Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Prediction of Blood-Brain Barrier Permeability of Compounds by Machine Learning Algorithms

Tan Wei Feng[1], Raihana Edros[1,*], Ngahzaifa Ab Ghani[2], Siti Umairah Mokhtar[3], RuiHai Dong[4]

[1] Faculty of Chemical and Process Engineering Technology (FTKKP), Universiti Malaysia Pahang, 26300 Gambang, Pahang Darul Makmur, Malaysia
[2] Faculty of Computing, Universiti Malaysia Pahang, 26600 Pekan, Pahang Darul Makmur, Malaysia
[3] Faculty of Industrial Science and Technology, Universiti Malaysia Pahang, 26300 Gambang, Pahang Darul Makmur, Malaysia
[4] The Insight Centre for Data Analytics, School of Computer Science, University College Dublin, Dublin, Ireland

ARTICLE INFO

Article history:
Received 25 June 2023
Received in revised form 8 October 2023
Accepted 17 October 2023
Available online 3 November 2023

Keywords:
Machine Learning; Blood Brain Barrier; classification

ABSTRACT

In the drug development for the Central Nervous System (CNS), the discovery of the compounds that can pass through the brain across the Blood-Brain Barrier (BBB) is the most challenging assessment. Almost 98% of small molecules are unable to permeate BBB, reducing the pharmacokinetics of the drugs in the CNS by affecting its absorption, distribution, metabolism, and excretion (ADME) mechanisms. Since the CNS is often inaccessible to many complex procedures and performing in-vitro permeability studies for thousands of compounds can be laborious, attempts were made to predict the permeation of compounds through BBB by implementing the Machine Learning (ML) approach. In this work, using the KNIME Analytics platform, 4 predictive models were developed with 4 ML algorithms followed by a ten-fold cross-validation approach to predict the external validation set. Among 4 ML algorithms, Extreme Gradient Boosting (XGBoost) overperformed in BBB permeability prediction and was chosen as the prediction model for deployment. Data pre-processing and feature selection enhanced the prediction of the model. Overall, the model achieved 86.7% and 88.5% of accuracy and 0.843 and 0.927 AUC, respectively in the training set and external validation set, proving that the model with high stability in prediction.

1. Introduction

The transport of the water, plasma components, amino acid, and glucose across BBB is facilitated via several transport mechanisms including passive diffusion and active transport. Nutrient transporter such as glucose transporter (Glut-1) and monocarboxylic acid transporter (MCT) are involved in transporting specific solutes via influx transport systems which are Carrier-mediated (SLC) transport, Receptor-mediated Transport (RMT), and endocytosis by Adsorptive-mediated Transcytosis Transport [1]. Meanwhile, efflux transporters such as P-glycoprotein (P-gp) and multidrug resistance (MDR1) are involved in the efflux system which is responsible for the efflux of endogenous molecules including toxic xenobiotics and neural waste products such as amyloid-beta.

[*] *Corresponding author.*
E-mail address: rzahirah@umpsa.edu.my

https://doi.org/10.37934/araset.33.2.269276

Nevertheless, drugs that are substrates to the efflux transporter, especially P-gp, may be effluxed out of CNS and reduce the pharmacokinetics of the drugs [2]. As result, the characteristic of the drugs such as physicochemical properties including molecular size and lipophilicity and substrate to the efflux transporter should be taken into consideration in the BBB permeability prediction.

Generally, Lipinski's Rule of Thumb guides the prediction of the potential of the compounds to permeate BBB. A molecule that passes through BBB via transcellular diffusion should have molecular weight < 500 Da, C log P within the range of 5, the sum of nitrogen and oxygen (N+O) atoms ≤ 5, and C log P - (N+O) should be positive, indicating log BB is likely positive and able to permeate BBB [3]. Meanwhile, other studies proposed that the log D value should be in the range of 1-3 and topological polar surface area (PSA) is limited to 90 Å2 or 60 Å2 [4,5]. The permeability of many discovered compounds across the BBB can be examined according to their physico-chemical properties through ML approach. However, the prediction of the compound's penetration should not be limited to molecular parameters but also need to include the structural properties of compounds to have a complete view of BBB permeability. Concurrently, this implementation is also crucial to ensure the effectiveness during the prediction of BBB permeability for a large dataset.

In the beginning of the CNS-targeted drug development, the prediction of BBB permeability is mostly focused on single learning algorithms, where each algorithm has its limitations in learning the input datasets and is prone to inconsistencies, especially in actual applications. Imbalanced and noisy data are the causes of inconsistencies in datasets, resulting in less reliable and inadequate models to produce a forecast with the highest accuracy. Additionally, massive data that cause diverse distribution of datasets is one of the critical issues as the greatly varying data may complicate the classification, reducing the performance and accuracy of the classification.

In this study, the fingerprints and molecular descriptors were calculated from 7236 compounds SMILES strings from previous resources for classification by using four predictive models known as Random Forest (RF), Decision Tree (DT), Gradient Boosted Trees, and Extreme Gradient Boosting (XGBoost). The dataset was divided into modelling set of 6150 compounds and a validation set of 1086 compounds. To build an ML model with robust performance, data pre-processing such as feature selection is required to filter out the redundant and irrelevant data before being applied for the training ML model. Tanimoto correlation coefficient and Variance threshold are used with defined threshold values to remove similar fingerprints with high Tanimoto coefficient and features with low variance. As result, around 25% of features were removed from total descriptor features to ensure the robust performance of the ML model. With that, the dataset can be used in ML model construction to predict BBB permeability.

2. Methodology

2.1 Collection and Preparation of Data

In this study, the data set was collected from the literature published in 2020 [6] which was integrated from other eight studies [7-14]. This data set consists of 7236 compounds (5492 BBB+ and 1744 BBB-) and each compound was prepared in the format of canonical Simplified Molecular-Input Line-Entry System (SMILES) from the database. The missing values in input data and redundant data were filtered and removed for classification of BBB permeability [15]. For the external validation set, the data set were divided into 90% of the training set (6513 compounds) and 10% of the test set (723 compounds) to execute 10-fold cross-validation in the KNIME Analytics Platform.

2.2 Generation of Fingerprints and Molecular Descriptors

To generate the fingerprint and molecular descriptors, CDK KNIME Extension was used to generate 2D coordinates for all compounds based on their SMILES. With that, the molecular descriptor can be generated by using the RDKit Descriptor Calculation node in KNIME RDKit Extension. A total of 124 1D and 2D descriptors including hydrogen bond and acceptor count, TPSA, Lipinski Rules of Five and molecular weight are generated which are important in BBB permeability prediction [16]. For fingerprint, the CDK Fingerprint node was used to generate the molecular fingerprint for each of the compounds based on their SMILES file [17, 18]. CDK Fingerprint Similarity node was used to generate the Tanimoto coefficient to filter out similar fingerprints in the feature selection step.

2.3 Feature Selection

Feature Selection was used for data pre-processing to remove the redundant data in machine learning. Removing the irrelevant data in a large number of features may help in improving the performance of algorithms in model construction [9, 15, 19, 20]. In this study, the Low Variance Filter node was used to remove the data below the variance threshold at 0.90. The threshold of the variance was set at 0.90 and the features with variance below the threshold were deleted, providing 94 features were used for model training.

2.4 Model Construction

Four machine learning algorithms in the KNIME Analytics Platform were used for constructing the best predictive model for BBB permeation of compounds. The algorithms include Random Forest (RF), Decision Tree (DT), Gradient Boosted Trees (GBM), and Extreme Gradient Boosting (XGBoost) [20–22]. In this context, the supervised binary classification of the datasets (BBB±), which is discrete data, was partitioned into 85% of the training set (5423 compounds) and 15% of the internal test set (957 compounds) for model training. Each of the algorithm learners was executed by using training set to predict the permeability of compounds through BBB based on the descriptors and binary classification through the respective predictor in the KNIME Analytics Platform. To train a robust model, hyperparameter tuning was required to optimize the performance of algorithms hence achieving better accuracy in prediction. Figure 1 shows the workflow of the model construction in KNIME Analytics Platform.

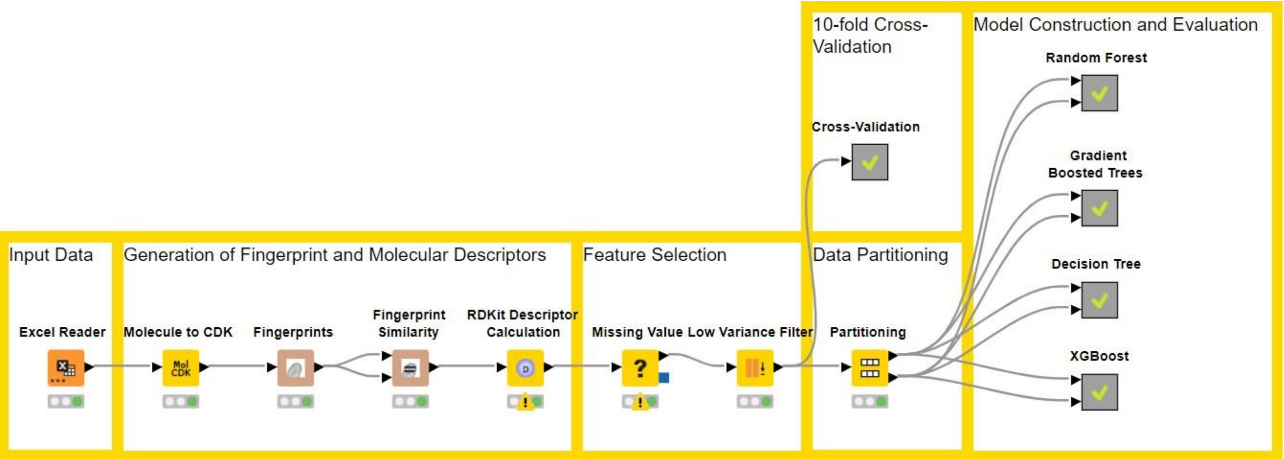


Fig. 1. Model construction workflow for prediction of BBB permeability

2.5 Model Evaluation

In the model evaluation, an external validation test set was used to validate the model to analyse the stability of the predictive model in performing the prediction by using 10-fold cross-validation. An internal test set was also used to figure out the influence of feature selection in constructing the model from machine learning algorithms [6, 23].

To analyze the model performance, the statistical parameter including Accuracy, Sensitivity, Specificity, and F1 score are calculated by using the following equations:

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TP+FP+TN+FN}} \tag{1}$$

$$\text{Sensitivity/Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP+TN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{F} - \text{measure} = \frac{2\times\text{P}\times\text{R}}{\text{P+R}} \tag{5}$$

where TP is true positive (BBB+), FP is false positive (BBB+), TN is true negative (BBB-) and FN is false negative (BBB-). The area under the receiver-operating characteristic (ROC) curve (AUC) was also used to evaluate the overall performance of the model [22, 24]. From the value of the statistical parameter, the evaluation of the model can be conducted to select a predictive model with robust performance in predict BBB permeability of compounds.

3. Results

3.1 Dataset and Feature Selection

The BBB dataset was integrated from previous studies and the compounds in the format of SMILES were divided into BBB+ and BBB- to train the machine learning algorithms. To assist in the prediction performance, BBB permeability can be expressed in the fingerprint and descriptors based on their physicochemical and topological properties. From feature selection, 25% of the features were remove includes physicochemical descriptors that describing number of heterocycles in aromatic, aliphatic and saturated hydrocarbon, and number of carbocycles in aromatic and saturated compounds. Topological properties such as SlogP_VSA9 and SMR_VSA2 that explain the binned accessible Van der Waals surface area that is contributed by lipophilicity and molecular refractivity were also removed because the values in these features converge to the average values or constant, hence do not provide useful information in contributing ML pattern in modelling with the variance below 0.90. In contrast, 94 remaining features with variance higher than 0.90 have higher degree of spread, indicating they are good predictors to provide useful information to ML patterns. By dropping constant features, the prediction for BBB permeability can be improved with stronger strength of association between features.

3.2 Evaluation of Prediction Performance of The Model

In this study, there are four algorithms were trained for prediction model construction and the model performance were evaluated by using the statistical metrics. As shown in Table 1, XGBoost model overperformed other algorithms with 0.927 of AUC, 88.5% of Accuracy, 0.910 of Sensitivity, 0.780 of Specificity and 0.928 of F-measure. The second highest from the perspective of overall performance is RF with 87.6 of ACC and 0.905 of AUC followed by DT and GBM. Boosting technique applied in XGBoost reduce bias of the data and reweighing the misclassified BBB classification result from previous iterations. The results showed that XGBoost model is the most reliable model to be selected for BBB permeability prediction in deployment. Furthermore, this model achieved accuracy of 86.7%, specificity of 0.722, sensitivity of 0.903, F-measure of 0.916 and AUC of 0.843 in 10-fold cross-validation, showing that the model constructed have acceptable performance in prediction of BBB permeability. The Figure 2 below visualized the ROC curve of each of the model as well as the cross validation.

Table 1
Comparison of the performance of models in BBB permeability prediction

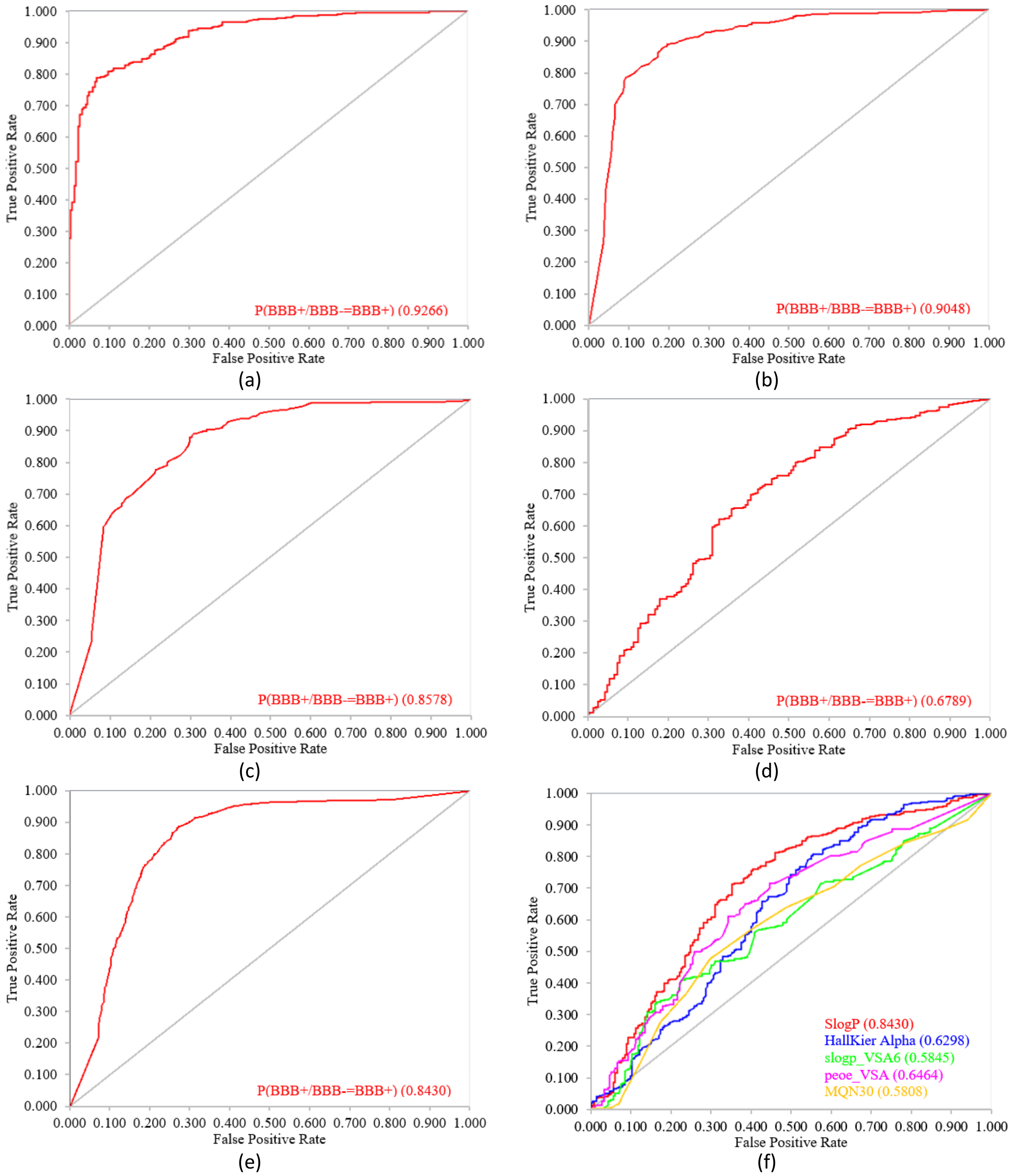
Model Evaluation	AUC	ACC	SEN	SPE	F-measure
Test Set					
Extreme Gradient Boosting	0.927	88.5	0.910	0.780	0.928
Random Forest	0.905	87.6	0.898	0.772	0.922
Decision Tree	0.858	84.8	0.899	0.663	0.903
Gradient Boosted Trees	0.678	88.5	0.902	0.804	0.928
External Validation					
10-fold Cross-validation	0.843	86.7	0.903	0.722	0.916

3.3 Molecular Descriptors

In the prediction of BBB permeability, the physicochemical and topological descriptors are significant to explain the influences of properties on the BBB penetration, which is numbered in AUC-ROC curve as visualized in Figure 2. Among 94 descriptors, SlogP is the most influential descriptor in prediction which show acceptable discrimination according to the AUC value, 0.716. There are 19 descriptors showed poor discrimination where their AUC values fall between 0.501 to 0.682 including HallKierAlpha, SlogP_VSA_6, PEOE_VSA6 and MQN30 whereas 74 descriptors cannot be used as reference as their AUC value are lower than 0.5, showing no discrimination on the classification.

SlogP is partition coefficient of octanol/water that measure lipophilicity and it is critical to be considered in prediction as the compounds must be sufficiently hydrophobic in water and lipophilic to solubilize in the hydrophobic core of the luminal and abluminal membrane of BBB [25]. SlogP_VSA provides information on the intermolecular interaction caused by different polarity of functional groups which affects the intermolecular radii and van der Waals surface area whereas Partial Equalization of Orbital Electronegativities (PEOE_VSA6) and Molecular Quantum Number 30 (MQN30) summarize the overall charge density within the molecule and atom valency that contribute to the electronegativity of partially distributed electrons within their orbital, determining the van der Waals radii of a compound [26, 27]. Hall-Kier α shows the influences of atom hybridization state in the molecule on the covalent radii and molecular size [28].

The most important factors in determining a molecule's ability to cross the blood-brain barrier (BBB) are its lipophilicity and molecular size. Molecules with unhybridized p-orbitals have a higher electron-sharing capacity, making them less likely to attract electrons and reducing their charge density. This leads to shorter, stronger intermolecular bonds, making them smaller in size and more lipophilic to pass through the BBB via the transcellular lipophilic pathway.



(a) (b) (c) (d) (e) (f)

Fig. 2. The ROC curve of model constructed for BBB permeability prediction by using algorithm (a)Extreme Gradient Boost (XGBoost), (b)Random Forest (RF), (c)Decision Tree (DT), (d)Gradient Boosted Trees, (e)Cross Validation and (f) the most influential descriptors in the prediction

4. Conclusions

Four ML algorithms were trained in constructing a classification model after feature selection for prediction of BBB permeation of compounds. Based on the result in this work, XGBoost model achieved highest accuracy of 88.5% and AUC of 0.927 in test set whereas the accuracy and AUC of external validation were recorded as 86.7 and 0.843, respectively. This indicates that XGBoost model is reliable to be implemented in the deployment of BBB permeability prediction. Nevertheless, descriptors should be considered in the prediction as it provides insight to the properties of molecules from the perspective of physicochemical and topology that are critical to BBB penetration.

Acknowledgement

The authors would like to thank Ministry of Education Malaysia in providing the fund for this project under the Fundamental Research Grant Scheme (FRGS; Grant Number: RDU210118; FRGS/1/2021/SKK06/UMP/02/1) and Universiti Malaysia Pahang for the research funding under the grant number RDU1703111. The authors would also like to express gratitude to the Faculty of Chemical and Process Engineering Technology (FTKKP) for the equipment and training provided to accomplish this research.

References

[1] Wong, Andrew D., Mao Ye, Amanda F. Levy, Jeffrey D. Rothstein, Dwight E. Bergles, and Peter C. Searson. "The blood-brain barrier: an engineering perspective." *Frontiers in neuroengineering* 6 (2013): 7. https://doi.org/10.3389/fneng.2013.00007

[2] Pulido, Robert S., Roeben N. Munji, Tamara C. Chan, Clare R. Quirk, Geoffrey A. Weiner, Benjamin D. Weger, Meghan J. Rossi et al. "Neuronal activity regulates blood-brain barrier efflux transport through endothelial circadian genes." *Neuron* 108, no. 5 (2020): 937-952. https://doi.org/10.1016/j.neuron.2020.09.002

[3] Clark, David E., A. Doherty, M. Bock, M. Desai, J. Overington, and J. Plattner. "Computational prediction of blood-brain barrier permeation." *Annual reports in medicinal chemistry* 40 (2005): 403. https://doi.org/10.1016/S0065-7743(05)40026-3

[4] Kelder, Jan, Peter DJ Grootenhuis, Denis M. Bayada, Leon PC Delbressine, and Jan-Peter Ploemen. "Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs." *Pharmaceutical research* 16 (1999): 1514-1519. https://doi.org/10.1023/A:1015040217741

[5] van de Waterbeemd, Han, Gian Camenisch, Gerd Folkers, Jacques R. Chretien, and Oleg A. Raevsky. "Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors." *Journal of drug targeting* 6, no. 2 (1998): 151-165. https://doi.org/10.3109/10611869808997889

[6] Shaker, Bilal, Myeong-Sang Yu, Jin Sook Song, Sunjoo Ahn, Jae Yong Ryu, Kwang-Seok Oh, and Dokyun Na. "LightBBB: computational prediction model of blood–brain-barrier penetration based on LightGBM." *Bioinformatics* 37, no. 8 (2021): 1135-1139. https://doi.org/10.1093/bioinformatics/btaa918

[7] Yuan, Yaxia, Fang Zheng, and Chang-Guo Zhan. "Improved prediction of blood–brain barrier permeability through machine learning with combined use of molecular property-based descriptors and fingerprints." *The AAPS journal* 20 (2018): 1-10. https://doi.org/10.1208/s12248-018-0215-8

[8] Martins, Ines Filipa, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao. "A Bayesian approach to in silico blood-brain barrier penetration modeling." *Journal of chemical information and modeling* 52, no. 6 (2012): 1686-1697. https://doi.org/10.1021/ci300124c

[9] Wang, Zhuang, Hongbin Yang, Zengrui Wu, Tianduanyi Wang, Weihua Li, Yun Tang, and Guixia Liu. "In silico prediction of blood–brain barrier permeability of compounds by machine learning and resampling methods." *ChemMedChem* 13, no. 20 (2018): 2189-2201. https://doi.org/10.1002/cmdc.201800533

[10] Adenot, Marc, and Roger Lahana. "Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates." *Journal of chemical information and computer sciences* 44, no. 1 (2004): 239-248. https://doi.org/10.1021/ci034205d

[11] Gao, Zhen, Yang Chen, Xiaoshu Cai, and Rong Xu. "Predict drug permeability to blood–brain-barrier from clinical phenotypes: drug side effects and drug indications." *Bioinformatics* 33, no. 6 (2017): 901-908. https://doi.org/10.1093/bioinformatics/btw713

[12] Plisson, Fabien, and Andrew M. Piggott. "Predicting blood–brain barrier permeability of marine-derived kinase

inhibitors using ensemble classifiers reveals potential hits for neurodegenerative disorders." *Marine drugs* 17, no. 2 (2019): 81. https://doi.org/10.3390/md17020081

[13] Singh, Manvi, Reshmi Divakaran, Leela Sarath Kumar Konda, and Rajendra Kristam. "A classification model for blood brain barrier penetration." *Journal of Molecular Graphics and Modelling* 96 (2020): 107516. https://doi.org/10.1016/j.jmgm.2019.107516

[14] EN, Wu Z. Ramsundar B. Feinberg, and Gomes J. Geniesse C. Pappu AS. "Leswing K Pande V." *MoleculeNet: a benchmark for molecular machine learning Chem Sci* 9, no. 2 (2018): 513. https://doi.org/10.1039/C7SC02664A

[15] Liu, Lili, Li Zhang, Huawei Feng, Shimeng Li, Miao Liu, Jian Zhao, and Hongsheng Liu. "Prediction of the blood–brain barrier (BBB) permeability of chemicals based on machine-learning and ensemble methods." *Chemical Research in Toxicology* 34, no. 6 (2021): 1456-1467. https://doi.org/10.1021/acs.chemrestox.0c00343

[16] Allouche, Abdul‐Rahman. "Gabedit—A graphical user interface for computational chemistry softwares." *Journal of computational chemistry* 32, no. 1 (2011): 174-182. https://doi.org/10.1002/jcc.21600

[17] Yu, Tzu-Hui, Bo-Han Su, Leo Chander Battalora, Sin Liu, and Yufeng Jane Tseng. "Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying CNS drugs with high prediction power." *Briefings in bioinformatics* 23, no. 1 (2022): bbab377. https://doi.org/10.1093/bib/bbab377

[18] Carracedo-Reboredo, Paula, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J. Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, and Carlos Fernandez-Lozano. "A review on machine learning approaches and trends in drug discovery." *Computational and structural biotechnology journal* 19 (2021): 4538-4558. https://doi.org/10.1016/j.csbj.2021.08.011

[19] Perez-Riverol, Yasset, Max Kuhn, Juan Antonio Vizcaíno, Marc-Phillip Hitz, and Enrique Audain. "Accurate and fast feature selection workflow for high-dimensional omics data." *PloS one* 12, no. 12 (2017): e0189875. https://doi.org/10.1371/journal.pone.0189875

[20] Shi, Zhiwen, Yanyi Chu, Yonghong Zhang, Yanjing Wang, and Dong-Qing Wei. "Prediction of blood-brain barrier permeability of compounds by fusing resampling strategies and extreme gradient boosting." *IEEE Access* 9 (2020): 9557-9566. https://doi.org/10.1109/ACCESS.2020.3047852

[21] Kumar, Vinod, Sumeet Patiyal, Anjali Dhall, Neelam Sharma, and Gajendra Pal Singh Raghava. "B3pred: A random-forest-based method for predicting and designing blood–brain barrier penetrating peptides." *Pharmaceutics* 13, no. 8 (2021): 1237. https://doi.org/10.3390/pharmaceutics13081237

[22] Roy, Dipankar, Vijaya Kumar Hinge, and Andriy Kovalenko. "To pass or not to pass: predicting the blood–brain barrier permeability with the 3D-RISM-KH molecular solvation theory." *ACS omega* 4, no. 16 (2019): 16774-16780. https://doi.org/10.1021/acsomega.9b01512

[23] Andrius, Vabalas, Gowen Emma, and Poliakoff Ellen. "Casson Alexander J." *Machine learning algorithm validation with a limited sample size. PLoS ONE* 14 (2019): e0224365. https://doi.org/10.1371/journal.pone.0224365

[24] Hossin, M. "Sulaiman,“." *A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS,” IJDKP) Int. J. Data Min. Knowl. Manag. Process* 5, no. 2 (2020).

[25] Naylor, Matthew R., Andrew M. Ly, Mason J. Handford, Daniel P. Ramos, Cameron R. Pye, Akihiro Furukawa, Victoria G. Klein et al. "Lipophilic permeability efficiency reconciles the opposing roles of lipophilicity in membrane permeability and aqueous solubility." *Journal of medicinal chemistry* 61, no. 24 (2018): 11169-11182. https://doi.org/10.1021/acs.jmedchem.8b01259

[26] Chester, Karishma, Sultan Zahiruddin, Adil Ahmad, Washim Khan, Sarvesh Paliwal, and Sayeed Ahmad. "Bioautography-based identification of antioxidant metabolites of Solanum nigrum L. and exploration its hepatoprotective potential against D-galactosamine-induced hepatic fibrosis in rats." *Pharmacognosy Magazine* 15, no. Suppl 1 (2019): S104-S110.

[27] Awale, Mahendra, Ruud Van Deursen, and Jean-Louis Reymond. "MQN-mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13." (2013): 509-518. https://doi.org/10.1021/ci300513m

[28] Summers, Andrew Z., Justin B. Gilmer, Christopher R. Iacovella, Peter T. Cummings, and Clare McCabe. "MoSDeF, a Python framework enabling large-scale computational screening of soft matter: Application to chemistry-property relationships in lubricating monolayer films." *Journal of Chemical Theory and Computation* 16, no. 3 (2020): 1779-1793. https://doi.org/10.1021/acs.jctc.9b01183