Asian Journal of Pharmaceutical Research and Health Care, *Vol 13(4), 312-320, 2021*
ISSN (Online) : 2250-1460
DOI: 10.18311/ajprhc/2021/28497

Comparative Study for Prediction of Low and High Plasma Protein Binding Drugs by Various Machine Learning-Based Classification Algorithms

Sumit Govil[1], Sandesh Tripathi[2], Amit Kumar[1], Divya Shrivastava[1,] and Shailesh Kumar[3]*

[1]School of Life Sciences, Jaipur National University, Jaipur - 302025, Rajasthan, India
[2]Birla Institute of Applied Sciences, Bhimtal, Nainital - 263136, Uttarakhand, India
[3]National Centre for Cell Science, NCCS Complex, Pune University Campus, Pune - 411007, Maharashtra, India, shailesh_iiita@hotmail.com

Abstract

In the drug discovery path, most drug candidates failed at the early stages due to their pharmacokinetic behavior in the system. Early prediction of pharmacokinetic properties and screening methods can reduce the time and investment for lead discoveries. Plasma protein binding is one of these properties which has a vital role in drug discovery and development. The focus of the current study is to develop a computational model for the classification of Low Plasma Protein Binding (LPPB) and High Plasma Protein Binding (HPPB) drugs using machine learning methods for early screening of molecules through WEKA. Plasma protein binding drugs data was collated from the Drug Bank database where 617 drug candidates were found to interact with plasma proteins, out of which an equal proportion of high and low plasma protein binding drugs were extracted to build a training set of ~300 drugs. The machine learning algorithms were trained with a training set and evaluated by a test set. We also compared various machine learning-based classification algorithms i.e., the Naïve Bayes algorithm, Instance-Based Learner (IBK), multilayer perceptron, and random forest to determine the best model based on accuracy. It was observed that the random forest algorithm-based model outperforms with an accuracy of 99.67% and 0.9933 kappa value on training set and on test set as compared to other classification methods and can predict drug plasma binding capacity in the given data set using the WEKA tool.

Keywords: Drug Discovery, Machine Learning, Multilayer Perceptron, Pharmacokinetic Plasma Protein Binding, Random Forest

1. Introduction

It is well-reported fact that the activity of a drug is controlled by various Pharmacokinetic (PK) and Pharmacodynamic (PD) factors. PK denotes the response of the body towards the drug and PD represents the effect of the drug on the body[1,3]. PK parameters play a vital role in drug discovery and development. In the last decades, the binding capacity of drugs to blood components like albumin i.e., Human Serum Albumin (HSA) and α-acid glycoprotein (AAG) has been widely studied. Interaction of drugs to other components like lipoproteins (γ-globulin), and erythrocytes also contributes to their PK behavior. Plasma protein binding works as a rate-controlling factor for the PK of drugs[1]. Both PK and PD phenomena together contribute to the efficacy of the drugs. One of the important factors that assess the PK and PD profile of medicine is the interaction of drug with plasma or serum protein which is a saturable and reversible process[1,3].

In blood, the drug existence is broadly classified as bound and unbound form. Majorly drugs are classified

*Author for correspondence

into two types based on their interaction with plasma proteins. Drugs that interact with high affinity to the plasma protein are HPPB, they are also reported to have low distribution resulting in lower efficacy of the drug. Drugs that interact with low affinity to plasma protein are LPPB results in the high efficacy of the drug because a high concentration of a drug is freely accessible to tissue. The degree to which the drug binds to blood plasma is correlated to its efficiency[3]. This plasma binding capacity also contributes to t1/2 of the drug by influencing its distribution, metabolism, and excretion i.e., high plasma binding drugs are less prone to metabolism and excretion hence have high t1/2. Thus, the prediction of the plasma binding capacity of the drug is an important factor for drug designing. Though plasma binding capacity depends on the structural properties of the molecules, we have used the machine learning method and compared them to predict plasma binding.

Several studies were reported to predict the plasma binding capacity of the drug by various approaches where Yuan *et al.* applied machine learning methods on molecular descriptors to predict the plasma binding capacity[26]. The plasma binding profile of the molecules has been tested by Sun *et al.* where they used 967 molecules to build the QSAR model[27]. A similar approach was discussed by Zhivkova where the author has used 220 basic drugs building QSAR model for early prediction of plasma binding and reported model accuracy up to 59%[18]. Toma *et al.* have provided an insight on the effect of ionization for plasma protein binding via the QSAR model[21,28].

Waikato Environment for Knowledge Analysis (WEKA) software which is a Java-based software developed at Waikato University provided under the GNU General public License was used for the development of models and testing for accuracy[12]. Data Mining approach was used to extract useful knowledge from a large amount of dataset[17]. These methods were used in multipurpose fields such as pattern identification, computational performance, information technology, machine learning, data classification, artificial intelligence systems, information retrieval, and neural networks[6]. Association rule mining, classification, and clustering are the three widely used techniques to analyze the information[12]. Among these three, we used the classification technique to build a computational model for predicting whether a drug is LPPB or HPPB. Classification has been considered as an instance of a supervised learning method where a pre-classified training set of correctly identified observations were provided to classification algorithms available for learning[7]. Accurate predictions of the target class for every case in the data is the basic goal of classification which is further tested on test and evaluation sets[10].

Using this machine learning technique of data mining, drugs can be classified into two Plasma Protein Binding drug data sets. Thus, in the early stages of drug development, these methods can help in predicting the drug class i.e., HPPB or LPPB. It may help in reducing the drug candidate failure in the later stages of the drug development process. The objective of the current study is to build a prediction model for the plasma binding capacity of the drug for early screening of molecules and compare model performance for different classification algorithms.

2. Materials and Methods

2.1 Collection and Pre-Processing of Data

Drug data is collected from Drug Bank database[22] (https://www.drugbank.ca/) and an excel file created in CSV format. We downloaded plasma binding drugs along with physicochemical properties and labeled them as HPPB and LPPB. The dataset consists of a matrix of 617 compounds with 15 features and after filtering the resultant data to remove noises where 615 compounds were taken further that are limited to drug distribution. The resultant with an initial matrix of 615×15 was taken further to build training, test, or validation sets. The training data included 300 drugs with an equal proportion of HPPB and LPPB ligands with all 15 properties. The remaining data was taken to build a test or validation set. The workflow of the current study has been represented in Figure 1.

2.2 Features/Parameters Selection

There are several factors involved in determining the binding capacity of molecules to plasma proteins. It becomes challenging for the classifier to classify the data with a huge no. of parameters therefore, only 15 features were taken further to train the algorithm. The features/parameters that were selected are water solubility, logP, logS, pKa (strongest acidic), pKb (strongest basic), physiological charger, hydrogen acceptor count, hydrogen

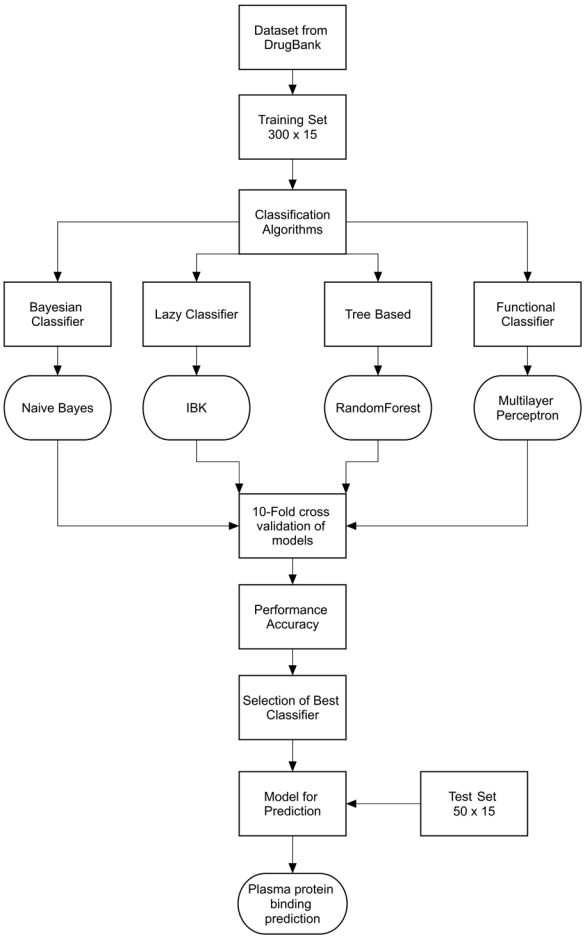

Figure 1. The workflow for classification of HPPB and LPPB drugs using machine learning-based classification algorithms.

donor count, polar surface area, refractivity, polarizability, number of rings, bioavailability, and weight.

2.3 Converting CSV File into ARFF file

WEKA prefers to load data in the input ARFF (Attribute-Relation File Format) file, which is an extension of the CSV (Comma Separated Values) file format where a separate header was represented which provides metadata about the data types in the columns[12]. A handy way is also provided by WEKA software to load CSV files which can further be saved as ARFF.

2.4 Train Model Using WEKA

WEKA is the most widely used data mining tool which supports a large amount of data mining algorithms for classification[8]. WEKA provides a feature where data can be loaded from various sources at the local system, including files, and can be extended to specific URLs and databases. ARFF, CSV, Lib SVM, and C4.5 are the supported file formats in WEKA[13]. WEKA comprises various tools for data filtering, pre-processing, classification models (inbuild and custom), regression analysis, clustering methodologies, association rules, and a few aspects of visualization[16]. It is also well-suited for developing new machine learning schemes. To train the model, select WEKA explorer opens the ARFF format file, and choose a classifier to train and save the model. The comparative analysis of four different algorithms was described based on various performance parameters.

2.4.1 Naive Bayes Classifier

It is one instance of the classification techniques based on the Bayes theorem of probability. It uses a probabilistic features-driven method that describes the condition-based probability of an event[11]. It assumes that the attribute values of given classwork an individual value, that simplifies the computations methodology involved thus it is called a Naïve Bayesian classifier[12].

2.4.2 IBK Classifier

IBK algorithm is implemented and derived from the K-Nearest Neighbor Algorithm (KNN) approach. In WEKA, it is called IBK (Instance-Based learning with parameter k)[12] and is available under the lazy class folder[11]. In machine learning it is sometimes referred to as memory-based learning. While classifying a test instance KKN (K-Nearest Neighbor) algorithm applies a mechanism to specify the number of nearest neighbors to use and the outcome is determined by vote majority toward a class.

2.4.3 Artificial Neural Networks

A multilayer perceptron with a free forward model on an artificial neural network defines a mapping on a set of input data on specified parameters onto a set of the appropriate output. The multilayered network comprises more than one layer of hidden perceptron that are not part of the input or output of the network. This hidden perceptron serves as a base unit of neural networks with interconnection to another perceptron in a complex network enable the learning of complex tasks by progressively extracting more meaningful features at each layer from the input patterns.

Machine learning methods such as Artificial Neural Networks (ANNs) imitate the processing behavior of neurons in a biological system where they perform function approximation and pattern recognition from a set of prototypes, in such a way that it can generate its mapping over new data[9,15]. This network determines a significant improvement over traditional analytical methods hence providing an opportunity to build forecasting with a better and precise insight by an understanding of relations among different variables, particularly over non-linear relationships[4]. It works through initial learning via a known set of data from a given problem with a known possible label or outcome (training set) that results in weighted networks, which are inspired and mimicking the analytical learning processes adopted by the human brain, these also can restructure the proceeding inaccurate rules which are being resulted via a complex set of data[14].

2.4.4 Random Forest

It works on a supervised learning approach for classification and is one of the best tree-based algorithms for classification. It works on the methodology to break the dataset into many subsets and generate the number of decision trees for classification. The result from each tree is averaged out to generate the final result. The greater number of subtrees helps in the better accuracy of the algorithms[24]. The random forest has added advantage for its processing speed where it takes less time for training in comparison to others. It can classify the dataset and predict the values by regression analysis[25]. This has been utilized for various applications in various filed, and extensively used for the classification of chemical compounds and prediction[23].

2.5 Model Evaluation

Two test modes are used to evaluate the selected tool:

a. The K-fold Cross-Validation Mode

The database is divided into K disjoint blocks of objects randomly, then the K-1 blocks are used to train the data mining algorithm and the performance of the algorithm is tested by the remaining blocks[13]. This process is repeated K times and the recorded measures were averaged in the end. The value of K depends on the size of the original dataset and commonly it is used as K=10.

b. Percentage Split (Holdout Method)

In data mining, it is common to split the dataset randomly into two disjoint datasets. In the first set viz training set, every data mining system derives knowledge from the pre-defined training set, and the resultant extracted knowledge is furthermore tested against the second set which is referred to as the test set. Usually, after the split, 66% of the objects of the original database are training sets and the remaining objects are represented in the test set. These results are collected and an overall comparison is conducted from the available classification and test modes, after the completion of the test on the selected dataset.

2.6 Performance of the Classifiers

The classifier's performance is evaluated under the following factors,

i) The accuracy parameter is the percentage of test set data that are correctly classified by the classifier.
ii) Kappa statistics of model build defines the chance agreement known as inter-rater reliability, lower the value the agreement by chance, where higher value represents the perfect agreement.
iii) Other parameters such as TP rate, FP rate, Accuracy, Sensitivity, Specificity, Recall, Precision, and F-measure were also compared to evaluate the models[13,19].

After choosing the best mode based on the above-mentioned parameter comparison. We further investigated the variable importance to develop the model which ranks the features for their influence on the model.

3. Results

Four strategies for data classification were investigated and compared: Naïve Bayes, IBK, ANN (multilayer perception), and random forest algorithm for classification of LPPB and HPPB. The classification techniques are used to find the most suitable algorithm for predicting the efficacy of the drug. We used these algorithms and compared them to develop a classifier model which can predict LPPB and HPPB. We have used a set of 615 drug candidates. Each drug was investigated with 15 descriptors. Test data and training data were prepared for

classification. The training set involved 300 data and the rest of the data was used to build test data.

The Naïve Bayes algorithm is a simple, clear, and fast classifier algorithm based on the Bayesian approach[16]. It assumes mutually independent attributes, therefore, called naïve. Practically, this is seldom true but is achievable by pre-processing the data to remove the dependent categories[16]. The correctly classified instances are 75% and incorrect prediction is for 25%, with observed kappa value as 0.5045 as shown in Figure 2 and represented in Table 1.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         225               75      %
Incorrectly Classified Instances        75               25      %
Kappa statistic                          0.5045
Mean absolute error                      0.242
Root mean squared error                  0.4788
Relative absolute error                 48.4194 %
Root relative squared error             95.7787 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.986     0.477     0.665      0.986     0.795      0.571     0.964     0.965     High PPB
                 0.523     0.014     0.976      0.523     0.681      0.571     0.963     0.958     Low PPB
Weighted Avg.    0.750     0.241     0.823      0.750     0.737      0.571     0.964     0.961

=== Confusion Matrix ===

   a   b   <-- classified as
 145   2 |   a = High PPB
  73  80 |   b = Low PPB
```

Figure 2. The capability of the Naïve Bayes algorithm for classification on 10-fold cross-validation.

The second algorithm is IBK and it is provided under the lazy class folder of WEKA. K- Nearest Neighbor (KNN) algorithm among the popular examples of an IBK Classifier. KNN algorithm works by specifying the number of nearest neighbors to use while classifying a test instance and the voting majority determines the outcome. WEKA applies cross-validation for the selection of the best value for KNN. The correctly classified instances are 87.67% and incorrect prediction accounts for 13.33 %, with an observed kappa value as 0.7334 as shown in Figure 3 and represented in Table 1.

The third algorithm was based on artificial neural networks. It is a modified version of the standard linear perceptron which applies three or more perceptron layers with nonlinear activation functions. It is more powerful than the perceptron as it can distinguish data that is not linearly separable or separable by an atypical hyperplane[2]. The high degree of connectivity of layers is determined by the network. A change in the population of synaptic connections based on their weight results in changes in the connectivity of the network[5], while the time taken to build the model is very less and correctly classified instances is 97.67% as shown in Figure 4. The kappa statistics observed to be 0.9533, it is performing better than NB and IBK algorithms as represented in Table 1.

The fourth algorithm used for classification was the random forest where 100 iterations were taken for evaluation of results. Random forest performance measures are represented in Figure 5. It had been

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         260               86.6667 %
Incorrectly Classified Instances        40               13.3333 %
Kappa statistic                          0.7334
Mean absolute error                      0.136
Root mean squared error                  0.3638
Relative absolute error                 27.2139 %
Root relative squared error             72.7711 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.884     0.150     0.850      0.884     0.867      0.734     0.867     0.808     High PPB
                 0.850     0.116     0.884      0.850     0.867      0.734     0.867     0.828     Low PPB
Weighted Avg.    0.867     0.133     0.867      0.867     0.867      0.734     0.867     0.818

=== Confusion Matrix ===

   a   b   <-- classified as
 130  17 |   a = High PPB
  23 130 |   b = Low PPB
```

Figure 3. The capability of the IBK algorithm for classification on 10-fold cross-validation.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         293               97.6667 %
Incorrectly Classified Instances         7                2.3333 %
Kappa statistic                          0.9533
Mean absolute error                      0.025
Root mean squared error                  0.1285
Relative absolute error                  5.0058 %
Root relative squared error             25.7102 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966     0.013     0.986      0.966     0.976      0.953     0.997     0.996     High PPB
                 0.987     0.034     0.968      0.987     0.977      0.953     0.997     0.997     Low PPB
Weighted Avg.    0.977     0.024     0.977      0.977     0.977      0.953     0.997     0.997

=== Confusion Matrix ===

   a   b   <-- classified as
 142   5 |   a = High PPB
   2 151 |   b = Low PPB
```

Figure 4. The capability of multilayer perceptron algorithm for classification on 10-fold cross-validation.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         299               99.6667 %
Incorrectly Classified Instances         1                0.3333 %
Kappa statistic                          0.9933
Mean absolute error                      0.0508
Root mean squared error                  0.0972
Relative absolute error                 10.1697 %
Root relative squared error             19.4505 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000     0.007     0.993      1.000     0.997      0.993     1.000     1.000     High PPB
                 0.993     0.000     1.000      0.993     0.997      0.993     1.000     1.000     Low PPB
Weighted Avg.    0.997     0.003     0.997      0.997     0.997      0.993     1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 147   0 |   a = High PPB
   1 152 |   b = Low PPB
```

Figure 5. The capability of random forest algorithm for classification on 10-fold cross-validation.

observed that random forest outperforms for plasma protein binding as compared to all tested algorithms as represented in Table 1. This works with an accuracy of 99.67% which is significantly higher than NB, IBK, and multilayer perceptron.

Comparison between various classification algorithms on data of plasma protein binding drugs is shown in Table 1. Here the on-all parameter comparison was carried to evaluate the best model for HPPB and LPPB protein prediction.

We also ranked the parameters on their importance in the model to understand the influence of each of the parameters for the prediction of protein binding this may help in determining the feature to be evaluated for degerming the probable drug candidate. The variable importance plot is represented in Figure 6, where Mean Decrease Accuracy and Mean Decrease Gini were reported for each feature. LogP, logS, water-solubility, polar surface area, and refractivity are the top five parameters that contribute to model generation.

4. Discussion

The data from the drug bank database along with molecular properties for 617 were downloaded and filtered for data corrections where two records are found with missing values which were removed from the analysis. Further, the resulting file of 615 records was taken with 15 properties like water solubility, logP, logS, pKa (strongest acidic), pKb (strongest

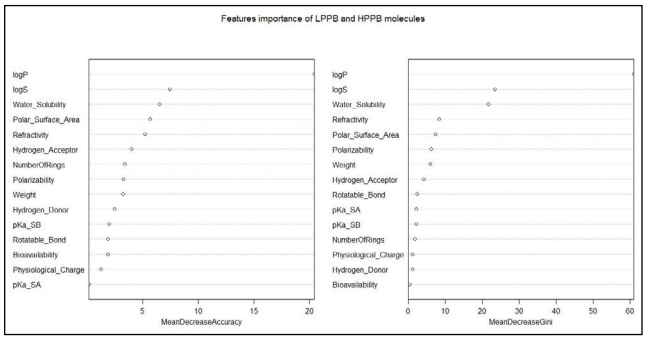

Figure 6. Feature importance of model building for LPPB and HPPB prediction, here features are ranked in order of their importance most influential feature is on top of the table i.e., logP and least one is at the bottom i.e., pKa strong acidic.

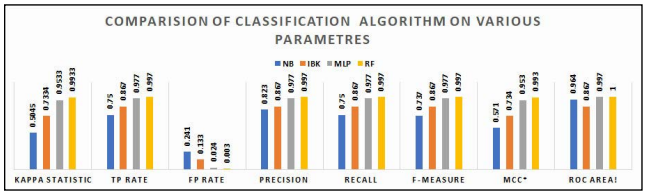

Figure 7. Comparison various performance metric for four different classification algorithms. On comparison RF model was found best for classification and prediction as compared to NB, IBK, MLP.

Table 1. Comparison between various classification algorithms on data of Plasma Protein Binding drugs on the training set with 300 drug candidates with 10-fold cross-validation and tested the resultant models on test set for 50 randomly selected drug candidates

	Classified Instances		Kappa statistic	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC*	ROC Area!
	Correct	Incorrect								
Training set evaluation metrics										
NB	225	75	0.5045	0.75	0.241	0.823	0.75	0.737	0.571	0.964
IBK	260	40	0.7334	0.867	0.133	0.867	0.867	0.867	0.734	0.867
MLP	293	7	0.9533	0.977	0.024	0.977	0.977	0.977	0.953	0.997
RF	299	1	0.9933	0.997	0.003	0.997	0.997	0.997	0.993	1
Test set evaluation metrics										
NB	42	8	0.6753	0.840	0.173	0.878	0.840	0.835	0.714	0.994
IBK	49	1	0.9599	0.980	0.022	0.981	0.980	0.980	0.961	0.979
MLP	50	0	1	1	1	1	1	1	1	1
RF	50	0	1	1	1	1	1	1	1	1

**Matthews correlation coefficient*

! Receiver Operator Characteristic

basic), physiological charger, hydrogen acceptor count, hydrogen donor count, polar surface area, refractivity, polarizability, number of rings, bioavailability, number of rotatable bonds and molecular weight. This file was transformed to ARFF file format which helps in loading data to WEKA software.

The resulting data was used to generate the training set and test set where the training set has ~150 molecules for each LPPB and HPPB category. This resulting file was taken with an equal number of LPPB and HPPB to minimize the model overfitting and bias. A training set of 300 records with properties was used to train the classification algorithm (Naïve Bayes, IBK, multilayer perceptron, and random forest) to generate models with 10-fold cross-validation methods. The resultant models were further exposed and tested for their performance along with a test set of 50 drug candidates. The results obtained based on various performance indicators or evaluation metrics for training and test dataset are enlisted in Table 1. and individual model details are provided in Figure 2-5.

It has been observed that the tree-based random forest model works exceptionally well for a given set of molecules as compared to other models. The precision and recall value of random forest has been found significantly higher than Naïve Bayes, IBK, and multilayer perceptron models over the training set. The number of correct classifications is also higher for random forests (99.70%). The F-measure (0.997) for the random forest is also reported to be better than other models. As represented in Figure 7 the comparative graph represents that the random forest model was found best among others classification algorithms. These models were afterward evaluated on a training set of 50 drugs taken randomly from the remaining drug data and their performance evaluation results are enlisted in Table 1. On evaluation, random forest and multilayered perceptron were found to work exceptionally well with all correctly predicted plasma binding capacity. As random forest performed better in both the sets i.e., training and test set. Hence, the random forest model can be further utilized for the preliminary analysis of the molecules at early stages. These models can also help in high throughput screening of molecular databases for determining the plasma binding capacity of the molecules as they can work as potential drug targets. The important feature shown in Figure 6 has also represented that logP, logS, water-solubility, polar surface area, and refractivity are the top five contributors to the model. These predictive models with feature properties provide a better opportunity to accelerate the process of drug discovery and development.

5. Conclusion

The current study for predicting protein binding class, and analyzing the comparative analysis for performance of models result reveals that the random forest model outperforms in terms of various evaluation metrics i.e., identification of correct instances in the random forest also called accuracy which is found to be 99.67% as compared to other classifier filters. This has been also evaluated that both multilayer perceptron and random forest works with similar performance on evaluation set but random forest model has higher F1 measures and kappa values in both training (in Figure 7) and test set hence, defining the capability of the random forest-based model in the detection of large drug data set. Therefore, the random forest-based model should be used for the classification of unknown drug candidate molecules (whether the drug molecule is a High Plasma Protein Binding drug or Low Plasma Protein Binding drug). The accuracy of the prediction was cross-validated with a 10-fold cross-validation mechanism was found to be 99.67%, its performance was also found best as compared to other methods hence, this random forest model can be further used for virtual screening of large database and molecules based on physicochemical properties during the drug designing process. These computational prediction models can help in accelerating the drug discovery process with better lead identification with a lower rejection rate of lead molecules at later stages of drug development.

6. Acknowledgement

Authors gratefully acknowledge University of Waikato for WEKA tool availability as an open-source. This work was supported by Jaipur National University, Jaipur, and Birla Institute of Applied Sciences through infrastructure and guidance.

7. References

1. Bohnert T, Gan LS. Plasma protein binding: from discovery to development. J Pharm Sci. 1 Sep 2013; 102(9): 2953–94. https://doi.org/10.1002/jps.23614

2. Chakravarthy SV, Ghosh J. Scale-based clustering using the radial basis function network. IEEE transactions on neural networks. Sep 1996; 7(5): 1250–61. https://doi.org/10.1109/72.536318
3. Chauhan AS, Raj U, Varadwaj PK. Prediction of Plasma Protein Binding affinity by support vector machine and artificial neural network. World J Pharma Res. 2014; 3: 432–441.
4. Grossi E. Non-Linearity in medicine: a problem or an opportunity. BMJ. 2001; 323: 750. https://doi.org/10.1136/bmj.323.7315.750
5. Howell AJ, Buxton H. 1 network methods for face detection and attentional frames. Neural Process Lett. Jun 2002; 15(3): 197–211. https://doi.org/10.1023/A:1015743231018
6. Han J, Jian P, and Micheline K. Data mining: concepts and techniques. Elsevier, 2011.
7. Kalmegh SK. Analysis of WEKA data mining algorithm REPTree, Simple Cart, and Random Tree for classification of Indian News. Int J Innov Sci Eng Tech. Feb 2015; 2(2): 438–46
8. Karthikeyan T, Thangaraju P. Analysis of classification algorithms applied to hepatitis patients. Int J Comput Appl. 1 Jan 2013; 62(15): 25–30. https://doi.org/10.5120/10157-5032
9. McEvoy F J, Amigo J M. Using machine learning to classify image features from canine pelvic radiographs: evaluation of partial least squares discriminant analysis and artificial neural network models. Vet Radiol Ultrasound. Mar 2013; 54(2): 122–126. https://doi.org/10.1111/vru.12003
10. Patil PH, Thube S, Ratriaparkhi B and Rajeswari K. Analysis of Different Data Mining Tools using Classification, Clustering and association rule mining. Int J Comp Appl. 1 Jan 2014; 93(8): 35–39. https://doi.org/10.5120/16238-5766
11. Rana R, Pruthi J. Heart Disease Prediction using Naïve Bayes classification in data mining. Int J Sci Res and Dev. 2014; 2(05): 2321–0613.
12. Revathi KK, Kavitha KK. Comparison of classification techniques on heart disease Dataset. Int J Adv Res Comp Sci. 2017 Nov 1; 8(9): 276–280. https://doi.org/10.26483/ijarcs.v8i9.4870
13. Kumar S, Govil S, Kumar V, Kachhawah S and Kothari SL. Classification of 5' and 3' untranslated regions in the human transcriptome by machine learning methods. Res J Biotechnol. 1 Dec 2018; 13(12): 47–53.
14. Sharma TC and Jain M. WEKA Approach for comparative study of classification Algorithm. Int J Adv Res Comp Comm Eng, Apr 2013; 2(4): 1925–31.
15. Street ME, Grossi E, Volta C, Faleschini E, Bernasconi S. Placental determinants of fetal growth: identification of key factors in the insulin-like growth factor and cytokine systems using artificial neural networks. BMC Pediatr Dec 2008; 8(25): 1–11. https://doi.org/10.1186/1471-2431-8-24
16. Toma C, Gadaleta D, Roncaglioni A, Toropov A, Toropova A, Marzo M, Benfenati E. QSAR development for plasma protein binding: influence of the ionization state. Pharm Res. Feb 2019; 36(2): 1–9. https://doi.org/10.1007/s11095-018-2561-8
17. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Sigmod Rec. 1 Mar 2002; 31(1): 76–7. https://doi.org/10.1145/507338.507355
18. Zhivkova Z, Doytchinova I. Quantitative structure—plasma protein binding relationships of acidic drugs. J Pharm Sci. 1 Dec 2012;101(12): 4627–41. https://doi.org/10.1002/jps.23303
19. Tiwari M, Govil S, Kumar S. A Review on Predictive Models and Classification of Inhibitors using Bioinformatics Approach. Int J Pharm Technol Biotechnol. 2015; 2(1): 26–32.
20. Zhivkova ZD. Quantitative structure–pharmacokinetics relationships for plasma protein binding of basic drugs. J Pharm & Pharm Sci. 2017; 20: 349–59. https://doi.org/10.18433/J33633
21. Zhu XW, Sedykh A, Zhu H, Liu SS, Tropsha A. The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. Pharm Res. Jul 2013; 30(7): 1790–8. https://doi.org/10.1007/s11095-013-1023-6
22. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 1 Jan 2014; 42(D1): D1091–7. https://doi.org/10.1093/nar/gkt1068
23. Cano G, Garcia-Rodriguez J, Garcia-Garcia A, Perez-Sanchez H, Benediktsson JA, Thapa A, Barr A. Automatic selection of molecular descriptors using random forest: Application to drug discovery. Expert Syst Appl. 15 Apr 2017; 72: 151–9. https://doi.org/10.1016/j.eswa.2016.12.008
24. Breiman L. Random forests. Mach Learn. Oct 2001; 45(1): 5–32. https://doi.org/10.1023/A:1010933404324
25. Kaushal, Sharma K., Kumar Shailesh, Singh Brijendra, Bundela Saurabh, Patro Nisha, Patro K. Ishan, and Bisen S. Prakash. "Targeting fatty acid synthase protein by molecular docking studies of naturally occurring ganoderic acid analogues acting as anti-obesity molecule." Res J Biotechnol. July 2019; 14(7): 52-61.
26. Yuan Y, Chang S, Zhang Z, Li Z, Li S, Xie P, Yau WP, Lin H, Cai W, Zhang Y, Xiang X. A novel strategy for prediction of human plasma protein binding using machine learning techniques. Chemometrics and Intelligent

Laboratory Systems. 15 Apr 2020; 199: 103962. https://doi.org/10.1016/j.chemolab.2020.103962

27. Sun L, Yang H, Li J, Wang T, Li W, Liu G, Tang Y. In silico prediction of compounds binding to human plasma proteins by QSAR models. ChemMedChem. 2018; 13(6): 572–581. https://doi.org/10.1002/cmdc.201700582
28. Zhivkova ZD. Quantitative structure–pharmacokinetics relationships for plasma protein binding of basic drugs. J. Pharm. Pharm. Sci. 2017; 20: 349–359. https://doi.org/10.18433/J33633