Available online at *www.ujpronline.com*
Universal Journal of Pharmaceutical Research
An International Peer Reviewed Journal
ISSN: 2831-5235 (Print); 2456-8058 (Electronic)

RESEARCH ARTICLE

INTERPRETABLE BINARY CLASSIFICATION MODELS USING XAI AND FEW DESCRIPTORS FOR PREDICTING BLOOD-BRAIN BARRIER PERMEABILITY OF PHARMACEUTICAL COMPOUNDS BASED ON RESAMPLING, CLUSTERING, AND MACHINE LEARNING METHODS

Aubin N'guessan[1], DésiréMélèdje[1], Ludovic Akonan[1], Jean-Louis Kouakou Kouakou[1], Logbo Moussé[1], Melalie Kéita[1], Raymond Kré[1], Nahossé Ziao[2], Eugène Megnassan[1,3,4,5,6*]

[1]*Fundamental Applied Physics Laboratory (FAPL), Nangui Abrogoua University, Côte d'Ivoire.*
[2]*Laboratory of Thermodynamics and Physico-chemistry of the Environment, Nangui Abrogoua University, Côte d'Ivoire.*
[3]*International Center for Theoretical Physics, ICTP-UNESCO, Coastal Road 11, I-34151 Trieste, Italy.*
[4]*Laboratory of Crystallography and Molecular Physics, University of Cocody (Now Felix Houphouet-Boigny), Côte d'Ivoire.*
[5]*Laboratory of Material Sciences, The Environment and Solar Energy and Laboratory of Structural and Theoretical Organic Chemistry, University Felix Houphouet-Boigny, Abidjan 02, Côte d'Ivoire.*
[6]*QLS, ICTP-UNESCO, I 34151 Trieste, Italy.*

Article Info:

Article History:
Received: 3 August 2025
Reviewed: 11 September 2025
Accepted: 17 October 2025
Published: 15 November 2025

Cite this article:
N'guessan A, Mélèdje D, Akonan L, Kouakou JLK, Moussé L, Kéita M, Kré R, Ziao N, Megnassan E. Interpretable binary classification models using XAI and few descriptors for predicting blood-brain barrier permeability of pharmaceutical compounds based on resampling, clustering, and machine learning methods. Universal Journal of Pharmaceutical Research 2025; 10(5): 8-20.
http://doi.org/10.22270/ujpr.v10i5.1420

***Address for Correspondence:**
Eugène Megnassan, Fundamental Applied Physics Laboratory (FAPL), Nangui Abrogoua University, Côte d'Ivoire. Tel: (+225) 01-02-36-30-08
E-mail: *megnase@yahoo.com*

Abstract

Background: Designing pharmaceutical compounds to treat brain diseases, or drugs that interact with biological targets in peripheral organs without penetrating the blood-brain barrier, remains a very difficult task. It is evident that animal models are costly and unproductive; therefore, the pharmaceutical industries and/or regulatory bodies need reliable, accurate and interpretable predictive tools to assess the permeability of pharmaceutical compounds across the blood-brain barrier.
Method: This study proposes the development of artificial intelligence models characterized by greater accuracy and enhanced explanatory capacity, in the context of binary classification of blood-brain barrier permeability of drug candidate compounds. By applying a resampling approach and clustering technique, we developed five distinct artificial intelligence models support vector machine, k-nearest neighbor, classification and regression decision tree, random forest, and gradient boosting machine using only 10 molecular descriptors and a dataset of 1,726 molecular observations (comprising 1,000 originals and 726 synthetic compounds).
Results: Of all the models evaluated, Gradient Boosting Machine had the best 10-fold cross-validation statistics, achieving prediction accuracy (Q), MCC and AUC of 91.04%, 0.82 and 1.0 on the external test set respectively. The gradient boosting machine outputs are explained using Shapley additive explanation approach. This method allows the main modeling descriptors involved in predicting blood-brain barrier permeability to be ranked in order of importance.
Conclusion: Non-animal predictive models were designed to determine whether pharmaceutical compounds can penetrate the blood–brain barrier. The proposed model reached a reliable level of accuracy sufficient to prove extremely useful for virtual screening of large pharmaceutical compounds libraries. It revealed two key indicators for predictions: spatial distribution of atomic charges and electro negativity.
Keywords: blood-brain barrier permeability; curse of dimensionality, explainable AI, logBB, machine learning, QSAR.

INTRODUCTION

The blood-brain barrier (BBB) can be defined as a highly selective, semi-permeable barrier to the circulatory system. Its main role is to maintain the homeostasis of the central nervous system (CNS), by isolating the brain from systemic blood circulation. This isolation protects the CNS from the damaging effects of harmful substances[1]. Although the BBB is defensive in nature, the inability of drug candidates to

cross it remains challenging. Correct administration of these drugs is therefore essential for treating diseases of the central nervous system (CNS), such as Alzheimer's disease, Parkinson's disease or CNS infections, which act directly on specific targets in the brain[2]. Furthermore, pharmaceutical compounds designed to interact with their molecular targets in peripheral organs must not cross the blood-brain barrier (BBB), in order to avoid side effects in the central nervous system (CNS). Many drug candidates have failed to reach the market due to a poor pharmacokinetic profile. In both cases, it is essential to have a clear idea of whether pharmaceutical compound candidates can cross the blood-brain barrier (BBB), which is crucial for the research and development of new treatments.

Experimental determination of brain permeability provides more reliable data. However, its implementation remains complex, time-consuming and expensive, and requires access to highly sophisticated laboratory facilities, particularly in terms of equipment and animal resources[1].This dynamic has led to a growing need for predictive models that are reliable, efficient and easy to use. In this context, quantitative structure-activity relationship (QSAR) tools have proved to be relevant solutions for rapidly and efficiently predicting or estimating the blood-brain barrier (BBB) permeability of drug compounds. Indeed, QSAR relies on theoretical and computational methodologies to predict BBB penetration faster, cheaper and easier. Various model building tools used in QSAR have been satisfactorily implemented by researchers and in these approaches the development of artificial intelligence (AI) and its subfield machine learning (ML) techniques have been successfully used to predict whether a query compound is BBB permeable or not.

To date, several QSAR models that predict BBB permeability, grouped into two main categories, classification and regression, have been satisfactorily implemented by authors using machine learning techniques. As part of the research carried out by Shaker *et al.*[1], classification and regression models were developed with the aim of predicting both the class (permeable or non-permeable) and the concentration ratio of the drug compound in the brain to the compound in the blood, provided by logBB .The researchers designed and refined their models using a selection of machine learning algorithms, namely Light GBM, RF, k-NN, MLR, SVM, AdaBoost, XGBoost and ANN. The best LightGBM regression prediction model called LogBB_Pred for the test set showed an R^2 of 0.61 and mean square error (MSE) of 0.36. Implemented as classification, LogBB_Pred achieved on the independent test dataset an accuracy (Q) of 85%, an MCC (Mathews Correlation Coefficient) of 0.60, and a positive predictive value (PPV) of 1.0[1]. Two years previously, they used 1,119 molecular features for training and testing LightGBM machine algorithm to a large dataset of 7,162 compounds for the BBB permeability prediction with an accuracy of 89%, an area under the curve (AUC) of 0.93, specificity (Sp) of 0.77, and sensitivity (Se) of 0.93, when ten-fold cross-validation was performed[3]. Faramarzi and coworkers constructed two distinct binary QSAR models for logBB permeability prediction using 392 medicinal chemistry structural descriptors with a training set of 921 compounds[4]. The combined predictive performance of the two models obtained achieved an accuracy of 66 %, a sensitivity (Se) of 80%, a negative predictive value (NPV) of 70%, an Sp of 51%, a PPV of 64% and an MCC of 0.4. Singh *et al.*[5], employed three different machine-learning algorithms (RF, MLP, SVM) with descriptors and fingerprints calculated using PaDEL-Descriptorv2.21. They curated a dataset of 605 compounds and trained two classification models, based on two thresholds, with 389 2D molecular descriptors. The best-obtained consensus model achieved good predictive accuracies. Mauri *et al.*, attempted to estimate propensity of compounds to penetrate the BBB by training k-NN machine learning model using a dataset of 3,884 molecules, 2,239 molecular descriptors including 166 MACCS fingerprints, 2048 bits EFCP and 9 features. Their best consensus model showed good evaluation metrics (Q=82.7%, Se=76%, Sp=91.6%)[6]. Yuan *et al.*, developed SVM-based BBB permeability prediction models using a larger dataset of 1,990 compounds with 1,874 molecular descriptors and five different types of fragment descriptors ranging from 307 to 4860 bit. The best prediction accuracies, ranging from 94.9 to 97.5%, were obtained by combining the use of property-based descriptors and fingerprints[7]. Although highly accurate, these models share the same shortcomings: a large or very large number of descriptors, increasing the likelihood of overfitting and unexplainability. Furthermore, the classification models were built using unbalanced datasets, resulting in a high rate of false positives, creating models that failed to save experimental costs[8].

Given these critical deficiencies in building more reliable machine learning models, we implemented hierarchical clustering of descriptors using the *ClustOfVar* algorithm provided by the R programming software to solve the problem of the *curse of dimensionality* caused by the large number of descriptors used. To improve the accuracy of our model, we used a resampling method based on SMOTE (Synthetic Minority Oversampling Technique), which uses information from the data to generate synthetic samples from the minority class[9]. In addition to the performance of model, it is its explicability that is a determining factor for the implementation of computational methods in the field of pharmaceutical research. In this work, the shapley additive explanations (SHAP) values were used to explain the best proposed *black box* model predictions at both local and global levels to identify the significant molecular descriptors that influence BBB permeability prediction.

MATERIALS AND METHODS

Data collection

In the field of QSAR modeling, binary classification is the process of classifying compounds on the basis of two predefined classes. Here, observations or

compounds were divided into two classes using logBB as a criterion: BBB+ (substances that tend to cross the BBB) if logBB≥ -1 or BBB– (substances that do not tend to cross the BBB) if logBB< -1, respectively. In our binary BBB permeability prediction investigation, the dataset was obtained and integrated from a previous study[1]. In their study, they collected the largest logBB data set of 1000 organic compounds separated in a training set of 913 compounds, a validation set of 27 compounds and additional molecules from MedChemExpress (https://www.medchemexpress.com/). In binary classification modelling, the next crucial step is to transform the compounds into vectors of physical and chemical properties. These vectors are determined from the chemical structures represented in SMILES (Simple Molecular Input Layer) format. In this study, for each compound of the final dataset, 919 structural 2- and 3-D descriptors have been calculated using Mordred software; a publicly molecular descriptors calculator. Thus, the entire data set of our study, consisting of a 1000X920 matrix, obtained from Shaker and coworkers' study stands as starting point for the development of our QSAR models for the BBB permeability prediction[1].

Feature selection methods

Increasing the number of descriptors amplifies the effect of the error terms, and consequently increases the correlation between the explanatory variables, with potentially spurious results. In machine learning, feature selection plays a crucial role. It aims to reduce the size of the feature space, speed up the learning process, improve accuracy and make the learning results more explainable. In this work, the hierarchical clustering algorithm, implemented in the *hclustvar* function of the R package *ClustOfVar*, was used for partitioning or clustering the chemical descriptors[i]. Based on the PCAMIX method, a principal component analysis for a mixture of *p1* quantitative ($\{x_1, \dots, x_{p1}\}$) and *p2* qualitative ($\{y_1, \dots, y_{p2}\}$) variables, the *hclustvar* function calculates synthetic quantitative variables that summarize as well as possible the variables in the clusters of the partition obtained. As described by Chavent *et al.*[10], the synthetic variable s_k is defined as the quantitative variable most related to all variables in cluster C_k:

$$s_k = argmax_{u\in\mathbb{R}^n}\left\{\sum_{x_j\in C_k} r^2_{u,x_j} + \sum_{y_j\in C_k} \eta^2_{u|y_j}\right\} = \sqrt{n}\,u^1_k\lambda^1_{Ck} \quad (1)$$

Where u^1_k is the first eigenvector of U_k matrix, λ^1_{ck} represents the first eigenvalue of D_k matrix, n is observation number, r^2 represents the squared Pearson correlation for the quantitative variables and η^2, the correlation ratio for the qualitative variables[10]. The two previous matrices (U_k and D_k) are obtained after singular value decomposition(SVD) of the matrix M_k, obtained by concatenating two matrices corresponding to the quantitative and qualitative data matrices X_k and Y_k, respectively, with their standardized versions $\tilde{X}_k$ and $\tilde{Y}_k$:

$$M_k = \frac{1}{\sqrt{n}}\left(\tilde{X}_k|\tilde{Y}_k\right) = U_k D_k V^T \quad (2)$$

The concept of hierarchical grouping of variables is applied to machine learning and data analysis methods. This methodical approach is based on the construction of a nested tree hierarchy, which is built from a set of variables. These approaches organize descriptors or variables into hierarchical representations in which the clusters at each level of the hierarchy are created by merging the clusters at the level immediately below[10]. To build a hierarchy of $p = p1+p2$ variables, *hclustvar* function optimizes two homogeneity functions. The first homogeneity function h (Eq.3) measures adequacy between the variables in the cluster C_k and its central synthetic quantitative and/or qualitative variable:

$$h(C_k) = \sum_{x_j\epsilon C_k} r^2_{x_j,s_k} + \sum_{y_j\epsilon C_k} \eta^2_{s_k|y_j} = \lambda^1_{ck} \quad (3)$$

The second function, H, defined as the sum of the homogeneities applied to the k clusters of the partition P_k, is obtained as follows:

$$H(P_k) = \sum_k h(C_k) = \lambda^1_{c1} + \cdots + \lambda^1_{ck} \quad (4)$$

Finally, two clusters (C_1 and C_2) are aggregated by choosing the smallest aggregation criterion, d, defined as:

$$d(C_1, C_2) = h(C_1) + h(C_2) - h(C_1 \cup C_2) = \lambda^1_{c1} + \lambda^1_{c2} - \lambda^1_{c1\cup c2} \quad (5)$$

The maximum of the second homogeneity function *(H)* is reached when this procedure is repeated among all the remaining groups. As a result, once the recursive algorithm has been completed, a new partition is generated. The *hclustvar* function also provides a *boostrap* process to obtain the appropriate number of clusters. This is evaluated by the stability of the p-nested partitions of the resulting dendrogram, since each variable is considered as a cluster at the start[10].

Data set standardizing

The standardization of data sets is of crucial importance for the optimal operation of machine learning algorithms. Such algorithms or estimators may exhibit suboptimal performance if the features do not resemble standard normal data (mean of 0 and a standard deviation of 1). Given that the range of values in the raw data varies considerably, the input variables need to be normalized so that higher numerical values do not dominate lower numerical values, while preserving the full informational structure of the data being studied[11]. The normalization procedure is carried out autonomously for each feature, which requires the relevant statistics to be calculated on the samples in the dataset. In this study, the standard Z-score of feature X is computed as follows:

$$Z = \frac{X-\mu}{\sigma} \quad (6)$$

Where μ and σ stand for the average and standard deviation value of descriptor X respectively. In this work, we use *StandardScaler* protocol offered by python *scikit-learn* module[12].

Data balancing

As our dataset is imbalanced, we use the Synthetic Minority Oversampling Technique (SMOTE) provided by python *imbalanced-learn* module to have same ratio of target variable. In most cases, conventional machine learning algorithms are not suited to this type of dataset. This is because they favor samples from the

majority class, which results in poor predictive accuracy for the minority class and limited generalization capability. SMOTE is a popular oversampling approach that handles imbalance by analyzing minority class similarity in near-neighbor feature space and generating new synthetic minority data into the original set. This methodological approach involves inserting synthetic examples along line segments linking all the *k* nearest neighbors of the minority sample, where *k* =5[13]. A synthetic sample, x_s, was generated by selecting a minority instance, x_i, identifying its *k* nearest neighbors using Euclidean distance, and constructing a vector toward one neighbor, x_k. This vector was scaled by a random coefficient $\alpha(0,1)$ and added to x_i[9]. The resulting process of minority class synthesis is summarized by the following equation:

$$x_s = x_i + \alpha(x_i - x_k) \quad (7)$$

Handling data balancing and applicability domain (AD) with statistical methods

QSAR models are mathematical representations that correlate the biological or physicochemical responses of compounds with their structural and molecular descriptors generally expressed as numerical values. Although each numerical value is an individual data point, the data distribution, on the other hand, provides insight into the underlying statistical behavior of the descriptors considered for all molecular observations, thus describing how these values are distributed, concentrated, or shaped in the dataset. In this study, the Synthetic Minority Oversampling Technique (SMOTE) was employed to augment and balance the dataset by generating additional samples for the underrepresented class. To ensure that the synthetic data accurately reflect the distribution of the original experimental data, the Jensen–Shannon Distance (JSD), a robust and widely used statistical measure, was calculated to assess the similarity between the two datasets[9,14].The JSD that measures the degree of overlap or dissimilarity between two distributions P and Q as defined mathematically as follows:

$$\mathrm{JSD(P,Q)} = \sqrt{\frac{1}{2}\mathrm{KL(P \,||\, M)} + \frac{1}{2}\mathrm{KL(Q \,||\, M)}} \quad (8)$$

Where

$$\mathrm{M} = \frac{1}{2}(P + Q) \quad (9)$$

M is a mixed distribution of the P and Q distributions; KL(·||·) represents the Kullback-Leibler divergence. After the quantitative comparison with JSD score, kernel density estimation (KDE) was applied to derive the corresponding probability density functions, enabling a qualitative assessment of the data distributions. This method was successfully employed in previous work[9].

PCA is a linear statistical transformation technique that projects all data (observations and variables) into a lower-dimensional orthogonal space defined by principal components (PCs), which successively capture a significant portion of the information or variance of the original dataset[9]. The PCA bounding box method, categorized among range-based and geometric approaches, is one of several techniques proposed for defining the applicability domain (AD) of QSAR models. An ideal AD approach should delineate the interpolation regions within a multivariate descriptor space, ensuring reliable model predictions for compounds structurally similar to those in the training set[15].

Following the completion of all data processing steps, the data must be split prior to executing machine learning methods. The training and validation sets were selected at random using the *train_test_split* function from the *sklearn* python (version 3.9.2) library. The value assigned to the *test_set size* parameter is 0.2, which is defined as 80% for training and 20% for validation or test subsets with the shuffled option[12].

Model implementation

This work applied five machine learning estimators namely SVM, k-NN, CART-DT, RF, and GBM implemented with the *scikit-learn* package (Python 3.9.2) to model and predict the BBB permeability of drug molecules[12].

1. The support vector machine (SVM) is a supervised learning algorithm applied to both classification and regression problems. The fundamental objective of this process is to determine an optimal hyperplane that separates at most two classes in a *p*-dimensional feature space. This improves the overall ability of the model to generalize when dealing with unknown data[16]. For a dataset of labeled pairs (x_1, y_1) …… (x_n, y_n) where $x_i \in R^p$ and $y_i \in \{-1,1\}$, the decision function or optimal separating hyper plane: $g(x) = w^T x_i + b$ is obtained by estimating the weight vector $w^T = (w_1 \ldots \ldots w_n)$ and the intercept b. For linear SVM, support vectors x_i^* that meet the conditions $w^T x_i^* + b = 1$ and $w^T x_i^* + b = -1$, define the outer limits of the two classes, and the separation distance between these two hyperplanes given by, $\frac{2}{\|w\|}$, is maximized. Once the parameters *w* and *b* have been determined, any input vector x_i can be classified using the function: $sign[w^T x_i + b]$; a positive result assigns the sample to the positive class (BBB+), while a negative value corresponds to the negative class (BBB-).When the training data is not linearly separable, a non-linear SVM projects the input vectors into a higher dimensional feature space using a kernel function. The radial Gaussian basis function (RBF) kernel is a common choice defined as follows:

$$K(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \quad (10)$$

Where γ is the kernel parameter.

2. k-nearest neighbors (k-NN) algorithm is a supervised non-parametric approach used for both classification and regression modeling. Unlike parametric methods, it assumes no underlying data distribution. For a given input, x_j, the algorithm identifies the *k*-nearest training data points according to a predefined distance metric and assigns a class label or predicted value based on the majority vote or average response of these neighbors. In the present study, the nearness is

measured by the Euclidean distance between x_j and x_k as follows:

$$d(x_j, x_k) = \sqrt{\|x_j - x_k\|^2} \quad (11)$$

The 1-NN algorithm represents the simplest form of *k*-NN, where only one neighbor is considered. The input x_j is classified by assigning it the same label as its nearest sample[17].

3. A decision tree (DT) can be defined as a flexible supervised learning algorithm that is used for classification and regression, based on the division of the data. The process of partitioning the data, which is carried out recursively, involves subdividing the dataset according to the feature that allows the most efficient division at each stage. This approach results in a hierarchical tree structure, where internal nodes represent feature-based decisions and leaf nodes correspond to final predictions. Over the last few decades, a set of algorithmic methods dedicated to the construction of decision trees has emerged. The aim of these algorithms is twofold: firstly, to increase the accuracy of the models, and secondly, to adapt to the diversity of data sources[18]. Among them an optimized version of CART (Classification and Regression Tree), implemented as *Decision Tree Classifier,* is available in *scikit-learn python* package.
4. The Random Forest (RF) algorithm is an ensemble-based machine learning approach that aggregates the predictions of multiple decision trees to improve accuracy and minimize overfitting. The tree generation process relies on random sampling of subsets of the training data and features available for each tree. This random process has the effect of increasing model diversity and consolidating generalization performance. During prediction, each tree contributes to a result. The final result is obtained by averaging the predictions in regression tasks or by applying majority voting in classification. As a result, Random Forest models demonstrate greater robustness and generalization capability than individual decision trees.
5. The concept of "boost" refers to a set of algorithms designed to optimize the predictive capabilities of a learning system by increasing its performance, from weak to strong. Intuitively, these algorithms merge a number of weak performance learnings into a single strong performance model, significantly improving the results. Thus, boosting algorithms work by sequentially training a set of weak learning models and combining them for prediction where subsequent learners focus more on the errors of previous learners improving prediction performance to ultimately obtain, through this model, strong learners. The superiority of boosting lies in its serial learning nature, which enables excellent approximation and generalization[19]. Among the various kinds of boosting approaches, the highly effective tree boosting methods, Gradient Boosting Machine (GBM), have been used for binary classification-based QSAR models of logBB permeability predictions.

Binary classification assessment methods

Internal 10 fold cross-validation scheme was applied to the training dataset in order to identify the models with the best predictive performance. The final evaluation of the classifiers was carried out using an independent test set, the aim of which was to assess their generalization capability. For binary classification performance evaluation, several scalar measures were considered, including accuracy (Q), precision (Pr), recall (Re), specificity (Sp), F-score (F) and Matthews correlation coefficient (MCC). These measures are defined mathematically as follows:

$$\mathrm{Q}(\%) = 100 \times \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \quad (12)$$

$$\mathrm{Pr}(\%) = 100 \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (13)$$

$$\mathrm{Re}(\%) = 100 \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (14)$$

$$\mathrm{F}(\%) = 100 \times \frac{2 \times \mathrm{TP}}{2 \times \mathrm{FN} + \mathrm{TP} + \mathrm{FP}} \quad (15)$$

$$\mathrm{Sp}(\%) = 100 \times \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (16)$$

$$\mathrm{MCC} = \frac{\mathrm{TN} \times \mathrm{TP} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (17)$$

The quantities TP, FN, TN and FP are defined as true positives, false negatives, true negatives and false positives respectively. Beyond standard metrics, model performance was also assessed using the receiver operating characteristic (ROC) curve and its associated area under the curve (AUC), which provides a summary measure of classification accuracy.

Models explainability

"Black box" models are characterized by their inability to provide decisions that can be clearly interpreted and/or explained. Transparent models, on the other hand, have the ability to allow direct understanding of their internal reasoning. In drug research, for example, the explainability of models plays a crucial role, as the decisions taken must be justifiable. Interpretability is therefore just as important as the accuracy of predictions[20]. In recent years, Lundberg and Lee have developed a unified framework for interpretability prediction namely SHAP (SHapley Additive exPlanations)[21]. This explanation model suggests taking an additive feature contribution method as a weighted sum of the binary features:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \quad (18)$$

With M, the number of simplified input features, $z' \in \{0,1\}^M$ and the shapely values (weights) are defined as follow:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M-|z'|-1)!}{M!}[f_x(z') - f_x(z'\backslash i)] \quad (19)$$

Where f is the original model, $|z'|$is the number of non-zero entries in z', and $z' \subseteq x'$represents all z'vectors where the non-zero entries are a subset of the non-zero entries in a simplified input x'. According to Lundberg et *al.*, only one possible explanation model g satisfies equation 18 and three properties (local accuracy, missingness and consistency)[21]. SHAP, based on a concept from the field of cooperative game theory (Eq.19), provides model transparency for any machine-learning algorithm to define feature influence at the individual prediction level (i.e., local interpretability). In the perspective of BBB permeability prediction, Shapley values are assigned to each feature in order to estimate their importance and the direction of their impact for a particular prediction. Strongly positive SHAP values indicate that the molecular descriptor helps to predict molecules that cross the BBB, whereas strongly negative SHAP values indicate that the molecular descriptor helps to predict molecules that do not cross the BBB. Several variants of the SHAP algorithm have been reported: Kernel SHAP (model-agnostic), Tree SHAP (specifically applicable to model derived from trees) and Deep SHAP (specialized for deep learning models). Model explanation and analysis were performed using the SHAP package in Python (v. 0.43.0), which enables the quantification of each contribution of feature to model predictions[22].

RESULTS AND DISCUSSION

Data set distribution analysis

According to data gathered from published studies, a total of 1,000 compounds linked to BBB permeability were compiled with 137 BBB- and 863 BBB+ after data separation. The chemical diversity of the compounds used in both the training and external validation sets was extensively analyzed by Shaker and coworkers to support the construction of a robust and reliable binary classification model. Thus, similar compounds were discarded on the basis of Tanimoto similarity, preserving the uniqueness of the compounds and avoiding biased and over fitted models with an abundance of similar compounds[1]. The success of the machine-learning algorithms depends on the quality of the data in order to obtain a generalized predictive model of the classification problem. Therefore, to ensure optimum data quality and optimize the performance of the machine learning models, a rigorous normalization procedure was implemented. This involved centering each feature around a mean equal to zero and scaling it to a standard deviation of one. Like redundancy and non-standardizing, unbalanced datasets have a serious impact on the optimization performance of binary classification machine learning models. Thus, using the SMOTE method, we obtained 1,726 compounds (1000 evaluated chemicals and 726 synthethic compounds), divided into two balanced groups for the implementation of models to assess the ability of drugs to penetrate the BBB. Three hundred and forty-six (346) molecules (166 BBB- and 180 BBB+) serve as the test set to evaluate the generalization ability and reliability of the model and the remaining 1,380 molecules were used to build the prediction models, which were divided into 697 BBB+ and 683 BBB.

Feature involved in models

Feature selection methods have been used for dimension reduction, and this technique is essential for mitigating the effects of the *curse of dimensionality* and improving the performance of algorithms[9]. The tree-like diagram, constructed using squared Pearson correlation coefficients, was divided horizontally to form forty (40) clusters of descriptors, each containing strongly correlated variables sharing similar information[10]. After grouping the descriptors, we selected one variable from each group, i.e. the one that best correlated with its centroid. Thus, the gain in cohesion was 50.77%. This percentage measures the correspondence between the descriptors of the cluster and its centroid (the first PC obtained by applying PCA to it). *ClustOfVar* algorithm for detecting a partition extracted from a tree-like diagram obtained by hierarchical representation of quantitative variables have been used. Each stage of the hierarchy is thus created by successively merging the clusters of the lower stage. This merge is initiated by the lowest stage with the most homogeneous partition, i.e. the partition whose cluster contains only one variable[10]. The difference between PCA and our clustering method built using the *ClustOfVar* algorithm is that the centroids of the obtained clusters can be correlated. Therefore, the correlation matrix of the 40 descriptors was performed to detect residual correlations, which could negatively affect the models by increasing variance. Finally, after removing redundant, non-informative and irrelevant descriptors from the original high-dimensional dataset, we obtained 10 informative descriptors for BBB permeability prediction. The correlation matrix of the ten most informative and best selected descriptors was then constructed using the method described by N'guessan *et al.*[9]

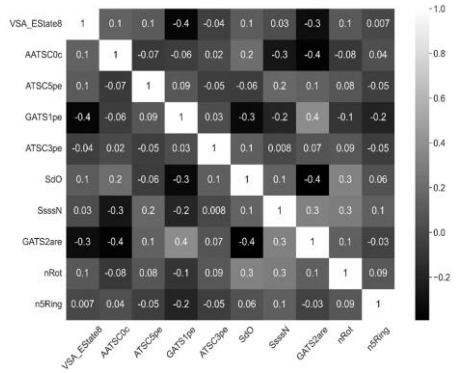

Figure 1: Correlation matrix of 10 non-redundant and informative selected descriptors.

It appears highly improbable that any of the selected descriptors will be correlated with another, with all R^2 measuring no more than 0.40 (Figure 1). This suggests that multi-collinearity, a consequence of the *curse of dimensionality*, has been addressed.

Table 1 shows the 10 molecular descriptors obtained using a data mining procedure that integrates hierarchical clustering and correlation-based analysis. This approach allows us to assess both the strength and direction of relationships between variables. These descriptors are similar to those used in previously published qualitative QSAR models designed to predict the permeability of pharmaceutical compounds across the blood-brain barrier[1,23]. As shown in Table 1, a large proportion of the obtained descriptors belong to the autocorrelation descriptor class, with descriptors VSA_Estate8, AATSC0c, ATSC5pe, GATS1pe, ATSC3pe, and GATS2are.

Table 1: Molecular descriptors using in this work.

No	Descriptors	Description
1	VSA_Estate8	Van der Waals surface area_electrotopological State 8
2	AATSC0c	Averaged and centered Moreau-Broto autocorrelation of lag 0 weighted by Gasteiger charge
3	ATSC5pe	Moran autocorrelation - lag 2 / weighted by Sanderson electronegativities
4	GATS1pe	Geary coefficient of lag 1 weighted by Pauling EN
5	ATSC3pe	Centered Moreau-Broto autocorrelation of lag 3 weighted by Pauling EN
6	SdO	Sum of atom-type E-State: =O
7	SsssN	Sum of atom-type E-State: >N-
8	GATS2are	Geary coefficient of lag 2 weighted by Allred-Rochow EN
9	nRot	Rotatable bonds count
10	n5ing	Number of 5-membered rings

In molecular modelling and QSAR, it is common practice to use autocorrelation descriptors to describe how the physicochemical properties of molecules vary according to their spatial distribution structure. These descriptors are derived from a conceptual partitioning of the structure of the molecule and the application of an autocorrelation function. Typically, spatial autocorrelation descriptors are computed by considering the atoms of a molecule as discrete spatial points, with an atomic property assigned to each point. The descriptors are then weighted according to physicochemical parameters such as atomic weight, volume of van der Waals of considered atom, atomic electronegativity, atomic polarizability, atomic charge, or covalent radius[24]. In this work, E-state indices (SdO and SsssN) that encode electronic and topological environment of each atom are used as the most informative descriptors in classification-based QSAR models predicting blood-brain barrier permeability (logBB)[25].

Class imbalance handling

As indicated in the previous study, we use the SMOTE approach to solve the class imbalance problem. We then proceed to a quantitative and qualitative assessment of its impact on the initial or original dataset[9]. A quantitative comparison between the synthetic (Ds) and true (Do) probability distributions was performed using the Jensen–Shannon Distance (JSD) across the ten most informative molecular descriptors. The resulting JSD scores, reflecting the degree of similarity between synthetic and original distributions, are presented in Table 2. As outlined in Table 2, the two distributions – the original one prior to SMOTE and the synthetic one following SMOTE algorithm are indistinguishable, with all JSD values approaching 0. Before modeling, a qualitative comparison between the two minority class distributions was performed for two informative descriptors selected based on their JSD scores (minimum and maximum) to further study the impact of the SMOTE algorithm.

Table 2: JSD to quantitatively compare the original (Do) and synthetic (Ds) distributions of the minority class dataset for each descriptor involved in models.

Descriptors(D)	JSD(Do, Ds)
VSA_EState8	0.10
AATSC0c	0.13
ATSC5pe	0.09
GATS1pe	0.08
ATSC3pe	0.08
SdO	0.06
SsssN	0.08
GATS2are	0.07
nRot	0.11
n5Ring	0.09

Kernel density estimation (KDE), a technique that estimates and plots probability distribution functions, was applied[9]. As demonstrated in Figure 2, there is a significant overlap in the probability distribution functions between the original and synthetic distributions. This indicates that the SMOTE algorithm effectively preserves data quality and maintains the local structure of features[9]. The finalized dataset encompasses a total of 1,726 molecular observations, which are associated with 10 informative descriptors. As illustrated in Table 3, a concise overview of the samples employed in the experimental procedures, both for training and testing, is provided.

Hyperparameter values of ML models

Following the identification of the optimal methods for dimensionality reduction and handling class imbalance problem, ML-based QSAR models for BBB penetration prediction were implemented. Prior to model development, a deep grid search with 10-fold cross-validation was constructed to adjust the hyperparameters of each classifier, except for the RF model where default values were used[9]. Table 4 provides the best hyperparameters that gave the best QSAR performance based on the statistical training of the models used.

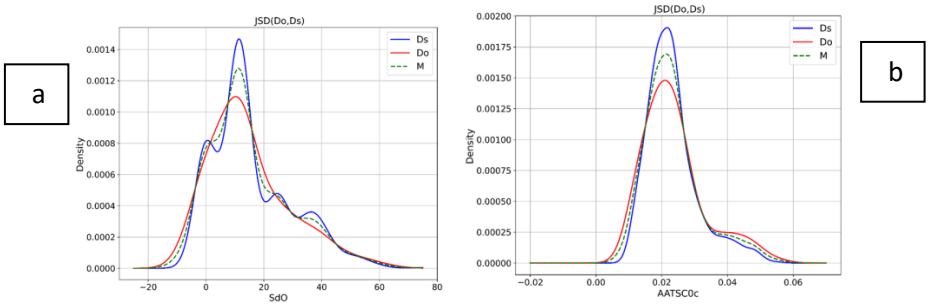

Figure 2: Qualitative assessment of the impact of SMOTE algorithm on our dataset with KDE method. (a, b) plot of density functions of the descriptors SdO and AATSC0c for the original (Do), synthetic (Ds) and mixture (M) distributions colored blue, red and green respectively.

Performance of ML models

In this study, classification models were constructed utilizing five distinct machine learning (ML) algorithms. After model construction, both internal and external validation schemes were employed to assess model reliability. 10-fold cross-validation scheme was used for internal validation using training dataset consists of 1,380 molecular observations divided into 697 BBB+ and 683 BBB- compounds (Table 3) and 10 descriptors. In this procedure, the training dataset is stochastically partitioned into ten distinct splits. For each iteration, nine of these splits are used for training the model, while the remaining split is used to assess its performance. This process is repeated ten times so that every split was used once for validation, and the average of all results was taken as the final performance measure.

Table 3: Details of the training and testing datasets used in this work.

	Sample size				Descriptors	
	Before SMOTE	After SMOTE	Training set	Test set	Before ClustofVar	After ClustofVar
BBB+	863	877	697	180		
BBB-	137	849	683	166	919	10
Total	1000	1726	1380	346		

Table 4: Optimal hyperparameters obtained for all classification-based ML models.

Classifiers	Hyperparameters	Values
SVM	C	3.0
	Kernel	rbf
	Tolerance	1e-3
	Gamma	0.1
k-NN	k	3
	Weights	uniform
	p	2
CART-DT	criterion	gini
	max_depth	5
	min_samples_leaf	1
	max_features	None
	min_samples_split	5
	min_impurity_decrease	0
RF	n_estimators	100
	criterion	gini
	max_depth	none
	min_samples_split	2
	max_features	sqrt
	min_impurity_decrease	0.0
	bootstrap	True
GBM	n_estimators	300
	max_depth	9
	max_features	sqrt

The performance of ML binary classifiers was compared based on an internal validation scheme using the training dataset. As shown in Table 5, the results of the five ML binary classification models are presented according to the evaluation measures defined and used in equations 12 to 17. The table clearly reveal that the decision tree classifier exhibits the weakest performance across all evaluation measures. Conversely, SVM and *k*-NN classifiers appear to have similar performance, although they differ in their ability to correctly classify BBB+ molecules. According to Table 5, GBM model outperforms all other binary classifiers in terms of Q (92.90%), Pr (94.84%), Re (90.61%), Sp (92.65%), and MCC (0.86).

The next best classifier is RF with Q = 91.74% for10-fold internal cross-validation. Considering Q (%), Pr (%), F (%), and MCC, the classifiers were ranked in descending order of performance as GBM, RF, k-NN, SVM, and CART-DT. Based on the evaluation metrics, the GBM estimator was identified as the most effective logBB permeability prediction model of drug molecules. To further confirm its stability and predictive strength, GBM model was tested on an independent dataset that had not been used during training. As shown in Table 5, he GBM classifier achieved a correct classification rate of 91.04% on the test dataset, accurately identifying 88.89% of the BBB+ molecular observations.

Table 5: Performances of binary classification models for five machine learning methods on 10-fold internal cross-validation.

Validation scheme	ML Models	Q (%)	Pr (%)	Re (%)	F (%)	Sp (%)	MCC
Internal	SVM	85.73	89.81	80.22	90.89	84.61	0.72
	k-NN	86.23	94.38	76.68	95.60	84.53	0.74
	CART-DT	77.68	85.07	67.35	88.04	74.76	0.57
	RF	91.74	93.19	89.92	93.52	91.50	0.84
	GBM	92.90	94.84	90.61	95.09	92.65	0.86
External	SVM	84.10	88.34	80.00	88.55	83.97	0.69
	k-NN	83.24	96.21	70.56	96.99	81.41	0.69
	CART-DT	72.25	88.18	53.89	92.17	66.90	0.49
	RF	89.31	92.81	86.11	92.77	89.34	0.79
	GBM	91.04	93.57	88.89	93.37	91.17	0.82

Table 6: Comparison with previous classification-based ML models.

No.	Model Name	Dataset shape (line x column)	Q(%)	SE / RE(%)	SP(%)	MCC	AUC
1	Light GBM[1]	1000 X 396	85.00	42.00	99.00	0.60	-
2	Light GBM[3]	7162 X 1119	90.00	85.00	94.00	-	0.90
3	LS/CU[4]	921 X 392	66.00	80.00	51.00	0.40	-
4	Consensus KNN[6]	3884 X 2240	82.70	76.00	91.60	-	-
5	SVM[7]	1990 X 1874	92.90	93.70	90.70	0.83	-
6	Model proposed	1726 X 10	91.04	88.89	91.17	0.82	1.00

The Ability of recommended classifier to recognize false-alarm molecule on the test dataset is good with Sp = 91.17 %; and the Pr, F and MCC scores are 93.57%, 93.37% and 0.82 respectively.

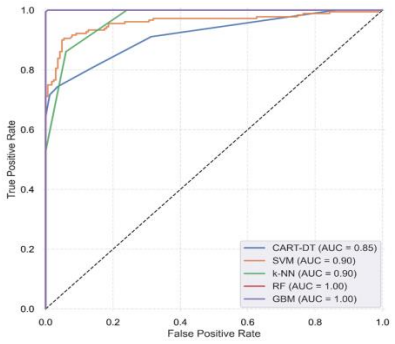

Figure 3: Receiver operating characteristics (ROC) curve for the test set using five machine learning methods and AUC (the area under ROC curve).

As illustrated in Figure 3, the ROC curves of each classifier on the external test set are presented. It appears that all classifiers demonstrate superior performance in comparison to the random classifier, which is represented by the diagonal line $f(x) = x$. Therefore, the area AUC scores are as follows: 0.85 for the CART-DT classifier, 0.90 for the SVM and *k*-NN classifiers, 1.0 for the RF and GBM classifiers. The study exhibits the effectiveness of our classifiers, particularly ensemble models, in accurately distinguishing between blood-brain barrier permeable and non-permeable pharmaceutical compounds. This result therefore highlights the crucial role of effective feature engineering methodologies in improving model accuracy and overall predictive performance.

Applicability domain

In this study, binary classification-based ML models were designed to evaluate the blood-brain barrier (BBB) penetration potential of pharmaceutical compounds encompassing broad-spectrum chemical diversity. Since QSAR models are not universal, defining the applicability domain AD is essential to distinguish reliable interpolations from less reliable extrapolations. Following validation, the applicability domain of GBM classifier was analyzed through the PCA bounding box method. The first three principal components (PCs), derived from the ten most informative descriptors obtained, capture more than half of the total variance of the dataset[26]. As can be seen in Figure 4, test set observation points are colored in red and the molecular observations of training dataset are colored in blue. An analysis of the prediction reliability using PCA bonding box shows that only a few molecular observations reside outside the AD. This incorrect prediction could be a consequence of the oversampling method implementing by SMOTE algorithm that inserts synthetic examples on the original dataset. Consequently, it is assumed that the predictions for 3 of the 346 compounds will be incorrect, thereby

suggesting that the selected model captures the majority of the information present in the ten informative features. Furthermore, this result reveals that the test set molecular observations exhibit a structural similarity greater than 99% to those in the training set molecular observations, confirming strong overlap between the two datasets and reliable representation within the applicability domain of classification-based ML models.

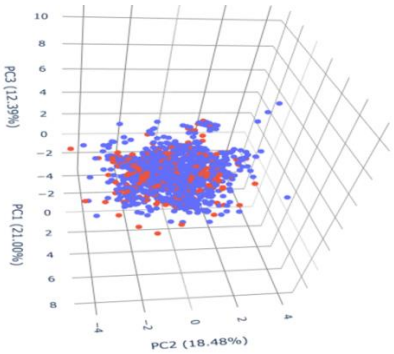

Figure 4: Applicability domain assessment for training (blue) and test data (red) observations with the first three PCs representing more than 50% of total variance. PC1 captures 21.00% of the total variance in the data, PC2 captures 18.48%, and PC3 captures 12.39%.

In conclusion, we can say that our models can be used with high accuracy to predict whether a compound can effectively penetrate the brain or not.

Model comparisons

Using a larger number of descriptors (features) when training machine-learning models can introduce several important drawbacks especially in drug discovery. Although the models display satisfactory predictive performance, binary classification approaches that incorporate a large number of descriptors are susceptible to overfitting and exhibit limited generalization to unseen data. This is because high-dimensional descriptors often contain redundant or highly correlated variables, a phenomenon that is often referred to as the *curse of dimensionality*. Therefore, in our study, we trained our models with the few descriptors make it simpler to extract biological or chemical meaning from model outputs. Thus, the predictive capabilities of our binary classification-based GBM model exceed those of previously published ML models for blood–brain barrier (BBB) penetration prediction, highlighting its improved accuracy and robustness (Table 6).

Explaining ML model

In recent years, the need for interpretable models has been increasingly recognized in research, industry, and regulatory contexts[27]. Given the potential risks of deploying opaque or "black box" models in clinical and preclinical applications, explainable artificial intelligence (XAI) approaches have become a top priority. The practice of XAI models is essential to justify predictive results and ensure the reliability, safety and transparency of preclinical or clinical decision-making[28]. In order to meet this objective, SHAP was developed and validated to interpret how the proposed GBM estimation algorithm predicts class labels for chemical compounds. Here, Tree-SHAP, a variant of SHAP algorithm, is applied to study the effect or influence of selected informative descriptors on the prediction of chemical class (BBB+ vs BBB-) of pharmaceutical compounds studied with GBM model. Thus, multiple visualization techniques can be applied to examine and illustrate the distribution of SHAP values, providing both local (instance-level) and global (model-level) explanations of the predictive behavior. As illustrated in Figure 5(a), a sample-wise SHAPE summary plot is employed to demonstrate which features are the most significant overall. In this plot, the *x*-axis represents the Shapley values, whereas the *y*-axis lists the descriptors and their corresponding value distributions, sorted according to their mean absolute Shapley values, highlighting the relative importance of each feature. Each point represents a Shapley value corresponding to a specific molecular observation, with the color indicating the magnitude of the associated descriptor. As shown in the color bar, sky blue indicates the lowest values and magenta the highest. The descriptors are displayed along the *y*-axis in descending order of importance, reflecting their relative contribution to the model's predictions[29]. With GBM classifier, averaged and centered Moreau-Broto autocorrelation of lag 0 weighted by Gasteiger charge (AATSCOc), geary coefficient of lag 2 weighted by Allred-Rochow EN (GATS2are) and geary coefficient of lag 1 weighted by Pauling EN (GATS1pe) are the top three important descriptors.

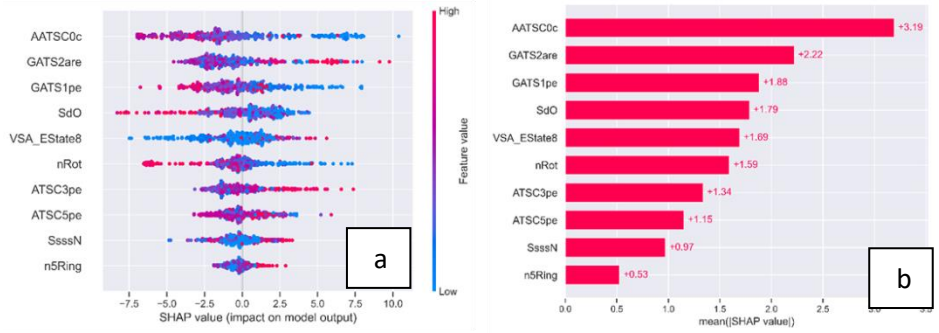

Figure 5: Local and global explanation of the GBM classifier using SHAP values corresponding for the test dataset.
(a) sample wise SHAP values; (b) mean SHAP value for each selected descriptor.

Furthermore, Figure 5(b) shows the MAS (mean absolute SHAP) value for specific informative descriptors, serving as a metric of feature importance. MAS values provide an effective measure of the effect of selected informative descriptors in decision making to classify a query compound. The more the mean absolute values, the more the selected descriptors influence overall in separating compounds into class BBB- *vs* class BBB+. This will help interpret the sample-wise SHAP values shown in Figure 5(a). Shapley values are a means to describe the influence of selected descriptors in the model prediction, and the direction of this influence can be determined using the positive or negative values assigned to a particular descriptor for each molecular observation[29]. As the SHAP values indicate the direction of the predictions (towards BBB- for negative values and towards BBB+ for positive values), it can be concluded that compounds with higher AATSCOc values decrease the probability of the BBB+ class, while lower values of this descriptor appear to increase the probability of the BBB+ class. In the other words, it appears that highly charged molecules, such as macromolecular drugs, recombinant proteins and nucleic acid, are not likely to cross the blood-brain barrier[30]. Electronegativity auto-correlation is a graph-based molecular descriptor that quantifies how the electronegativity values of atoms in a molecule are correlated at a specific topological distance (number of bonds apart)[24]. In this work, GATS2are and GATS1pe are identified as the next most influential descriptors for predicting BBB permeability, reflecting electronegativity auto-correlation. These two descriptors are slightly correlated with $R^2 = 0.4$ because they reflect the same properties calculated in two different scales[31]. Electronegativity scale formulated by Pauling analyses or reflects single or multiple bond dissociation energies. And, as we can see in Figure 5(a), GBM classifier concludes that lower values of GATS1pe have high SHAP values. Therefore, the likelihood of BBB+ permeability increases as the amount of energy required to break a bond decreases. Electronegativity molecular property implemented in GATS2are uses the formulation of Allred and Rochow electronegativity that measures an atom's tendency to attract electrons in a chemical bond. It defined in terms of the electrostatic force or Coulombic attraction exerted by the effective nuclear charge (Z_{eff}) on valence electrons located at the covalent radius (r_{cov}) of the atom[31]. Therefore, the higher the effective nuclear charge of the atom, the higher the electronegativity. If atoms with a high electronegativity value are often connected at a distance d = 2 Å in the molecular graph, the value of the descriptor will be high and the autocorrelation will be strong, which will increase the probability of BBB+. Whilst the general trend between the top three descriptors values and the Shapley values allows for the identification of linear relationships, saturation effects emerge in the impact of these characteristics on the model's predictions.

Limitations of the study

The first limitation of this study is its dependence on a single dataset from Shaker *et al.*[1] Although GBM classifier achieved strong predictive performance, its clinical or preclinical relevance remains limited by the dataset's size and diversity. While the dataset provides a solid foundation for algorithm development, its restricted scope warrants caution when generalizing these findings to real-world settings[3,6]. The second limitation of this QSAR investigation pertains to the quality of the underlying data, which is inherently dependent on the accuracy and reliability of the molecular descriptors employed in model development. Molecular descriptors mathematically capture the chemical information embedded in molecular structures. As molecules may exist in various conformations, choosing the correct conformer is as important as selecting suitable descriptors, since conformational changes can alter descriptor values. Therefore, accurate molecular geometries are fundamental to constructing reliable QSAR models, particularly those employing quantum-chemical or 3D descriptors, as they ensure several benefits: (i) enhances data quality and model robustness; (ii) reduces overfitting and training complexity by avoiding the time-consuming hyper-parameter adjustment process; (iii) provides better biological relevance for structure–activity relationships and (iv) improves comparability and transparency of model development[32]. Another notable limitation of this study stems from the approach used to balance the dataset. Specifically, the application of the SMOTE algorithm, while effective in mitigating class imbalance, may introduce synthetic samples that do not fully represent the actual data distribution or the fit between the training and test data. This may increase the risk of overfitting and potentially distort class boundaries, thus affecting the generalizability of the model to unobserved data.

CONCLUSIONS

In this study, we developed non-animal predictive models to assess the ability of drug or pharmaceutical compounds to penetrate the blood–brain barrier (BBB), providing an alternative to traditional *in vivo* testing. The construction of robust and accurate predictive models necessitates the use of a dataset that is sufficiently large, chemically diverse, and well-balanced across classes. Using *ClustOfVar* algorithm and correlation matrix technique, only 10 molecular informative descriptors of 1,726 (original and synthetic) compounds with different structures were used. Then, five binary machine learning classifiers (SVM, k-NN, CART-DT, RF and GBM) used to predict whether a query compound is BBB permeable or not were developed and validated using 10-fold cross-validation. Since with a large or very large number of descriptors the risk of likelihood of overfitting and unexplainability increases, our models were trained with the few descriptors to make it simpler to extract relevant biological or chemical meaning from model outputs. The accuracy of these classifiers ranged from 77.68 ($\pm$1.25) to 92.90% ($\pm$0.72), and the MCC ranged from 0.57($\pm$0.02) to 0.86 ($\pm$0.01) in internal cross-validation. The best

model, GBM, has a Q of 91.04%, a Pr of 93.57%, a Re of 88.89%, a F-score of 93.37%, a Sp of 91.17%, a MCC 0.82 and AUC of 1.0 in external validation demonstrating that the collection of few and more informative descriptors can more accurately distinguish whether pharmaceutical compound can cross the blood-brain barrier. Additionally, the SHAP interpretability framework was employed to enhance model transparency and to elucidate the relative importance of key molecular descriptors influencing prediction results. The SHAP analysis revealed that two primary factors, such as the spatial distribution of atomic charges and atomic electronegativity, play a critical role in determining BBB penetration predictions. Overall, the explainable GBM classification model developed in this study shows strong potential as a predictive and screening tool to identify drug candidates targeting the central nervous system (CNS) or having a better pharmacokinetic profile.

ACKNOWLEDGEMENTS

The authors wish to thank the Laboratory of Fundamental and Applied Physics (LFAP) at Nangui ABROGOUA University, Côte d'Ivoire, for making available the facilities that supported this research.

AUTHOR'S CONTRIBUTION

N'guessan A: performing the data collection, curation, the study, and writing the original draft. **Mélèdje D:** checking and correcting python code**. Akonan L:** supervision methodology and intellectual input. **Kouakou JLK:** literature review, research analysis and data inspection. **Moussé L:** formal analysis and conceptualization. **Kéita M:** formal analysis and conceptualization. **Kré R:** data investigation and intellectual input**. Ziao N:** supervision methodology and review. **Megnassan E:** formal analysis, supervision methodology and review. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY

The empirical data used to support the study's results can be obtained upon request from the corresponding author.

CONFLICT OF INTEREST

The authors declare that no conflict of interest is associated with this study.

REFERENCES

1. Shaker B, Lee J, Lee Y, *et al.* A machine learning-based quantitative model (LogBB_Pred) to predict the blood-brain barrier permeability (logBB value) of drug compounds. Bioinfo (Oxford, England) 2023; 39(10):btad577. *https://doi.org/10.1093/bioinformatics/btad577*
2. Huang ETC, Yang JS, Liao KYK, *et al.* Predicting blood-brain barrier permeability of molecules with a large language model and machine learning. Sci Repo 2024;14(1):15844. *https://doi.org/10.1038/s41598-024-66897-y*
3. Shaker B, Yu MS, Song JS, *et al*. Light BBB: Computational prediction model of blood–brain-barrier penetration based on LightGBM. Bioinfo 2021;37(8):1135–1139. *https://doi.org/10.1093/bioinformatics/btaa918*
4. Faramarzi S, Kim MT, Volpe DA, *et al*. Development of QSAR models to predict blood-brain barrier permeability. Front Pharm 2022;13:1040838. *https://doi.org/10.3389/fphar.2022.1040838*
5. Singh M, Divakaran R, Konda LSK, *et al*. A classification model for blood brain barrier penetration. J Mole Graph Model 2020; 96:107516. PMID: 31940508 *https://doi.org/10.1016/j.jmgm.2019.107516*
6. Mauri A, Bertola M. Alvascience: A new software suite for the qsar workflow applied to the blood–brain barrier permeability. Int J Mol Sci 2022;23(21):12882. *https://doi.org/10.3390/ijms232112882*
7. Yuan Y, Zheng F, Zhan CG. Improved prediction of blood-brain barrier permeability through machine learning with combined use of molecular property-based descriptors and fingerprints. AAPS J 2018;20(3):54. *https://doi.org/10.1208/s12248-018-0215-8*
8. Wang Z, Yang H, Wu Z, *et al*. *In-silico* prediction of blood-brain barrier permeability of compounds by machine learning and resampling methods. Chem Med Chem 2018;13(20):2189-2201. *https://doi.org/10.1002/cmdc.201800533*
9. N'guessan A, Dali B, Esmel EA, *et al*. Pollution risk assessment by designing predictive binary classification models of substituted benzenes centered on data mining and machine learning techniques. Env Sci Pollu Res Int 2025;32(35):21092–21116. *https://doi.org/10.1007/s11356-025-36874-7*
10. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. Clustofvar: An r package for the clustering of variables. J Stat Soft 2012; 50(13):1-16. *https://www.jstatsoft.org/index.php/jss/article/view/v050i13*
11. Singh D, Singh B. Investigating the impact of data normalization on classification performance. App Soft Comp 2020;97:105524. *https://doi.org/10.1016/j.asoc.2019.105524*
12. Pedregosa, Fabian, *et al*. Scikit-learn: Machine learning in python. J Mac Learn Res 2011 ;12: 2825-2830.
13. Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. J Arti Intel Res 2002;16:321-357. *https://doi.org/10.1613/jair.953*
14. Wood DJ, Carlsson L, Eklund M, Norinder U, Stålring J Wood, David J. *et al*. QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. J Computer-Aided Mol Design 2013; 27 (3): 203–219. *http://dx.doi.org/10.1007/s10822-013-9639-5*
15. Kar S, Roy K, Leszczynski J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. In Methods in Molecular Biology 2018;141–169. Springer New York. *https://doi.org/10.1007/978-1-4939-7899-1_6*
16. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20:273-297. *https://doi.org/10.1007/BF00994018*
17. Peterson LE. K-nearest neighbors. Scholar 2009;4(2):1883. *https://doi.org/10.4249/scholarpedia.1883*
18. Patel BR, Rana KK. A survey on decision tree algorithm for classification. Int J Eng Dev Res 2014;2(1):1-5. *https://rjwave.org/IJEDR/papers/IJEDR1401001*
19. Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: The elements of statistical learning. Sprin Seri Statis 2009;337-387. *https://doi.org/10.1007/978-0-387-84858-7_10*
20. Arrieta AB, Díaz-Rodríguez N, Del Ser J, *et al.* Explainable artificial intelligence (XAI): Concepts, taxonomies,

opportunities and challenges toward responsible AI. Inform Fusion 2020;58 :82-115.
https://doi.org/10.1016/j.inffus.2019.12.012

21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neur Inform Process Sys 2017;30.
https://arxiv.org/abs/1705.07874
22. Lundberg SM, Erion G, Chen H *et al*. From local explanations to global understanding with explainable AI for trees. Nat Mach Intel 2020;2(1):56–67.
https://doi.org/10.1038/s42256-019-0138-9
23. Dehnbostel FO, Dixit VA, Preissner R, *et al.* Non-animal models for blood–brain barrier permeability evaluation of drug-like compounds. Sci Repo 2024;14:8908.
https://doi.org/10.1038/s41598-024-59734-9
24. Puzyn T, Leszczyński J, Cronin MTD. Recent advances in QSAR studies: Methods and applications. Springer 2010.
https://doi.org/10.1007/978-1-4020-9783-6
25. Roy K, Mitra I. Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment. Curr Comp Aid Drug Des 2012;8(2):135-158.
https://doi.org/10.2174/157340912800492366
26. Kar S, Roy K, Leszczynski J. Applicability domain: A step toward confident predictions and decidability for qsarmodeling. In: Nicolotti, o. (eds) Comp Toxi Meth Mole Bio 2018;1800.
https://doi.org/10.1007/978-1-4939-7899-1_6
27. De P, Kar S, Ambure P, *et al.* Prediction reliability of QSAR models: An overview of various validation tools. Arch Toxic 2022; 96(5):1279-1295.
https://doi.org/doi:10.1007/s00204-022-03252-y
28. Qadri YA, Shaikh S, Ahmad K, *et al.* Explainable artificial intelligence: A perspective on drug discovery. Pharma 2025; 17(9):1119.
https://doi.org/10.3390/pharmaceutics17091119
29. Ponce-Bobadilla AV, Schmitt V, Maier CS, *et al.* Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clin Transl Sci 2024;17(11):e70056.
https://doi.org/10.1111/cts.70056
30. Pardridge WM. Blood-brain barrier and delivery of protein and gene therapeutics to brain. Front Agi Neuro 2019;11:373.
https://doi.org/10.3389/fnagi.2019.00373
31. Lang PF. Revisiting electronegativity and electronegativity scales. J Chem Edu 2024;102(1):424–29.
https://doi.org/10.1021/acs.jchemed.4c01353
32. Önlü S, ürker Saçan M. Impact of geometry optimization methods on QSAR modelling: A case study for predicting human serum albumin binding affinity. SAR and QSAR in Environmental Research 2017; 28(6): 491–509.
https://doi.org/10.1080/1062936X.2017.1343253