A Computational Prediction Model of Blood-Brain Barrier Penetration Based on Machine Learning Approaches

Deep Himmatbhai Ajabani
Application Developer, Lead, Source InfoTech Inc., Atlanta, Georgia, United States

***Abstract*—Within the field of medical sciences, addressing brain illnesses such as Alzheimer's disease, Parkinson's disease, and brain tumors poses significant difficulties. Despite thorough investigation, the search for truly successful neurotherapies continues to be challenging to achieve. The blood-brain barrier (BBB), which is currently a major area of research, restricts the passage of medicinal substances into the central nervous system (CNS). It is crucial in the field of neuroscience to create drugs that can effectively cross the blood-brain barrier (BBB) and treat cognitive disorders. The objective of this study is to improve the accuracy of machine learning models in predicting BBB permeability, which is a critical factor in medication development. In recent times, a range of machine learning models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Artificial Neural Networks (ANN), and Random Forests (RF) have been utilized for BBB. By employing descriptors of varying dimensions (1D, 2D, or 3D), these models demonstrate the potential to make precise predictions. However, the majority of these studies are biased to the nature of datasets. To accomplish our objective, we utilized three BBB datasets for training and testing our model. The Random Forest (RF) model has shown exceptional performance when used on larger datasets and extensive feature sets. The RF model attained an overall accuracy of 90.36% with 10-fold cross-validation. Additionally, it earned an AUC of 0.96, a sensitivity of 77.73%, and a specificity of 94.74%. The assessment of an external dataset resulted in an accuracy rate of 91.89%, an AUC value of 0.94, a sensitivity rate of 91.43%, and a specificity rate of 92.31%.**

Keywords*—Central Nervous System (CNS); Blood-Brain Barrier (BBB); Machine Learning (ML); Simplified Molecular Input Line Entry System (SMILES); Support Vector Machine (SVM); K-Nearest Neighbor (KNN); Logistic Regression (LR); Multi-Layer Perceptron (MLP); Light Gradient Boosting Machine (LightGBM); Random Forest (RF)

I. Introduction

In the last few decades, human brain diseases like tumors in the brain, dementia, Alzheimer, and other brain disorders are the most common fastest-growing issue nowadays that causes disability in humans and received the most attention from the research community in medical sciences. As there are no effective treatments have been made by neurotherapist to treat these kinds of serious diseases. Almost all macro and small molecule drugs are blocked by a barrier named the BBB [1]. The BBB is the most important key point in the treatment of brain diseases as it forcibly prevents the drugs from crossing this barrier and enters the CNS [2]. Many ML and Deep learning techniques have been made in the past and till now most researchers have been working on this problem of BBB permeability but still, the question arises on the performance of models and their precise results for drug formation in pharmaceutics [3], [4].

As the name indicates BBB, it is the barrier between blood and the brain. The barrier is made of endothelium cells [5] which can prevent large and even small molecules from entering the CNS [6]. It only allows some specific molecules like water molecules and some lipid-soluble to cross the barrier [7]. The BBB is divided into two classes labeled BBB+ and BBB-. The BBB+ shows the higher permeability and the BBB- shows the lower permeability respectively [8]. Developing a classification model requires a piece of detailed information and a complete understanding of issues or problems regarding BBB permeability. These issues are mainly caused by the selection of algorithms and the dataset on which these computations are performed. The problems faced in algorithms include their lower coverage, overfitting w.r.t dataset, and lower accuracy scores while predicting the molecular compounds i.e., (BBB-). The problems raised regarding datasets are duplication of compounds or improper class label distribution in the BBB dataset, which is a serious cause of inaccurate results [9].

In BBB there are molecular descriptors used as features in the dataset. The definition of molecular descriptor states that the transformation of chemical compounds by applying mathematical procedure converts these compounds into standardized numeric information that can be used for further experiments [10]. Molecular descriptors encompass various characteristics of molecules, such as their weight, amount of carbon atoms, and hydrogen bonds. The literature review primarily focused on the discussion of various classes of molecular descriptors, namely one dimension, two dimensions, and three dimensions. The representation of molecular descriptors is categorized into several kinds. Numerous classes have been discussed in the existing body of scholarly literature. The classes were categorized into three distinct groups, namely 1D, 2D, and 3D.

One-dimensional molecular descriptors, such as the number of certain atoms and molecular weights, are utilized to express the attributes of molecules [11]. The presentation of structural information is accomplished by the utilization of 2D molecular descriptors. It is computed from the 2D molecular structure like the number of donors in the H bond, the number of C6H6

rings, etc. [11]. The structural information is represented by 3D molecular descriptors like a positive partial charge structure of solvent-accessible surface area [11].

In this study, the main focus is given to the diversity of datasets in terms of size and nature of datasets, further applying a variety of machine learning algorithms. The novelty of the work is the generation of chemical features from SMILES and testing them as unseen data for the best model. Further, we have tested machine learning algorithms with different hyperparameters and chosen the best hyperparameter for each algorithm that was missing in the previous literature. In previous studies, experiments were conducted with default hyperparameters [7], [12-15]. Also, the model is evaluated on several different evaluation metrics to validate the performance of the best-chosen model.

II. Related Work

BBB is an up-and-coming research area that is widely used in the formation of drug discovery. In the last decades, several approaches to the BBB have been proposed. These approaches are based on ML algorithms and have followed their method of technique. Permitting the literature study, there were several approaches have been proposed which have their methods and techniques. These techniques vary with the number of compounds used and the selection of important features related to these chemical compounds. So, the proposed system will give outcomes in terms of results of model accuracy, sensitivity, specificity, and the robustness of testing scores.

Dai et al. 2021 [1] proposed a feature representation in sequential-based prediction for BBB peptides. In this study, 16 classes of peptide sequence feature descriptors have been used. For finding the best solution three-step method was used. In the three-step model, features were selected based on the F1 score, and Spearman's rank-order correlation and a sequential forward selection strategy were implemented. In this study, many ML models were compared i.e., ERT, XGB, LR, MLP, RF, and SVM. While comparing the results of each model the LR has the best prediction ability to gain an overall AUC and AUPR score of 0.87 with 10-fold cross-validation. But dataset contains only 119 BBPs compound datasets having only seven features for classification and mainly focusing on peptide-based molecular compounds. On the contrary, Zou 2022 [2] uses the physicochemical properties of amino acids and through these amino properties, the author identifies blood-brain barrier peptides (BBPs) and also applies the features fusion method. In this research, SVM was implemented on a dataset that represents peptide sequences based on 100 samples from BBPs, and 100 samples from non-BBPs were used, together with 10 physicochemical characteristics descriptors. For the selection of discriminative features, the Fisher algorithm was used. The highest accuracy, specificity, and sensitivity achieved by the model on the training dataset is 100%, while MCC and AUC are also 1.00, while on the independent dataset, it was 89.47%. The limitation of this work is limited samples were used and they just employed the correlation information between two different types of physicochemical properties. Also, there is a lack of biological experiments to validate the predicted results.

Similarly, Shaker et al. 2021 [7] proposed a LightGBM algorithm that was implemented on a 7162 compounds dataset with BBB permeability in which 5453 BBB+ and 1709 BBB- class with 1119 molecular descriptors of SMILES format. 10-fold cross-validation was implemented after splitting the dataset into 10% testing and 90% training and the results show an accuracy score of 0.89, specificity of 0.77, sensitivity of 0.93, and area under the curve of 0.94 respectively. However, the accuracy can be improved in the future by testing other ML models as it is critical to decide which molecular compound can penetrate from CNS through BBB. However, the use of these many features increases the complexity of the models. Therefore, Alsenan et al. 2021 [9] proposed a model that used the Kernel Principal Component Analysis (KPCA) method for finding descriptors. The author also compares the deep learning (DL) models with ML models and comes with the result that deep learning models show more accurate results than ML models. The FFDNN and CNN achieve accuracy of 100%, specificity of 98.11 and 99.87, and sensitivity of 96.78 and 98.76 respectively. The AUC was also calculated which was 97.7 and 99.71 and Matthew's correlation coefficient was 95.55 and 92.85 respectively. However, the dataset was composed of 2500 molecular compounds with 6,394 molecular descriptors which are small datasets as a large dataset has a direct impact on accuracy and only focuses on the KPCA feature extraction technique.

Furthermore, various variants of ML models are tested by Kumar et al. 2021 [12]. The author proposed an RF-based method for the prediction of the BBB by using chemical peptides. Different algorithms were implemented i.e., DT, RF, LR, KNN, GNB, XGB, and SVC. Three datasets were used in this study i.e., dataset-1 had 269 B3PPs and CPPs respectively. Dataset-2 was having 269 B3PPs and non-B3PPs respectively, while dataset-3 was having 269 B3PPs and 2690 non-B3PPs. The highest accuracy, specificity, and sensitivity were achieved by RF, which is 85.08, 85.08, and 86.97 respectively. Matthew correlation and AUC are also calculated which are 0.51 and 0.93. But the author collected only 465 peptides from the B3Pdb database which is a small dataset of cell-penetrating peptides with 80 selected features.

Also, Liu et al. 2021 [13] proposed the SMOTE technique on 1757 chemical compounds, and the feature descriptors were produced by PaDEL-Descriptor software for nine molecular fingerprints and 2D and 3D descriptors on five-fold cross-validation with 100 iterations. Three algorithms were implemented i.e., SVM, RF, and XGBoost from which RF shows the higher scores in terms of accuracy of 0.910, specificity of 0.867, sensitivity of 0.927, and AUC of 0.957 respectively. But there are a smaller number of descriptors used and this model is not a quantitative approach to identifying which BBB chemical compound can penetrate or not. In comparison, Shi et al. 2021 [14] in this approach, 2354 drug molecules of SMILES format were used with 33 molecular features. 10-fold cross-validation was used and six types of methods were used for training of imbalance dataset i.e., Upsampling, RUS, Weight parameter, SMOTE, SMOTECENN, and ADASYN. The results clearly show that XGBoost outperforms other approaches in terms of precision 0.92, recall 0.96, F1-score 0.94, Accuracy 0.95, specificity

0.93, sensitivity 0.98, and AUC 0.98 respectively. It is worth mentioning that, using too many resampling methods can lead to overfitting and inaccuracy. So, it may harm the model's outcomes.

Saber et al. 2020 [15] proposed a comparative approach to ML algorithms. The algorithms that are implemented in this research study are SVM with linear, polynomial, radial basis function kernels, LDA and QDA, and KNN. The author concludes that a genetic algorithm with SVM outperforms other approaches. It shows an accuracy of 96.23, a specificity of 86.67, and a sensitivity of 98.45. All algorithms compiled on 1593 drug compounds and eight molecular descriptors were generated by sequential feature selection and genetic algorithm. There is a lack of a greater number of features and training the model on fewer features has a great impact on the outcomes. Similar dataset dimension biases also happened in the study proposed by Ciura et al. 2020 [16]. The authors suggested a technique that focuses on micellar electrokinetic chromatography and has 50 2D and 3D molecular descriptors from a collection of market available 45 chemical drugs. MLR and SVM were implemented on a given dataset and showed the same results for prediction, by showing the same results of RMSE and cross-validation of 0.310 and 0.314 respectively. But if a large dataset and a greater number of features were applied this will affect the results as model accuracy directly depends on the size of the dataset.

Moreover, a study proposed by Singh et al. 2020 [17] comprised a novel validation approach of QSAR. In this approach, RF has been implemented on a 605 compounds dataset with 1444 molecular descriptors of 1D and 2D generated by PaDEL software 2.21. In the proposed methodology 10-fold cross-validation QSAR approach was used. Two types of thresholds were employed to divide the dataset. Specifically, threshold-1 was defined as (B/P>=0.6 classified as BBB+ and B/P<0.6 classified as BBB-), while threshold-2 was defined as (B/P>0.6 classified as BBB+ and B/P<0.3 classified as BBB-). Threshold-1 and threshold-2 attained precision of 86% and 87% accordingly. However, this study defined a specific range of thresholds to specify the classes of BBB and only focused on the QSAR approach may other techniques improved the results of the proposed model.

A few other researchers such as Radchenko et al. 2020 [18] implemented an artificial neural network on 529 molecular compounds datasets based on their LogBB values and 100 to 1000 descriptors were generated by using substructures of molecular compounds. The silico LogBB-based model used fragmental substructural descriptors representing the occurrence number of the various substructures. The results show that Q2 has a value of 0.815 and an RMSE of 0.318. However, this research work only concentrates on LogBB values of compounds with a small dataset of compounds. Saxena et al. 2019 [19] presented an ML model for permeability prediction of the BBB. In this study, SVM, KNN, RF, and NB were implemented in 1978 molecular compounds. Physicochemical characteristics, MACCS fingerprints, and substructure fingerprints were included in 1917 feature vectors. With an accuracy score of 96.77 percent, SVM with RBF kernel performs better as compared to other proposed ML techniques. However, the dataset used in this research study has a smaller number of chemical compounds which can affect the results. Roy et al. 2019 [20] proposed an approach SVM, KNN, gradient boost machine, and the statistical importance analysis method used to select 37 descriptors, and a generalized linear model was implemented on it. The results show that SVM surpasses other approaches with an accuracy of 96%, a sensitivity of 99%, a specificity of 87%, a precision of 96%, and an F1 score of 97% respectively. The dataset contains 1800 molecules and was divided into 75% training data and 25% testing data. Rui Miao et al. 2019 [21] proposed three clinical phenotypes data of 1000 molecular compounds were used. DL method, SVM with sigmoid, polynomial, radial basis kernel functions, KNN, and DT were implemented. The dataset was utilized for both training and testing with five-fold cross-validation. As 70 percent is used for training and 30 percent is used for testing. The author concludes that the deep learning method outperforms other ML algorithms in terms of area under curve, accuracy, and F1-score i.e., 98%, 97%, and 92% respectively. Saber et al. 2019 [22] implement SVM, ANN, and KNN models with 1593 drug compounds. For the selection of molecular features, a genetic algorithm was used which generated 8 descriptors. The highest overall accuracy was obtained with both Quadratic Discriminant Analysis and SVM classifiers at 96.23%. But this research study only focuses on ADMET characteristics of compounds, and it is not clear how well the system detects permeable compounds because of the small dataset.

By Analyzing the related work, it is observed that the above research have some drawbacks. First, most of the researchers target datasets having a smaller number of molecular compounds. Second, the number of features used in the research is too small, and vice versa.

III. Methods and Methodology

A. Algorithm for Proposed Study

The algorithm for the proposed study was given below which inputs the dataset and applies preprocessing techniques to the given dataset. After preprocessing the dataset is divided into 90% training and 10% testing purposes and for each molecular compound feature values were generated. By applying the ML models on preprocessed datasets if the molecular compound belongs to class BBB+ it would be updated to the permeable list and else the molecular compound belongs to class BBB- it would be updated to the non-permeable list. For validation of our model, we test it on a test dataset and evaluate these results by using an evaluation matrix.

Algorithm 1: Algorithm for Prediction of BBB Permeability

Input: BBB Dataset

Output: List of compounds into permeable and non-permeable

1 *Initialization*
2 *Input dataset*
3 *Refining an initiated dataset*
4 *Selection of 90% dataset for training data*
5 *Selection of 10% dataset for testing data*
6 *Filtration of data for required features*

```
Applying ML models
For (i=1; i<= n; i++),
    If (Comp(i) are permeable == Yes)
        Updating permeable list
    Else
        Updating non-permeable list
    End If
```

B. Results of Flow Charts

The flow chart of the proposed methodology is discussed in Fig. 1. The dataset contains molecular compounds loaded for the filtration process. After filtration, the dataset was divided into 90% training and 10% testing. After the division of the dataset feature extraction process was applied in which 1D and 2D features were extracted for each compound. The most well-known ML models i.e., SVM, KNN, LR, ANN, and RF were applied to each chemical compound. ML models classify the dataset into two class labels i.e., BBB- (0) non-penetrating list and BBB+ (1) penetrating list. For evaluation and validation of results accuracy, specificity, sensitivity, precision, and recall, the F1-score is applied.

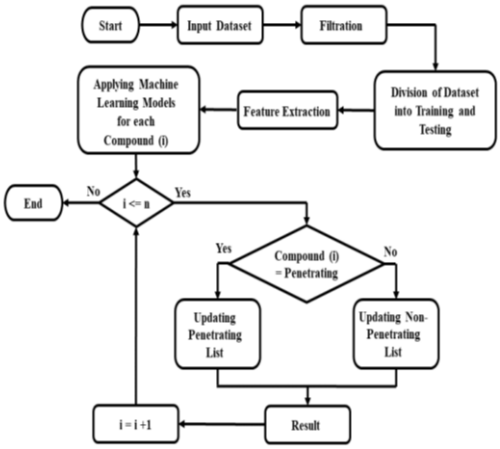

Fig. 1. Flow chart of the BBB permeability prediction model.

C. Data Collection

The datasets used in the proposed study were collected from the online repository of the LightBBB [23] web server which was in SMILES format [24]. In these datasets, the compounds were grouped as BBB+ which belongs to class 1, and BBB- which belongs to class 0. There are numerous descriptors available for expressing the BBB permeability chemicals. It's crucial to pick efficient descriptors for model training to prevent overfitting and poor performance. The training dataset included chemical compounds with 1D and 2D descriptors for each compound after dataset pre-processing.

D. Data Preprocessing

Preprocessing is an essential task for each ML model. To obtain effective results preprocessing is to be done on the dataset. The accuracy of the model is directly impacted by the size of the dataset. In this research study, molecular compounds were compiled with the experimental BBB permeability which leads to compounds belonging to the class of BBB+ and BBB-. A SMILES format was used to prepare each molecule. The BBB+ belonged to class 1 and BBB- belongs to class 0. The dataset was preprocessed to remove duplicates inconsistent compounds and missing structural information data were also removed. The dataset was split into 90% training and 10% testing using ten-fold cross-validation, with each iteration of validation being repeated ten times. For testing purposes, the external dataset contains molecular compounds of which some compounds belong to BBB+ and BBB- classes.

E. Feature Set

The physical and chemical characteristics of substances were described using molecular descriptors as features. As a result, these aspects provide more information to create a reliable BBB model [7].

F. Machine Learning Models

In recent decades, there has been a numerous growth in ML models and most of all have been used in the prediction of the BBB. Some BBB prediction models show good performance with a high accuracy score. Therefore, it was a challenging task to develop an ML model for the prediction of BBB permeability as the dataset of BBB available in biological science gives the impression of being limited [19]. ML approaches are classified as supervised, unsupervised, or reinforced. There are many supervised learning techniques some of them are; SVM, KNN, LR, ANN, and RF are mainly used for classification or regression problems and some deep learning-based algorithms are used for the prediction of BBB. Scikit-learn, a Python-based toolkit, were used in the implementation of the model.

1) Support Vector Machine (SVM): The support vector machine (SVM) is a widely recognized approach in supervised learning, commonly employed for solving classification and regression problems. The proposal was initially forth during the decade of the 1990s and has since been effectively utilized within the fields of bioinformatics and computer-aided diagnosis. Therefore, the SVM classification model was utilized in this research work to analyze the BBB dataset. The support vector classifier (SVC) is implemented using the support vector machines toolkit. The algorithm typically accommodates the supplied attributes' points of information and identifies the optimal hyperplane for classifying the data into two distinct categories [12]. The SVM model's effectiveness over traditional methods can be attributed to its inherent structure and the risk management philosophy it employs. Multiple kernel functions, such as polynomial, linear, radial basis function, and sigmoid, are utilized in Support Vector Machines (SVM) to facilitate the transformation of data into a higher-dimensional space where a distinct separation between classes can be achieved [20]. This research paper examines the performance of the Support Vector Machine (SVM) algorithm, specifically utilizing a linear kernel function, in the context of classification problems.

2) K-Nearest Neighbors (KNN): The k-nearest neighbor (KNN) approach is an example of supervised machine learning that may be applied to both classification and

regression tasks. The technique, which was introduced in 1951, has gained significant popularity over the years as a reliable method for predicting drug penetration and blood-brain barrier (BBB) permeability. Unlabeled datasets are classed through the assignment of a class based on the similarity to neighboring data points. The K-nearest neighbors (KNN) algorithm computes the distances among the information points, namely the feature values, using metrics such as Euclidean distance or Manhattan distance. In the context where a desirable value of k is sought, it is necessary to evaluate multiple neighboring values. The decision to choose one of these neighbors has a significant influence on the general efficacy of the prediction system that is being constructed. Typically, the value of k is limited to an integer not exceeding 20 [19].

3) Logistic Regression (LR): Logistic regression is one of the most popular statistical models that uses a logistic function for binary dependent variables [14]. This algorithm comes under the tree of supervised machine techniques. LR is like linear regression as linear regression is used for regression problems and LR is used for solving classification problems.

4) Random Forest (RF): Random Forest (RF) is a machine-learning technique that is based on decision trees. The bootstrap resampling approach allows for the extraction of multiple samples from the original set of data. Following the choice of specimens, a prediction decision-making structure was constructed for each sample, which was subsequently aggregated through a voting mechanism to obtain the ultimate outcomes. Random Forest (RF) can be utilized to address both classification and regression problems. The primary advantage of the RF model is its ability to mitigate errors resulting from asymmetrical data during training, particularly when there is a substantial disparity between the two types of class compounds. Additionally, this model has superior performance in mitigating the overfitting phenomena, as well as exhibiting enhanced capability in effectively addressing outliers and noisy data [13].

5) Multi-Layer Perceptron (MLP): A multi-layer perceptron (MLP) refers to an artificial neural network characterized by a forward architecture, wherein it transforms a given set of input vectors into a corresponding set of output vectors. The MLP can be conceptualized as a directed graph including multiple tiers of nodes. The subsequent layer is interconnected with the preceding layer. Every node, except the input node, represents a neuron that possesses a non-linear activation function. [16].

6) Light Gradient Boosting Machine (LGBM): Gradient boosting decision trees are a popular machine learning algorithm that combines the strengths of decision trees and gradient boosting. This algorithm iteratively builds an ensemble of weak decision trees, where There are various manifestations of trees, one of which is LightGBM (GBDT). This technique is commonly employed for classification, regression, and efficient parallel training. The LightGBM algorithm is widely recognized as a rapid and efficient variation of the Gradient Boosting Decision Tree (GBDT) technique. The proposed approach involves partitioning the tree into individual leaves and thereafter identifying the leaf that exhibits the highest delta loss. Hence, under the LightGBM framework, the leaf-wise approach can minimize loss to a greater extent compared to the level-wise strategy when expanding on the identical leaf. [7]. The description of all the parameters applied to ML models is discussed below in Table I.

G. Description of Datasets

ML models are implemented on three BBB datasets. These datasets have two classes i.e., class 0 belongs to (BBB-) which specifies non-permeable compounds to BBB and class 1 belongs to (BBB+) permeable compounds to BBB. All the molecular compounds were in SMILES format. The datasets were randomly divided into 90% training and 10% testing data on which ML models were trained.

Dataset 1 contains 1072 (317 BBB+ and 755 BBB-) molecular compounds in SMILES format with a variety of 196 1D descriptors generated from RDKit library [25] which is a Python built-in library mainly used for molecular descriptors. The test dataset contains 266 molecular compounds with 196 1D descriptors that are extracted for each compound.

TABLE I. DESCRIPTION OF PARAMETERS FOR MACHINE LEARNING MODELS

Name	Value	Description
Kernel	Linear	Defines the type of kernel to be used in the algorithm
Random_state	None	Controls the creation of pseudo-random numbers used to shuffle the data used to calculate probabilities.
N_neigbors	7	The numbers of neighbours that k-neighbors queries will by default utilize.
Metric	Minkowski	Metric to employ for distance calculations that, when p = 2, yields the usual Euclidean distance.
P	2	Minkowski metric's power parameter
Solver	liblinear	The optimization problem's algorithm.
Random_state	None	Used when Solver = liblinear
Hidden_layer_sizes	(8, 8, 8)	The number of neurons in the ith hidden layer is represented by the ith element.
Activation	Relu	The buried layer rectified linear unit function's activation function gives the result f(x) = max (0, x)
Solver	Adam	A stochastic gradient-based optimizer is referred to in the solution for weight optimization.
Max_iter	2000	The maximum number of iterations.
n_estimator	100	It means how many numbers of trees can be generated.
Max_depth	None	The depth of trees.
Min_sample_split	2	The needed minimum number of samples
Random_state	42	It means how many times the function calls for the same instance.
Max_features	Sqrt	It means max_features= sqrt(n_features)

Dataset 2 contains 7162 (5453 BBB+ and 1709 BBB-) molecular compounds in SMILES format with 1119 1D and 2D descriptors that are extracted for each compound. The test dataset contains 74 (39 BBB+ and 35 BBB-) molecular compounds with 1119 1D and 2D descriptors that are extracted for each compound [23]. These features were generated using Dragon software (version 7.0.10) [26].

Dataset 3 was constructed by adding more chemical compounds in dataset 2 which contains 9230 (6852 BBB+ and 2378 BBB-) molecular compounds in SMILES format with 1119 1D and 2D descriptors that are extracted for each compound. The test dataset contains 74 (39 BBB+ and 35 BBB-) molecular compounds with 1119 1D and 2D descriptors that are extracted for each compound. These features were generated using Dragon software (version 7.0.10).

H. Evaluation Matrices

1) Confusion matrix: The utilization of the confusion matrix is frequently observed in machine learning to assess and illustrate the performance of algorithms in supervised classification tasks. The matrix in question is a square matrix whereby the rows correspond to the actual class and the columns correspond to the predicted class. The confusion matrix establishes a quantitative assessment of the concordance between observed and forecasted data.

2) Sensitivity: The sensitivity is defined as the percentage of chemical compounds that the model properly classifies as BBB+ [9] and it is calculated by the given formula as shown in Eq. (1).

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \quad (1)$$

3) Specificity: The specificity is defined as the percentage of chemical compounds that the model properly classifies as BBB- [9] and it is calculated by the given formula as shown in Eq. (2).

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (2)$$

4) Accuracy: The accuracy shows the overall performance of the model [9] and it is calculated by the given formula as shown in Eq. (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (3)$$

5) Receiving Operating Characteristics (ROC): The model was graphically evaluated using an ROC curve [27], which is a highly efficient approach for determining how well the model can accurately distinguish between classes [7].

6) AUC: The AUC is used to assess how well the classifier separates the classes by calculating the area under the ROC curve and its output will always be between 0 and 1 [9].

IV. Results and Discussion

Dataset 1 contains 1072 chemical compounds with 196 1D descriptors on which ML models were trained. The dataset was divided into 90% for training and 10% for testing with 10-fold cross-validation and 10 times iterated the whole process. The results are demonstrated below in Table II.

On cross-validation, the training dataset contains 964 chemical compounds whereas the testing dataset contains 108 chemical compounds. The results on Dataset 1 show that we achieved an overall accuracy of the RF of 93.52, an AUC of 0.97, a sensitivity of 95.95, and a specificity of 88.24 on 10-fold cross-validation. The higher AUC value indicates that our model has a high level of accuracy in predicting BBB permeability and is suitable for use in BBB prediction. In contrast with other ML models, the RF model outperforms other ML models as shown in Fig. 2. The results of ML models are demonstrated by using ROC Curve for cross-validation on dataset 1 as shown in Fig. 2.

For validation of the models, we test the ML models on an external dataset 1. Fig. 3 shows ROC curve of ML models for cross-validation of dataset 1. The RF model shows an accuracy of 78.38, an AUC of 0.83, a sensitivity of 94.29, and a specificity of 64.1. Comparing RF results with other ML models clearly shows that RF outperforms in the prediction of BBB permeability compounds.

The LightBBB dataset contains 7162 molecular compounds and was divided into 90% training and 10% testing. After the division of the dataset feature extraction process was applied. The LightBBB dataset contains 1119 1D and 2D descriptors extracted for each chemical compound. These descriptors were generated using Dragon software (version 7.0.10). The most well-known ML models i.e., the SVM, KNN, LR, ANN, and RF were applied to each chemical compound. ML models classify the dataset into two class labels i.e., BBB- (0) non-penetrating list and BBB+ (1) penetrating list. For evaluation and validation of results accuracy, specificity, sensitivity, precision, and recall, the F1-score has been computed. The results are demonstrated below in Table III.

TABLE II. Cross-Validation Results of ML Models on Dataset 1

Models	AUC	Specificity	Sensitivity	Accuracy
SVM	0.91	88.24	83.78	85.19
KNN	0.95	94.12	81.08	85.19
LR	0.91	85.29	85.14	85.19
MLP	0.91	76.47	79.73	78.07
LGBM	0.97	88.24	94.59	92.59
RF	0.97	88.24	95.95	93.52

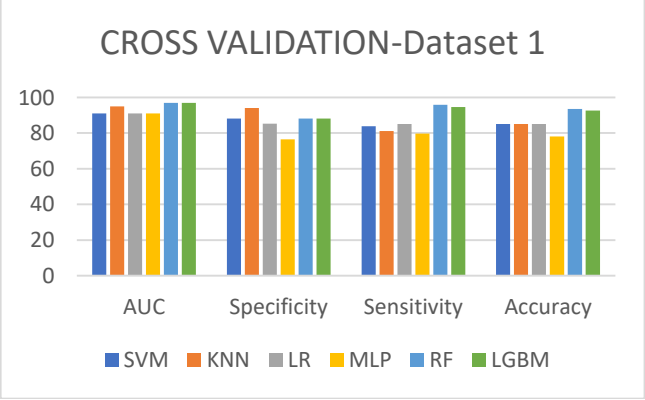

Fig. 2. Performance of ML models on cross validation dataset 1.

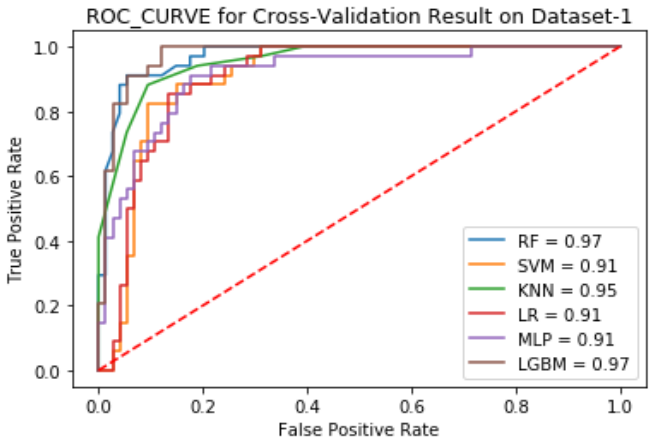

Fig. 3. ROC curves of ML models for cross validation on dataset 1.

TABLE III. CROSS-VALIDATION RESULTS OF ML MODELS ON DATASET 2

Models	AUC	Specificity	Sensitivity	Accuracy
SVM	0.90	93.49	73.78	88.98
KNN	0.92	96.02	62.8	88.42
LR	0.91	94.03	73.78	89.4
MLP	0.92	93.41	75.44	89.12
LGBM	0.94	0.77	0.93	89
RF	0.95	95.12	75.0	90.52

On cross-validation, the training dataset contains 6445 chemical compounds whereas the testing dataset contains 717 chemical compounds. The results on Dataset 2 show that we achieved an overall accuracy of the RF of 90.52, an AUC of 0.95, a sensitivity of 75.0, and a specificity of 95.12 on 10-fold cross-validation. The higher AUC value indicates that the RF model has a high level of accuracy in predicting BBB permeability and is suitable for use in BBB prediction. In contrast with other ML models, the RF model outperforms other ML models as shown in Fig. 4. The results of ML models are demonstrated by using the ROC Curve on cross-validation dataset 2 as shown in Fig. 5. For validation of the models, we test the ML models on an external dataset 2. The RF model shows an accuracy of 93.24, an AUC of 0.96, a sensitivity of 91.43, and a specificity of 94.87. Comparing RF results with other ML models clearly shows that RF outperforms in the prediction of BBB permeability compounds.

Dataset 3 contains 9230 molecular compounds and was divided into 90% training and 10% testing. After the division of the dataset feature extraction process was applied. The dataset contains 1119 1D and 2D features extracted for each chemical compound. The most well-known ML models i.e., the SVM, KNN, LR, ANN, and RF were applied to each chemical compound. ML models classify the dataset into two class labels. The results are demonstrated below in Table IV.

On cross-validation, the training dataset contains 8307 chemical compounds whereas the testing dataset contains 923 chemical compounds. The results on Dataset 3 show that we achieved an overall accuracy of the RF of 90.36, an AUC of 0.96, a sensitivity of 77.73, and a specificity of 94.74 on 10-fold cross-validation. In contrast with other ML models, the RF model outperforms other ML models as shown in Fig. 6. The results of ML models are demonstrated by using the ROC Curve on cross-validation dataset 3 as shown in Fig. 7. For validation of the models, we test the ML models on an external dataset 3. RF shows an accuracy of 91.89, an AUC of 0.94, a sensitivity of 91.43, and a specificity of 92.31. Comparing RF results with other ML models clearly shows that RF outperforms in the prediction of BBB permeability compounds.

While comparing our results with previously published BBB permeability prediction models it seems that our technique outperforms the existing methods. The uniqueness of our technique is the use of optimal hyperparameters and a high density of data. We compared the models by considering all the evaluation parameters i.e., AUC, specificity, sensitivity, and accuracy as shown in Table V.

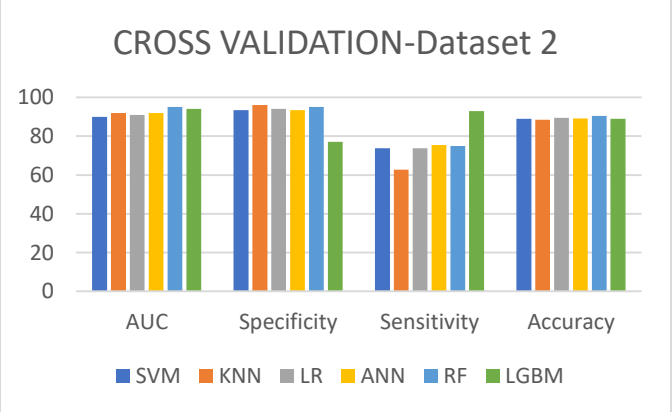

Fig. 4. Performance of ML models on dataset 2.

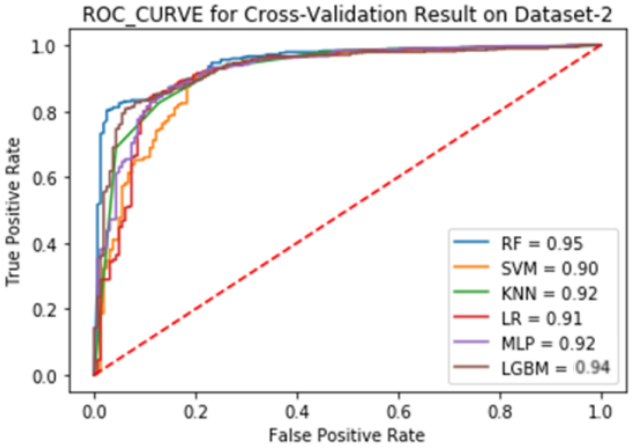

Fig. 5. ROC curves of ML models for cross validation on dataset 2.

TABLE IV. CROSS-VALIDATION RESULTS OF ML MODELS ON DATASET 3

Models	AUC	Specificity	Sensitivity	Accuracy
SVM	0.90	94.31	70.17	88.08
KNN	0.94	93.14	65.55	86.02
LR	0.92	93.72	68.49	87.22
MLP	0.96	91.94	85.09	90.25
LGBM	96	95.18	74.37	89.82
RF	0.96	94.74	77.73	90.36

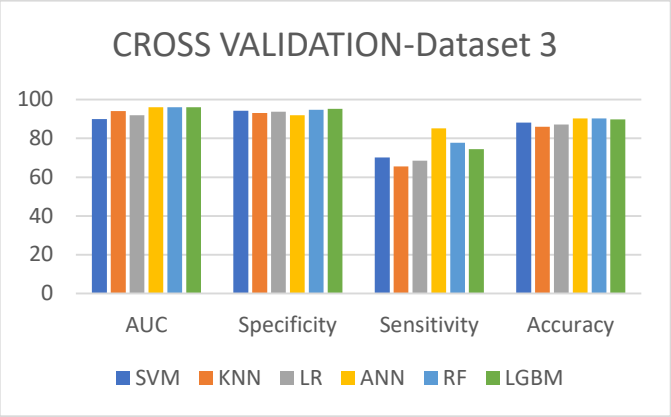

Fig. 6. Performance of ML models on dataset 3.

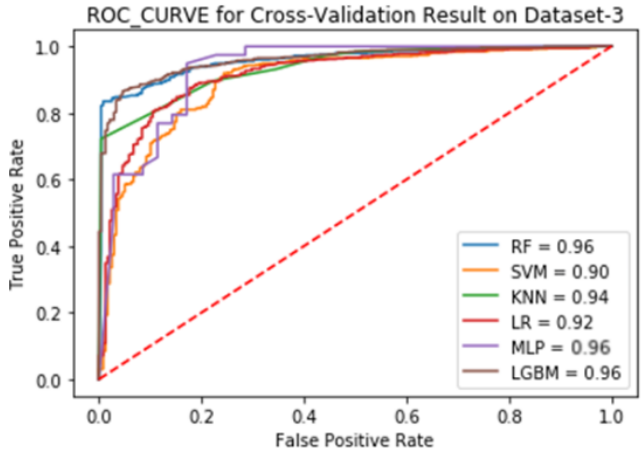

Fig. 7. ROC curves of ML models for cross validation on dataset 3.

TABLE V. COMPARISON OF ML MODELS WITH PREVIOUSLY PUBLISHED BBB MODELS

Reference	AUC	Specificity	Sensitivity	Accuracy
[7]	0.94	0.77	0.93	89%
[12]	0.93	0.85	0.86	85.08%
[13]	0.95	0.86	0.92	91%
[17]	-	0.71	0.92	87%
[21]	0.98	-	-	97%
[28]	0.90	0.83	0.98	94%
[29]	-	0.88	0.85	86%
[30]	0.78	-	-	82%
[31]	-	0.65	0.90	74.7%
[32]	-	0.80	0.82	81.5%
[33]	-	0.72	0.82	95%
[34]	-	0.80	0.72	83%
[35]	-	0.37	0.91	82.5%
[36]	-	0.79	0.84	82%
[37]	0.85	-	-	85%
Proposed Technique	**0.96**	**94.74**	**77.73**	**90.36%**

V. CONCLUSION

In the proposed study five machine learning models were applied to highly accurate small and large datasets with a larger number of features. The dataset is balanced and free from inconsistent and redundant data with accurate class labeling. On the contrary, other ML models were trained on a smaller dataset and fewer features, leading to differing accuracy levels but being unable to compensate for the variety of molecular components. The model uses 10-fold cross-validation with 10 iterations to assure correctness. The dataset contains molecular compounds and features. It was concluded that our ML model RF for the prediction of BBB penetration shows more accurate results on both small and large datasets than other ML algorithms.

The higher accuracy achieved by RF on dataset 1 is 93.52, with an AUC of 0.97, a sensitivity of 95.95, and a specificity of 88.24 on 10-fold cross-validation.

The higher accuracy achieved by RF on dataset 2 is 90.52, with an AUC of 0.95, a sensitivity of 75.0, and a specificity of 95.12 on 10-fold cross-validation. On testing our model on an external dataset RF shows an accuracy of 93.24, an AUC of 0.96, a sensitivity of 91.43, and a specificity of 94.87.

The higher accuracy achieved by RF on dataset 3 is 90.36, with an AUC of 0.96, a sensitivity of 77.73, and a specificity of 94.74 on 10-fold cross-validation. On testing our model on an external dataset RF shows an accuracy of 91.89, an AUC of 0.94, a sensitivity of 91.43, and a specificity of 92.31. The greater the value of AUC, the higher the accuracy of the model will be. Our model outperforms previously reported models.

VI. FUTURE WORK

To encourage future study, it may focus on the features of BBB datasets which can also be increased. It may also focus on applying feature extraction techniques for finding the most important features that were highly influential to the prediction compounds and how these models can be applied to treatments of the brain. Moreover, it may also focus on applying DL models to large datasets and comparing their outcomes with other ML models. Future work may intend to combine two techniques i.e., swarm algorithms with RF to obtain more precise results for this problem.

REFERENCES

[1] R. Dai et al., “BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression,” J Chem Inf Model, vol. 61, no. 1, pp. 525–534, 2021, doi: 10.1021/acs.jcim.0c01115.Ren, Y., et al. (2019). "Data storage mechanism based on blockchain with privacy protection in wireless body area network." Sensors 19(10): 2395.

[2] H. Zou, “Identifying blood-brain barrier peptides by using amino acids physicochemical properties and features fusion method,” Peptide Science, vol. 114, no. 2, 2022, doi: 10.1002/pep2.24247.Ren, C., et al. (2020). "Achieving Near-Optimal Traffic Engineering Using a Distributed Algorithm in Hybrid SDN." IEEE Access 8: 29111-29124.

[3] W. M. Pardridge, “The blood-brain barrier: Bottleneck in brain drug development,” NeuroRx, vol. 2, no. 1, pp. 3–14, 2005, doi: 10.1602/NEURORX.2.1.3.

[4] A. G.-C. opinion in drug discovery & development and undefined 1999, “The design and molecular modeling of CNS drugs.,” europepmc.org, Accessed: Oct. 01, 2022. [Online]. Available: https://europepmc.org/article/med/19649956.

[5] N. Abbott, L. Rönnbäck, E. H.-N. reviews neuroscience, and undefined 2006, "Astrocyte–endothelial interactions at the blood–brain barrier," nature.com, Accessed: Oct. 01, 2022. [Online]. Available: https://www.nature.com/articles/nrn1824.

[6] H. Davson, "History of the Blood-Brain Barrier Concept," Implications of the Blood-Brain Barrier and Its Manipulation, pp. 27–52, 1989, doi: 10.1007/978-1-4613-0701-3_2.

[7] B. Shaker et al., "LightBBB: Computational prediction model of blood-brain-barrier penetration based on LightGBM," Bioinformatics, vol. 37, no. 8, pp. 1135–1139, 2021, doi: 10.1093/bioinformatics/btaa918.

[8] B. Hendricks, A. Cohen-Gadol, J. M.-N. focus, and undefined 2015, "Novel delivery methods bypassing the blood-brain and blood-tumor barriers," thejns.org, doi: 10.3171/2015.1.FOCUS14767.

[9] S. Alsenan, I. Al-Turaiki, and A. Hafez, "A Deep Learning Approach to Predict Blood-Brain Barrier Permeability," PeerJ Computer Science, vol. 7. pp. 1–26, 2021. doi: 10.7717/peerj-cs.515.

[10] M. C. Hutter, "Molecular Descriptors for Chemoinformatics (2nd ed.). By Roberto Todeschini and Viviana Consonni.," ChemMedChem, vol. 5, no. 2, pp. 306–307, Feb. 2010, doi: 10.1002/CMDC.200900399.

[11] H. Hong et al., "Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics," J Chem Inf Model, vol. 48, no. 7, pp. 1337–1344, 2008, doi: 10.1021/CI800038F.

[12] V. Kumar, S. Patiyal, A. Dhall, N. Sharma, and G. P. S. Raghava, "B3pred: A random-forest-based method for predicting and designing blood–brain barrier penetrating peptides," Pharmaceutics, vol. 13, no. 8, 2021, doi: 10.3390/pharmaceutics13081237.

[13] L. Liu et al., "Prediction of the Blood-Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods," Chem Res Toxicol, vol. 34, no. 6, pp. 1456–1467, 2021, doi: 10.1021/acs.chemrestox.0c00343.

[14] Z. Shi, Y. Chu, Y. Zhang, Y. Wang, and D. Q. Wei, "Prediction of blood-brain barrier permeability of compounds by fusing resampling strategies and extreme gradient boosting," IEEE Access, vol. 9, pp. 9557–9566, 2021, doi: 10.1109/ACCESS.2020.3047852.

[15] R. Saber, R. Mhanna, and S. Rihana, "A machine learning model for the prediction of drug permeability across the Blood-Brain Barrier: a comparative approach," pp. 1–15, 2020.

[16] K. Ciura, S. Ulenberg, H. Kapica, P. Kawczak, M. Belka, and T. Bączek, "Assessment of blood–brain barrier permeability using micellar electrokinetic chromatography and P_VSA-like descriptors," Microchemical Journal, vol. 158, p. 105236, 2020, doi: 10.1016/j.microc.2020.105236.

[17] M. Singh, R. Divakaran, L. S. K. Konda, and R. Kristam, "A classification model for blood brain barrier penetration," J Mol Graph Model, vol. 96, p. 107516, 2020, doi: 10.1016/j.jmgm.2019.107516.

[18] E. v. Radchenko, A. S. Dyabina, and V. A. Palyulin, "Towards Deep Neural Network Models for the Prediction of the Blood-Brain Barrier Permeability for Diverse Organic Compounds," Molecules, vol. 25, no. 24, 2020, doi: 10.3390/molecules25245901.

[19] D. Saxena, A. Sharma, M. H. Siddiqui, and R. Kumar, "Blood Brain Barrier Permeability Prediction Using Machine Learning Techniques: An Update," Curr Pharm Biotechnol, vol. 20, no. 14, pp. 1163–1171, Aug. 2019, doi: 10.2174/1389201020666190821145346.

[20] D. Roy, V. K. Hinge, and A. Kovalenko, "To Pass or Not to Pass: Predicting the Blood-Brain Barrier Permeability with the 3D-RISM-KH Molecular Solvation Theory," ACS Omega, vol. 4, no. 16, pp. 16774–16780, 2019, doi: 10.1021/acsomega.9b01512.

[21] R. Miao, L. Y. Xia, H. H. Chen, H. H. Huang, and Y. Liang, "Improved Classification of Blood-Brain-Barrier Drugs Using Deep Learning," Sci Rep, vol. 9, no. 1, pp. 1–11, 2019, doi: 10.1038/s41598-019-44773-4.

[22] R. Saber, S. Rihana, and R. Mhanna, "In silico and in vitro Blood-Brain Barrier models for early stage drug discovery," International Conference on Advances in Biomedical Engineering, ICABME, vol. 2019-Octob, pp. 1–4, 2019, doi: 10.1109/ICABME47164.2019.8940222.

[23] "LightBBB." http://bioanalysis.cau.ac.kr:7030/ (accessed Jul. 29, 2022).

[24] D. W.-J. of chemical information and computer and undefined 1988, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," ACS Publications, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

[25] "RDKit." Accessed: Oct. 01, 2022. [Online]. Available http://www.rdkit.org/.

[26] A. Mauri, A. Srl, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: An easy approach to molecular descriptor calculations," researchgate.net, Accessed: Oct. 01, 2022. [Online]. Available: https://www.researchgate.net/profile/Andrea-Mauri-5/publication/216208341_DRAGON_software_An_easy_approach_to_molecular_descriptor_calculations/links/5da5c4b692851caa1ba601d4/DRAGON-software-An-easy-approach-to-molecular-descriptor-calculations.pdf.

[27] A. B.-P. recognition and undefined 1997, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Elsevier, Accessed: Oct. 01, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422.

[28] Z. Wang et al., "In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods," ChemMedChem, vol. 13, no. 20, pp. 2189–2201, 2018, doi: 10.1002/cmdc.201800533.

[29] "A simple method to predict blood-brain barrier permeability of drug-like compounds using classification trees," ingentaconnect.com, Accessed: Oct. 01, 2022. [Online]. Available: https://www.ingentaconnect.com/content/ben/mc/2017/00000013/00000007/art00010.

[30] F. Plisson, A. P.-M. drugs, and undefined 2019, "Predicting blood–brain barrier permeability of marine-derived kinase inhibitors using ensemble classifiers reveals potential hits for neurodegenerative disorders," mdpi.com, Accessed: Oct. 01, 2022. [Online]. Available: https://www.mdpi.com/403028.

[31] P. Crivori, G. Cruciani, … P. C.-J. of medicinal, and undefined 2000, "Predicting blood− brain barrier permeation from three-dimensional molecular structure," ACS Publications, Accessed: Oct. 01, 2022. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/jm990968+.

[32] S. Doniger, T. Hofmann, and J. Yeh, "Predicting CNS permeability of drug molecules: Comparison of neural network and support vector machine algorithms," Journal of Computational Biology, vol. 9, no. 6, pp. 849–864, 2002, doi: 10.1089/10665270260518317.

[33] I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, "A Bayesian approach to in Silico blood-brain barrier penetration modeling," J Chem Inf Model, vol. 52, no. 6, pp. 1686–1697, Jun. 2012, doi: 10.1021/CI300124C.

[34] A. Guerra, J. A. Páez, and N. E. Campillo, "Artificial neural networks in ADMET modeling: Prediction of blood-brain barrier permeation," QSAR Comb Sci, vol. 27, no. 5, pp. 586–594, May 2008, doi: 10.1002/QSAR.200710019.

[35] L. Zhang, H. Zhu, T. Oprea, A. Golbraikh, T. I. Oprea, and A. Tropsha, "QSAR modeling of the blood–brain barrier permeability for diverse organic compounds," Springer, vol. 25, no. 8, pp. 1902–1914, Aug. 2015, doi: 10.1007/s11095-008-9609-0.

[36] S. Kortagere, D. Chekmarev, W. J. Welsh, and S. Ekins, "New predictive models for blood-brain barrier permeability of drug-like molecules," Pharm Res, vol. 25, no. 8, pp. 1836–1845, Aug. 2008, doi: 10.1007/S11095-008-9584-5.

[37] Z. Gao, Y. Chen, X. Cai, R. X.- Bioinformatics, and undefined 2017, "Predict drug permeability to blood–brain-barrier from clinical phenotypes: drug side effects and drug indications," academic.oup.com, Accessed: Oct. 01, 2022. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/33/6/901/26.