capacity of the classifier to detect BBB + drugs in the entire dataset
- The positive predictive value (PP) which expresses its ability not to deem non-crossing drugs as BBB+
- The specificity (SP) which expresses the ability of the model to detect BBB- drugs in the dataset
- The overall accuracy (ACC) which expresses the total true predictions over the total number of prediction

The receiver operating characteristic (ROC) [20] curve was also applied in our study.

Results

In this study, the first step was to train the classifiers without applying any type of feature selection algorithm. The results are tabulated in Table 1. The highest accuracy value obtained with the test set when training the classifiers with all the initial features was 93.35% acquired with SVM (RBF kernel function).

Table 1
Overall accuracy computed prior to applying feature selection

	SVM (Linear)	SVM (RBF)	SVM (polynomial)	LDA	QDA	k-NN
Accuracy without feature selection (%)	93.28	93.35	93.03	92.72	92.78	93.10

Feature selection

After obtaining the initial results, the following feature selection algorithms were performed and led to the following results:

Part 1: sequential feature selection

As described in Fig. 2, the convergence of the SFS towards the final feature subset is based on a criterion value extracted from the classifier itself: the number of misclassified observations in our study. Therefore, a specific feature subset was obtained with each classifier as reported in Table 2.

Table 2
Selected features by SFS algorithm

Classifier used	Features chosen
SVM (linear)	PSA, logP, HD, pKa (strongest acidic), NRB
SVM (RBF)	HD, HA, pKa (strongest acidic)
SVM (polynomial)	HD, HA, NRB
LDA	All but the HA
QDA	MW, PSA, HD, pKa (strongest acidic), pKa (strongest basic)
k-NN	All but pKa (strongest basic) and NRB

Part 2: genetic algorithm

Two different genetic algorithms were used in this study differing by the type of fitness value. In the first case, the fitness function returns the classification loss of a SVM, while in the next case, it returns that of a k-NN. In the former case, the selected features are PSA and HD, whereas in the latter case, the selected features are the PSA and pKa (strongest acidic).

Prediction performance evaluation

After running the feature selection algorithms, the same classifiers reported in Table 1 were trained separately using the selected features. The overall accuracy obtained on the test set is reported in table 3.

Table 3

Summary of the overall accuracy obtained with each classifier after applying SFS and GA

	SVM (Linear)	SVM (RBF)	SVM (polynomial)	LDA	QDA	k-NN
Without feature selection	93.28%	93.35%	93.03%	92.72%	92.78%	93.10%
Backward SFS	94.67%	92.79%	88.4013%	93.73%	94.98%	94.36%
GA: k-NN based Fitness function	94.67%	94.04%	84.01%	94.36%	96.23%	92.79%
GA: SVM based Fitness function	93.73%	94.98%	96.23%	94.98%	95.62%	93.42%

In the case of SFS, the overall accuracy increased with the linear SVM as well as LDA, QDA and k-NN. The highest accuracy value reached was 94.98%, obtained with the QDA. On the other hand, GA resulted in an increase of the overall accuracy of all the classifiers in the case of a fitness function based on the classification loss of a SVM. The highest accuracy value reached was 96.23% with the SVM (polynomial kernel function). Nevertheless, with the k-NN based fitness function, the SVM (polynomial kernel) and the k-NN witnessed a decrease in overall accuracy. The highest accuracy value was also 96.23%, obtained with the QDA which is higher than that received with the SFS. Table 2 compares in details the performance of the two classifiers that led to the 96.23% overall accuracy. **Fig. 4** ROC curves of the SVM classifier (polynomial kernel) before (red) and after (blue) feature selection using GA algorithm

Figure 4 represents the ROC curves of the SVM classifier trained with the entire feature set and after the application of GA for feature selection. It is clear that the area under the curve is much higher after applying GA.

Discussion

The dataset used in this study was chosen since drug permeability stratification is CNS-based [10], hence independent from logBB thresholds. The dataset was used to classify drugs into BBB + and BBB- while comparing two types of feature selection algorithms, the backward SFS and GA.

Table 4

Comparison of the performance of the best two classifiers (True +: actual BBB + drug predicted as BBB+, True -: actual BBB- drug predicted as BBB-, False +: actual BBB- drug predicted as BBB+, False -: actual BBB + drug predicted as BBB-)

	QDA + GA (k-NN based fitness function)	SVM + GA (SVM based fitness function)
True +	256	255
True -	51	52
False +	11	4
False -	1	8
SE	95.88%	98.45%
PP	99.61%	96.95%
SP	98.07%	86.67%
NP	82.25%	92.85%
ACC	96.23%	96.23%

In this study, the SFS resulted in larger feature subsets. Since at each iteration, SFS performs 10-fold cross validation and returns a criterion value specifically based on the classifier being trained, it is hypothetically tailored to optimize the results of a given classifier. It is to be noted that the number of hydrogen bond donors was selected with all the classifiers. This result is in line with previously reported findings on the significant role of hydrogen bonding characteristics in predicting drug permeability across the BBB [15].

However, GA showed that relying exclusively on the PSA and the number of hydrogen bond donors would lead to better results compared to those obtained while including other features. This is reflected by the overall accuracy (table 3) as well as the ROC curve (Fig. 4).

Moreover, it is to be noted that GA (with SVM based fitness function) led to an improvement of the results with all the reported classifiers unlike SFS. The highest overall accuracy was also found with GA.

Comparing the best two classifiers, one can note that the QDA trained using the PSA and pKa (strongest acid) has a slightly lower sensitivity and a higher specificity than the SVM trained with the PSA and the HD. Nevertheless, both classifiers have a better balancing between predicting BBB + and BBB- drugs than the binomial partial least squares implemented in [15] on the same data with selected molecular descriptors. However, the division of the dataset between training and test sets differs between the two studies. Furthermore, PSA was selected by GA in the two models that led to the highest overall accuracy. This finding is in accordance with previously reported results underlining the importance of PSA in stratifying drugs into BBB + and BBB- [13]. In fact, this descriptor appeared in the best four RF models created in [5] using up to four descriptors. Moreover, PSA was used in a classification tree model [7] as

the criterion allowing the first split of the tree. In addition, D. Zhang et al. [11] listed PSA among the ten most significant descriptors. All these findings reflect the robustness of GA in selecting the most relevant descriptors.

Given that the QDA trained with the PSA and pKa (strongest acid) led to fewer false negatives than the SVM trained with the PSA and the HD, we can speculate that this model is more useful when the research aim is to detect BBB + drugs such as drugs targeting CNS diseases. In fact, this model has a lower risk of classifying BBB + drugs as BBB- which would prevent a pharmaceutical company from moving forward with a BBB + drug candidate that could have made it through drug discovery phases. On the other hand, the latter model has fewer false positives than the former which makes it more valuable for the detection of BBB- drugs whose target is located outside the brain. However, these findings could be confirmed by increasing the size of the test set.

Conclusion

This work implements two different *in silico* BBB models that are mainly useful during the early phases of drug discovery. In fact, *in silico* BBB models allow pharmaceutical companies to reduce the number of hits that will undergo lengthy and expensive *in vitro* testing. Therefore, increasing the accuracy of *in silico* models is key. For this purpose, this paper applies and compares two different types of feature selection algorithms on a CNS-based dataset. The results show that GA enables an improvement of prediction accuracy over SFS. The best classifiers obtained after performing GA gave an accuracy of 96.23% and have a relatively good balance between predicting BBB + drugs to BBB- ones. Highly accurate BBB *in silico* models are needed in order to improve the stratification of drug candidates into BBB + and BBB- drugs at early phases of drug discovery and consequently save time and money associated with this complex, laborious process.

Abbreviations

ANN: Artificial Neural Network, BBB:Blood-Brain Barrier, CNS:Central Nervous System, DT:Decision Tree, GA:Genetic Algorithm, HA:Number of Hydrogen Bond Acceptors, HD:number of Hydrogen bond Donors, k-NN:k-Nearest Neighbors, LDA:Linear Discriminant Analysis, logBB:Logarithm of the ratio of the concentration in Brain to that in Blood, MLP:Multi-Layer Perceptron, MW:Molecular Weight, NRB:Number of Rotatable Bonds, PR:Permeation Rate, RBF:Radial Basis Function, RF:Random Forest, SFS:Sequential Feature Selection, SVM:Support Vector Machine

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The dataset used in this study is published and made publically available by Zhao et al. [15] free of charge via [<https://pubs.acs.org/doi/abs/10.1021/ci600312d#>]

Competing interests

The authors declare that they have no competing interests

Funding

This study was supported by the university research board (URB) at the American University of Beirut (AUB) and the higher center of research at the Holy Spirit University of Kaslik (USEK)

Authors' contributions

RM and SR guided and supervised the work performed in this paper. RS implemented the algorithms and evaluated the results.

Acknowledgements

This study was supported by the university research board (URB) at the American University of Beirut (AUB) and the higher center of research at the Holy Spirit University of Kaslik (USEK)

References

1. Zlokovic BV. The Blood-Brain Barrier in Health and Chronic Neurodegenerative Disorders. *Neuron*. 2008;57(2):178–201.
2. Banerjee J, Shi Y, Azevedo HS. In vitro blood–brain barrier models for drug research: state-of-the-art and new perspectives on reconstituting these models on artificial basement membrane platforms. *Drug Discovery Today*. 2016;21(9):1367–86.
3. Vastag M, Keseru GM. Current in vitro and in silico models of blood-brain barrier penetration: a practical view. *Curr Opin Drug Discov Devel*. 2009;12(1):115.

4. Burns J, Weaver DF. A Mathematical Model for Prediction of Drug Molecule Diffusion Across the Blood-Brain Barrier. *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques*. 2004;31(4):520–7.
5. Markus Muehlbacher GM, Spitzer, Klaus R, Liedl. Johannes Kornhuber: Qualitative prediction of blood–brain barrier permeability on a large and refined dataset. 2011.
6. Kunwittaya S, Nantasenamat C, Treeratanapiboon L, Srisarin A, Isarankura-Na-Ayudhya C, Prachayasittikul V. Influence of logBB cut-off on the prediction of blood-brain barrier permeability. *Biomedical Applied Technology Journal*. 2013;1:16–34.
7. Castillo-Garit JA, Casanola-Martin GM, Le-Thi-Thu H, Pham-The H, Barigye SJ. A Simple Method to Predict Blood-Brain Barrier Permeability of Drug- Like Compounds Using Classification Trees. *Med Chem*. 2017;13(7):664–9.
8. Wang Z, Yang H, Wu Z, Wang T, Li W, Tang Y, Liu G: In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods. *ChemMedChem* 2018, 13(20):2189–2201.
9. Singh M, Divakaran R, Kumar KLS, Kristam R. A classification model for blood brain barrier penetration. *Journal of Molecular Graphics and Modelling* 2019,:107516.
10. Adenot M, Lahana R. Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates. *J Chem Inf Comput Sci*. 2004;44(1):239–48.
11. Zhang D, Xiao J, Zhou N, Zheng M, Luo X, Jiang H, Chen K. A Genetic Algorithm Based Support Vector Machine Model for Blood-Brain Barrier Penetration Prediction. *BioMed research international* 2015, 2015:292683–13.
12. Yuan Y, Zheng F, Zhan C. Improved Prediction of Blood–Brain Barrier Permeability Through Machine Learning with Combined Use of Molecular Property-Based Descriptors and Fingerprints. *AAPS J*. 2018;20(3):1–10.
13. Brito-Sánchez Y, Marrero-Ponce Y, Barigye SJ, Yaber-Goenaga I, Morell†Pérez C, Le-Thi-Thu H, Cherkasov A. Towards Better BBB Passage Prediction Using an Extensive and Curated Data Set. *Mol Inf*. 2015;34(5):308–30.
14. Miao R, Xia L, Chen H, Huang H, Liang Y. Improved Classification of Blood-Brain-Barrier Drugs Using Deep Learning. *Scientific reports*. 2019;9(1):8802–11.
15. Yuan H, Zhao MH, Abraham A, Ibrahim PV, Fish. Predicting Penetration Across the Blood-Brain Barrier from Simple Descriptors and Fragmentation Schemes. *J Chem Inf Model*. 2007;47(1):170–5.
16. Raul Robu and Stefan Holban. A genetic algorithm for classification. In *Proceedings of the 2011 international conference on Computers and computing (ICCC'11)*. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA. 2011, 52–56.
17. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. *PLoS Comput Biol*. 2008;4(10):e1000173.

18. Sayad S. Linear Discriminant Analysis. Available from: <http://www.saedsayad.com/lda.htm>[Accessed 7 April 2020].
19. Boehmke B. Linear & Quadratic Discriminant Analysis. Available from: http://uc-r.github.io/discriminant_analysis[Accessed 7 April 2020].
20. Park SH, Goo JM, Jo C. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. Korean Journal of Radiology. 2004;5(1):11–8.

Figures

Figure 1

General workflow of the study: Drug Data collection followed by extraction of features which are the relevant molecular descriptors. Then, feature selection is performed on the gathered dataset using two different types of algorithms; the Sequential Feature Selection (SFS) and the Genetic Algorithm (GA). The resulting features set are fed into different classifiers (SVM, LDA, QDA, k-NN). Finally, the performance of each classifier is evaluated and compared with the others.

Figure 2

Flowchart of the forward sequential feature selection algorithm: the algorithm iteratively selects features to be added to a feature subset until reaching the subset with the lowest criterion value. In this work, the criterion value used is the number of misclassified observations in the test set.

Figure 3

Histogram showing the number of times each descriptor was chosen by the SFS algorithm among the six classifiers implemented. The number of hydrogen bond donors was selected with all the six implemented classifiers.

Figure 4

ROC curves of the SVM classifier (polynomial kernel) before (red) and after (blue) feature selection using GA algorithm