Prediction of Blood Brain Barrier Permeability

Kushagra Gupta
Roll no. 21075049
Computer Science and Engineering (B.Tech)
Indian Institute of Technology
Varanasi, Uttar Pradesh
kushagra.gupta.cse21@itbhu.ac.in

Sachin Kumar Gupta
Roll no. 21075074
Computer Science and Engineering (B.Tech)
Indian Institute of Technology
Varanasi, Uttar Pradesh
sachinkumar.gupta.cse21@itbhu.ac.in

Harmanjot Singh
Roll no. 21075037
Computer Science and Engineering (B.Tech)
Indian Institute of Technology
Varanasi, Uttar Pradesh
harmanjot.singh.cse21@itbhu.ac.in

Ruchira Naskar
Roll no. 21075072
Computer Science and Engineering (B.Tech)
Indian Institute of Technology
Varanasi, Uttar Pradesh
ruchira.naskar.cse21@itbhu.ac.in

***Abstract*—The Blood Brain Barrier (BBB) is a selectively permeable interface that regulates the transport of substances between the blood circulation and the central nervous system (CNS). Understanding the permeability of the BBB holds very high importance in various research and clinical domains. As the BBB presents a formidable obstacle for drug molecules seeking to access the central nervous system, accurately predicting whether a compound can cross the barrier and penetrate the brain is vital for optimizing drug candidates and improving therapeutic outcomes. Machine learning and deep learning models have emerged as powerful tools for assessing BBB permeability. Various molecular descriptors can be utilized to develop predictive models which are trained on a dataset of compounds with known BBB penetration status, to determine if other compounds can penetrate the barrier. In our proceeding, Principal Component Analysis (PCA), Support Vector Machine (SVM), Keras Neural Network, and Extreme Gradient Boosting (XGBoost) techniques have been applied to classify compounds as penetrant or non-penetrant.**

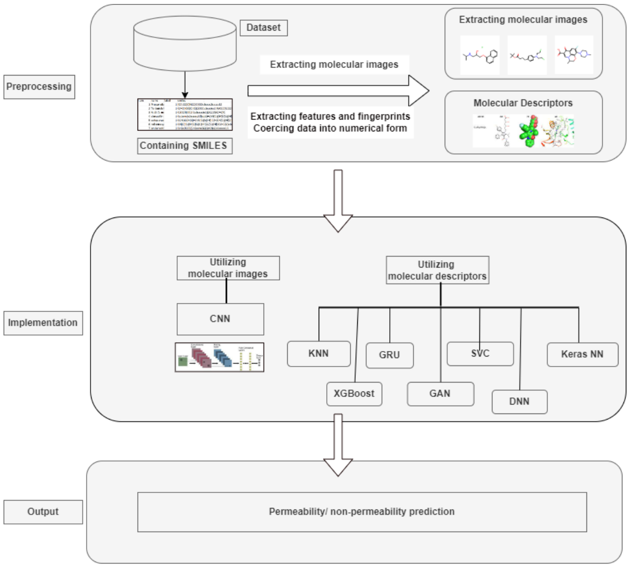
Preprocessing
Dataset
Containing SMILES
Extracting molecular images
Extracting features and fingerprints
Coercing data into numerical form
Extracting molecular images
Molecular Descriptors
Implementation
Utilizing molecular images
CNN
Utilizing molecular descriptors
KNN
GRU
SVC
Keras NN
XGBoost
GAN
DNN
Output
Permeability/ non-permeability prediction

Fig. 1. Block Diagram of Workflow

I. Introduction

Delivery of oxygen and nutrients to tissues and organs in the body depends on blood vessels. The blood vessels that vascularize the central nervous system (CNS) have unique properties, termed the blood-brain barrier, which allows these vessels to strictly regulate the movement of ions, molecules, and cells between the blood and the brain. Precise control of CNS homeostasis paves the way to proper neuronal function and protection of the neural tissue from toxins and pathogens, and alterations of the barrier properties are vital constituents of the pathology and progression of different neurological diseases [1]. The primary function of the blood-brain barrier is to regulate the transport of substances between the blood and the brain, to facilitate a stable and optimal environment for the brain to function properly. It acts as a protective barrier, preventing the entry of potentially harmful substances into the brain tissue. The tight junctions between the endothelial cells of the Barrier prevent leakage of substances between cells by restricting paracellular transport. The blood-brain barrier permeability (BBBP) refers to the ability of substances to cross the BBB and enter the brain tissue. The blood-brain permeability plays a key role in the pharmacokinetics of drugs and other substances that are intended to target the central nervous system. Thus, the efficacy of many medications is affected in the course of the treatment of various neurological disorders. The BBB also acts as a regulatory interface by maintaining tight control over the chemical composition of the extracellular fluid of the brain by actively transporting specific molecules across the barrier using specialized transporters and efflux pumps. This ensures a proper balance of ions, neurotransmitters, and other molecules which is necessary for optimal neuronal function.

BBB failure can both cause and lead to CNS inflammation, which affects the severity and course of many mental and neurological illnesses. An inflammatory response in the brain may result in endothelial cell destruction and increased BBB permeability. Numerous CNS illnesses, such as brain damage, ischemic stroke, multiple sclerosis (MS), epilepsy, Parkinson's disease, Alzheimer's disease, and significant depression, affect the permeability of the BBB. BBB impairment and various psychiatric conditions have been linked in studies [2].

Several methods have been developed to measure and predict BBB permeability, including in vitro cell culture, animal, and computational models. The in vitro models like cell monolayers and microfluidic devices are able to simulate the blood-brain barrier's physiological conditions and assess the permeability of test compounds. Animal models, such as rodents or non-human primates, can provide valuable insights into BBB function and permeability. Computational models with machine learning and molecular modeling techniques can predict barrier permeability based on molecular descriptors and physicochemical properties of compounds. The determination of the blood-brain barrier (BBB) permeability of compounds holds significant importance in various scientific and clinical contexts. One such major role is in drug development. Determining the permeability of the BBB helps identify compounds that can effectively penetrate the CNS and reach their target sites. Many therapeutic agents need to reach the brain to exert their intended effects. Drug delivery to the brain can be improved by overcoming the barrier, leading to better results from the treatment of various neurological disorders. Altered BBB permeability is associated with several conditions, for instance, neurodegenerative diseases, brain tumors, and inflammation. Getting insights into the changes in BBB permeability in these diseases provides knowledge of disease mechanisms and potential therapeutic targets. Neurotoxicity and cognitive impairments can be caused by certain environmental toxins and chemicals that can breach the BBB. The ability to predict BBB permeability using computational models and in vitro methods is valuable in the early stages of drug discovery. This capability assists researchers in utilizing their time and energy efficiently in focusing their efforts on compounds with a higher probability of successfully crossing the BBB, streamlining the drug discovery pipeline.

A. Motivation

- The blood-brain barrier is a very intriguing part of the anatomy and we have tried to implement predictive models based on machine learning and deep learning algorithms that can identify compounds with a higher likelihood of BBB penetration, resulting in increased efficiency of screening processes and reducing costs associated with the further development of non-penetrant compounds.
- Assessing BBB permeability will help in identifying potential neurotoxic agents and evaluating their impact on brain health.
- To enrich the medical field with information and developments related to blood-brain barrier activity.

B. Challenges

- It was challenging to obtain comprehensive data on blood-brain barrier permeability due to the complex nature of the barrier and restricted access to relevant experimental data. The limited dataset posed difficulties in training accurate and reliable models.
- Handling missing data added complexity to the feature engineering process.
- Selecting the optimal model architecture, optimization techniques, and hyperparameters required careful experimentation.
- The risk of overfitting needed to be carefully mitigated. Applying appropriate regularization techniques and ensuring the models' ability to handle diverse data distributions were necessary strategies implemented to improve generalization performance.
- Our project had to face challenges due to limited computational power and extended model training durations when working with complex model architectures.

C. Contributions

- We employed proper preprocessing techniques, such as standardization and handling missing values to ensure quality and integrity.
- We performed advanced techniques like molecular fingerprint algorithms were employed to extract informative features from the SMILES dataset to improve the predictive power of the models.
- We conducted rigorous experimentation and fine-tuning of hyperparameters to optimize the predictive performance of the models.
- We utilized deep learning and machine learning algorithms to improve prediction accuracy and discover intricate molecular patterns.

II. Related Work

In 2018, Wang et al. proposed a model based on insilico prediction of blood-brain permeability of compounds by machine learning and resampling methods. BBB permeability has recently emerged as a crucial problem in chemical ADMET (chemical absorption, distribution, metabolism, excretion, and toxicity) prediction; yet, practically all models were created using unbalanced data sets, leading to a high false-positive rate. So, in an effort to address the issue of biased data sets, constructed a trustworthy classification model was built using 2358 compounds. Models containing both 2D molecular descriptors and molecular fingerprints to represent the chemicals were refined simultaneously using machine learning and resampling techniques. It was concluded through a series of evaluations that resampling techniques like Synthetic Minority Oversampling Technique (SMOTE) and SMOTE+edited Nearest Neighbour could successfully address the issue of imbalanced datasets and that the best performance came from combining the MACCS fingerprint with support vector machine. The overall accuracy rate for the final external data set was raised to 0.966 following the creation of a consensus model. The model's accuracy rate for the test set was 0.919, and it had an extremely balanced capacity of 0.925 (sensitivity) to predict blood-brain barrier-positive compounds and of 0.899 (specificity) to predict blood-brain barrier-negative compounds. The implemented models reduced the false positive rate and performed well in predicting both BBB-positive and BBB-negative substances, which might be highly beneficial in the early stages of drug discovery.

In 2018, Yuan et al. proposed to develop a generally applicable SVM model by combining all the features of the molecular property-based descriptors and fingerprints to improve the predictive power of blood-brain barrier permeability. It was shown that the collaborative use of property-based descriptors and fingerprints can result in remarkable improvement of the SVM models compared to the corresponding SVM models implemented using the property-based descriptors and fingerprints individually.

In 2021, Alsenan et al. proposed a deep-learning model to predict the blood-brain barrier permeability of compounds. A deep learning-based classification model is offered that forecasts blood-brain barrier permeability. High dimensionality, class imbalances, and low specificity scores, which were the main problems with earlier models, are all addressed in the proposed approach. The imbalanced dataset is addressed using oversampling techniques, and the high dimensionality is addressed using a non-linear dimensionality reduction technique called kernel principal component analysis. With this method, valuable information is preserved when the high-dimensional dataset is converted to a low-dimensional Euclidean space. A convolutional neural network model and an improved feed-forward deep learning model were built for the classification job. The upgraded feed-forward deep learning model outperformed other models in the literature that were created using the same technique in terms of specificity scores. The issue of low specificity and a large number of false positives was resolved by the suggested strategy. An overall accuracy of 97.0% was achieved on external dataset by the CNN model and that of 96.5% for the FFDNN model.

In 2021, Liu et al. proposed in-silico ensemble learning models to predict the BBB permeability of compounds. To predict BBB permeability, in silico ensemble-learning models were created utilizing three machine-learning algorithms and nine molecular fingerprints from 1757 compounds (combined from two published datasets). A prediction model based on a random forest (RF) and a MACCS molecular fingerprint, with an ACC of 0.910, an area under the receiver-operating characteristic (ROC) curve (AUC) of 0.957, a SEN of 0.927, and a specificity of 0.867 in 5-fold cross-validation, had the best prediction performance of the base classifier models. The ensemble models' prediction performance is superior to that of the majority of the base classifiers. The final ensemble model showed impressive accuracy for external validation.

In 2022, Tong et al. studied ways in which uncertainty estimation methods develop in-silico blood-brain barrier permeability models. To assess the accuracy of the most recent algorithms, a brief review of the state-of-the-art in silico BBBp prediction and uncertainty estimation approaches of deep learning models was conducted. The findings revealed that, despite graph neural network-based deep learning models and traditional physicochemical-based machine learning models having equivalent performance on BBBp prediction, the GROVER-BBBp model performs significantly better when employing uncertainty estimations. Particularly, the strategy of combining Entropy and MC-dropout can increment the accuracy of distinguishing BBB positive from BBB negative to more than 99% by extracting predictions with a high confidence level (uncertainty score less than 0.1).

In 2022, Tang et al. proposed a merged molecular representation deep learning method for blood-brain barrier permeability prediction. A deep learning-based multi-model framework model was introduced, known as Deep-B^3, to predict the BBB permeability of potential drugs in order to complement the inadequacies of previous approaches. The samples in Deep-B^3, are encoded using three different types of features: molecular descriptors and fingerprints, molecular graphs, and text notation using the SMILES molecular input line entry system. To extract latent features from the molecular graph and SMILES, pre-trained models were created. These features provided tabular data, image, and text representations of the compounds. The performance of Deep-B^3, outperformed the most recent models, according to the validation findings obtained from an independent dataset.

In 2022, Parakkal et al. developed high-accuracy blood-brain barrier permeability prediction with a mixed deep-learning model. Here, Mol2vec was coupled with a mixed DL-based model made up of Multi-layer Perceptron (MLP) and Convolutional Neural Network layers. High-dimensional vector representations of molecules and their molecular substructures are created using the convenient and unsupervised machine learning technique known as Mol2vec. These brief vector representations serve as inputs for the mixed DL model, which generates predictions for BBBP. The mixed DL model has used a number of well-known benchmarks that include BBBP data for supervised training and prediction, achieving better results than other ML and DL methods currently in use for predicting BBBP.

III. Methodology

To determine the blood-brain barrier permeability of compounds, a series of processes were carried out to bring out the best results in our proceeding.

A. Dataset

The modelling and prediction of barrier permeability are done using the blood-brain barrier penetration (BBBP) dataset. In this dataset, more than 2000 chemicals with permeability qualities have binary labels. The raw data contains columns, namely, the ”name” column consists of the name of the compound, ”smiles” denotes the SMILES representation of the molecular structure, and the ”p_np” column has the binary labels associated with penetration or non-penetration of the compounds.

SMILES Representation: SMILES (simplified molecular-input line-entry system) is a specification in the form of a line notation to describe the structure of chemical species using short ASCII strings. SMILES can be converted to molecular structure by using the RDKIT module of Python. Atoms are frequently denoted by their atomic symbols, and they can also be further described by other characteristics. Different symbols, such as ”-” for a single bond, ”=” for a double bond, and ”#” for a triple bond, are used to represent bonds between atoms. Specific guidelines and syntax are used in the SMILES format. Beginning with the atom on the left and moving to the right, the atoms and bonds are listed sequentially, with brackets used to indicate branches.

B. Feature Extraction

Feature extraction is a process in which relevant and informative features are extracted from raw data to represent and capture important characteristics or patterns in the data. The goal is to transform the raw data into a lower-dimensional representation that retains the most relevant information for the given task. This reduction in dimensionality can simplify subsequent analysis and improve computational efficiency. Moreover, feature extraction aims to highlight distinctive patterns or properties in the data that are relevant to the specific problem at hand. We have extracted features for SMILES and stored them in an output file.

1) Molecular Descriptors: A well-defined algorithm applied to a defined molecular representation or a well-defined experimental process yields molecular descriptors, which are formally mathematical representations of a molecule [3]. Mordred is a descriptor-calculation software application that is capable of calculating more than 1800 2D and 3D descriptors [4]. All data types are coerced to numeric data type.

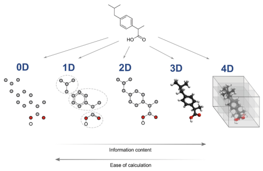

Fig. 2. Descriptors

2) Molecular Images: Molecular images play a vital role in detecting blood-brain barrier (BBB) permeability when combined with algorithms like CNNs. By providing detailed visual representations of molecular structures, molecular images offer valuable information of the factors that influence BBB permeability. These images are used to extract meaningful features and detect intricate patterns that impact permeability. The RDKit has built-in tools for creating images from molecules found in the rdkit.Chem.Draw package [8].

3) Scaling and Principal Component Analysis: Feature scaling is a technique for normalizing the variety of independent variables or features in data is called feature scaling. It is typically carried out during the data preprocessing step and is sometimes referred to as data normalization in the context of data processing. With a high number of dimensions or features per observation, principal component analysis (PCA) is a common method for analyzing huge datasets, improving data interpretation while retaining the most information, and enabling the visualization of multidimensional data. Standardization is a common preprocessing step in machine learning that aims to transform the data in a way that the features have zero mean and unit variance. It is often performed to bring different features onto a similar scale. The Find_Optimal_Cutoff function is used to determine the threshold (cutoff) probability that yields the optimal balance between the true positive rate and false positive rate, as measured by the ROC curve. The Find_Optimal_threshold calculates the threshold that maximizes the F1 score for each column of the predicted data, providing a threshold selection strategy based on F1 score optimization.

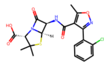

Fig. 3. Example of generated image

C. Algorithms

1) K-Nearest Neighbours: K-Nearest Neighbours is a non-parametric algorithm that does not make any assumptions about the underlying data distribution but relies on the local characteristics of the data to make predictions. The algorithm works based on the principle that similar instances tend to have similar outcomes. Given a new data point, KNN identifies its K nearest neighbors in the training set based on a distance metric. The prediction for the new point is then determined by the majority vote of the labels of its K neighbors in classification or by averaging the labels in regression.

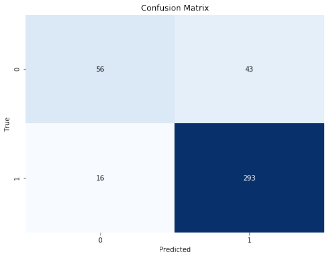

Fig. 4. Confusion Matrix of KNN

2) XGBoost: XGBoost, or Extreme Gradient Boosting, is a popular machine learning algorithm that leverages the gradient boosting framework to iteratively train a series of weak prediction models, gradually improving the overall model's performance. With its regularization techniques, customizable objective functions, and tree-pruning capabilities, XGBoost effectively prevents overfitting and enhances model generalization. It is capable of handling missing values and outliers, supports parallel processing for scalability, and provides feature importance analysis.

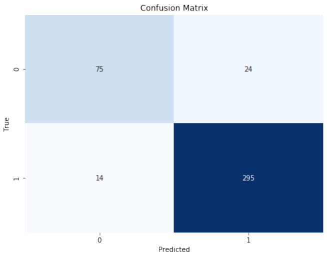

Fig. 5. Confusion Matrix of XGBoost model

3) Deep Neural Networks (DNN): A class of artificial neural networks called deep neural networks, commonly referred to as deep learning models, are able to learn and represent complex patterns and correlations in data. They are distinguished by their depth, which refers to the presence of numerous layers of interconnected nodes or neurons. Through the use of backpropagation and forward propagation, deep neural networks are built to automatically learn hierarchical data representations. Each layer in the network takes input from the previous layer, runs a number of calculations, and then transmits the results to the next layer. With the help of this hierarchical architecture, the network may gradually extract higher-level, more abstract properties from the input data.

4) Gated Recurrent Unit: Gated Recurrent Units (GRUs) are a type of recurrent neural network (RNN) architecture that addresses some of the limitations of traditional RNNs, such as the vanishing gradient problem, Gated Recurrent Units have become popular in various natural language processing tasks. GRUs are designed to capture long-term dependencies in sequential data by utilizing gating mechanisms which enable the model to selectively update and reset its internal state, allowing it to retain relevant information over long sequences.

5) General Adversarial Networks: Deep representations can be learned using generative adversarial networks (GANs) without the need for extensively annotated training data [4]. The basic idea behind GANs is to train two neural networks, called the generator and the discriminator, in a competitive setting. The generator network takes random noise as input and aims to generate synthetic data samples that resemble the real data distribution. The discriminator network acts as a binary classifier, distinguishing between real and fake samples.

6) SVC Model: The Sklearn SVC model refers to the Support Vector Classifier model provided by the scikit-learn library in Python. SVC is a supervised machine learning algorithm used for classification tasks. It is based on the concept of Support Vector Machines (SVM), which is a powerful algorithm for both classification and regression. SVC aims to find an optimal hyperplane that separates the different classes in the feature space in classification.

7) Keras Neural Network Model: A range of robust new methods for pattern recognition, data analysis, and control are made available by neural networks. High processing speeds and the capacity to learn a problem's solution from a series of instances are just two of their standout qualities [7]. Keras is a neural network Application Programming Interface (API) for Python that is tightly integrated with TensorFlow, which is used to construct machine learning models. The models offered by Keras often define a neural network in a simple manner, which will then be built for you by TensorFlow.

8) Gradient Boosting: Gradient Boosting, an ensemble method, is a popular machine learning technique used for building predictive models. It combines multiple weak predictive models (often decision trees) to create a strong predictive model. We have performed gradient boosting on Keras Model with XGBoost and with SVC.

9) Convolutional Neural Networks: CNNs are primarily used to solve complicated image-driven pattern recognition tasks with the help of a precise yet simple architecture [7]. CNNs automatically extract meaningful features from raw data. The key components of CNNs are convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply filters to input data, pooling layers downsample the output, and fully connected layers predict outcomes based on learned features.

10) DNN with Hybridisation: Hybrid features extracted from SMILES strings refer to a representation of the chemical structure encoded in the SMILES notation. Hybrid features are extracted to capture the hybridization state of each atom in a molecule. A deep learning model is trained and evaluated for binary classification based on hybrid features extracted from SMILES strings, as well as providing visualizations of the training process and performance metrics.

IV. Results

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

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

Misclassification = $\frac{FP+FN}{TP+TN+FP+FN}$

Precision = $\frac{TP}{TP+FP}$

Sensitivity (also known as Recall) = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{TN+FP}$

F1 Score = $2 \times \left(\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\right)$

AUC (Area Under the Curve) represents the area under the receiver operating characteristic curve (ROC curve). AUC score is a performance metric used in machine learning to evaluate the quality of a binary classification model.
We have obtained the following results from different models implemented on our BBBp dataset:-

Algorithm	*Accuracy*	*F1 Score*	*ROC AUC Score*	*Misclassification*	*Precision*	*Recall*	*Specificity*
KNN	0.855	0.908	0.757	0.145	0.872	0.948	0.566
XGBoost	0.909	0.941	0.858	0.090	0.925	0.958	0.757
DNN	0.885	0.925	0.828	0.115	0.912	0.935	0.717
GRU	0.757	0.862	0.5	0.242	0.757	1.0	0.0
GAN	0.702	0.821	0.533	0.298	0.696	1.00	0.067
SVC	0.863	0.915	0.744	0.137	0.868	0.968	0.521
Keras Neural network	0.857	0.912	0.733	0.142	0.863	0.968	0.5
Gradient Boosting of Keras NN and SVC	0.872	0.920	0.772	0.127	0.882	0.961	0.583
Gradient Boosting of Keras NN and XGBoost	0.868	0.915	0.791	0.132	0.896	0.936	0.646
CNN	0.998	0.999	0.999	0.002	0.999	0.998	0.997
DNN with Hybridization	0.816	0.794	0.791	0.183	0.805	0.816	0.816

TABLE I
RESULTS

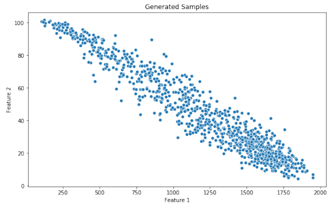

Fig. 6. Generated Samples, GAN

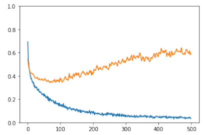

Fig. 7. Plot of Train loss and Valid loss with epochs, Keras Neural Network

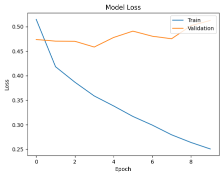

Fig. 8. Plot of DNN with hybridization model loss vs. epoch

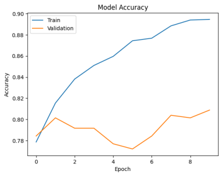

Fig. 9. Plot of DNN with hybridization model accuracy vs. epoch

V. CONCLUSION

The Convolutional Neural Network model to predict the permeability labels based on the images achieved the highest accuracy of 0.97008 among all models implemented. The ability of CNN to capture local patterns, learn hierarchical representations, and leverage weight sharing makes them more effective in tasks involving visual data proves to be more accurate in predicting labels than other models.

Future Directions

The prediction of the blood-brain barrier (BBB) permeability of substances using machine learning and deep learning techniques holds high significance for the future. These approaches can revolutionize drug discovery and therapeutic development by enabling more accurate and efficient identification of compounds that can effectively cross the barrier. With advancements in computational models and resources, the integration of diverse molecular descriptors, and the availability of large-scale datasets, machine learning, and deep learning algorithms can capture complex relationships and patterns that influence BBB permeability. These advancements will expedite the identification of potential therapeutics for neurological disorders and also facilitate the design of targeted drug delivery systems that can effectively cross the BBB, leading to more effective treatments for brain-related conditions, being of immense help in the medical field.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Dr. S.K. Singh, Head of the Department, of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, for providing us with the splendid opportunity to explore and innovate, with all the necessary information and knowledge required for the project. We express our heartfelt gratitude to Ma'am Sakshi Ranjan for her support and guidance throughout the project.

REFERENCES

[1] R. Daneman and A. Prat, ”The blood-brain barrier,” Cold Spring Harbor perspectives in biology, vol. 7, no. 1, p. a020412, 2015.
[2] M. A. Małkiewicz et al., ”Blood-brain barrier permeability and physical exercise,” Journal of neuroinflammation, vol. 16, pp. 1-16, 2019.
[3] V. Consonni and R. Todeschini, ”Molecular descriptors,” in Recent advances in QSAR studies: methods and applications, 2010, pp. 29-102.
[4] H. Moriwaki et al., ”Mordred: a molecular descriptor calculator,” Journal of cheminformatics, vol. 10, no. 1, pp. 1-14, 2018.
[5] C. M. Bishop, ”Neural networks and their applications,” Review of scientific instruments, vol. 65, no. 6, pp. 1803-1832, 1994.
[6] A. Creswell, et al., ”Generative adversarial networks: An overview,” IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018.
[7] K. O'Shea and R. Nash, ”An introduction to convolutional neural networks,” in arXiv preprint arXiv:1511.08458, 2015.
[8] G. Landrum, ”Rdkit documentation,” Release 1.1-79, 2013, pp. 4.