Brain disorders exhibit high prevalence and mortality rates. Globally, brain disorders contribute to 460 million disability-adjusted life years and 17 million deaths annually[1,2]. Despite the protective role of the blood-brain barrier (BBB) in maintaining brain integrity, it presents a formidable obstacle for drug delivery in treating brain disorders. Less than 2% of FDA-approved small molecules can penetrate the BBB[3], and active efflux mechanisms eliminate many that do[4], thereby limiting the efficacy of diagnostics and therapeutics. Current strategies to bypass BBB limitations, such as higher systemic administration or direct brain implantation, are associated with drawbacks such as severe side effects and risks of complications and infections. The prevalence of inadequate drug delivery has stimulated a dynamic market valued at $1.6 billion, which is projected to reach $5 billion within 5 years through a 26% annual growth rate[5]. Consequently, **successful prediction for a chemical's ability to penetrate the BBB has the potential to revolutionize drug development, inspire innovative delivery strategies, and improve patient outcomes**.

In this project, our objective is to utilize chemical structures to predict the BBB penetration capacity of a compound. The original dataset was sourced from a 2021 Nature publication[6], where the authors curated two datasets by pooling 50 literature sources. One dataset is for regression prediction, containing 1058 chemicals with numerical BBB data, while the other is for classification predication, encompassing 7807 chemicals with categorical BBB data. Among these, 4956 chemicals exhibit BBB penetration capability, while 2851 chemicals do not. Our project started with dataset preprocessing. We expanded the descriptors for each chemical using RDKit descriptors, Morgan fingerprints, and MACCS keys, then subsequently eliminating features with zero variance and setting aside 10% of the data as a holdout dataset.

To address the regression problem, we initially centered and standardized the data, followed by an 80/20 train-test split. Using the training set, we conducted an 8-fold cross-validation with GridSearch to train both a baseline support vector machine (SVM) regressor (SVR) and a more complex random forest (RF) regressor using the best parameters. Scoring was based on mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2). Model performances were compared through residual analysis and a Student's t-test on the MSE. Despite insignificant differences observed in the residual analysis between the two models, the performance on the test set indicated that the RF regressor outperformed the SVR in terms of mitigating overfitting. Additionally, the RF regressor identified important features in accessing the BBB, primarily those related to polarity and charge. The top-ranked feature was identified to be topological polar surface area (TPSA), which generally influences a chemical's ability to pass through membranes. Notably, this feature is also utilized by the commercial tool SwissADME to predict BBB penetration. Consequently, the RF regressor was selected as the better regressor.

For our classification model, following the centering and standardization of the dataset we addressed the heavily unbalanced nature of the original dataset. This was done by testing two methods: the ClusterCentroids method and the synthetic minority over-sampling technique with edited nearest neighbor (SMOTEENN). The former downsamples the larger category, while the latter upsamples the smaller category with subsequent cleaning. Each balanced dataset underwent PCA for dimensionality reduction until 95% of variance was retained. Subsequently, the datasets underwent an 80/20 stratified train-test split and were then fed into two SVM classifiers (SVC) with 8-fold cross-validation. We employed GridSearch to score different metrics including recall, precision, F1 score, accuracy, balanced accuracy, and area under the receiver operating characteristic curve (AUROC). AUROC was used for ranking the best SVC. After obtaining the ideal parameters for each trained model, we compared the performance of the two best SVCs. This was done with the holdout dataset in order to apply DeLong's test. This test revealed that balancing via ClusterCentroids yielded more reliable results. Therefore, the dataset balanced by ClusterCentroids was also employed to train a more complex model: the multilayer perceptron (MLP) classifier. Following the same train-test split and scoring metrics, the best MLP classifier was obtained. Both the baseline SVC and complex MLP classifiers were compared according to their performance on the test set to prevent overfitting. Since DeLong's test resulted in insignificance, we chose the SVC as the preferred model due to its simplicity and ease of training.

We finally assessed our best models, the RF regressor and SVC, on the holdout dataset. Their performance was comparable to that observed on the train and test sets, confirming our original findings. Additionally, we selected three new chemicals with known permeabilities for the BBB. Our SVC correctly predicted 2/3 of them when run through the model using the same feature set.

In conclusion, we trained an RF regressor that adequately captures the dataset structure in the regression dataset and demonstrates strong generalization. However, its performance could be further improved with additional regression data. On the classification side, we experimented with two balancing methods and opted for the more reliable ClusterCentroids balancing approach. The resulting SVC exhibited robust performance and generalization. Our validation checks, including consideration of orthogonal domain knowledge regarding key features related to polarity and charges, validation via the holdout dataset, and mostly successful cherry-picking, all suggest the efficacy of our models. Future studies will focus on determining causation relationships, acquiring more data to enhance model performance, and conducting in vitro and in vivo validations.

References

1 Álvaro Nuno, P., Suarez, J. & Michel Le, B. Sizing the brain. *Deloitte Insights* (2023).

2 Collaborators, G. B. D. N. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* **18**, 459-480 (2019). https://doi.org/10.1016/S1474-4422(18)30499-X

3 Mikitsh, J. L. & Chacko, A. M. Pathways for small molecule delivery to the central nervous system across the blood-brain barrier. *Perspect Medicin Chem* **6**, 11-24 (2014). https://doi.org/10.4137/PMC.S13384

4 Kusuhara, H. & Sugiyama, Y. Active efflux across the blood-brain barrier: role of the solute carrier family. *NeuroRx* **2**, 73-85 (2005). https://doi.org/10.1602/neurorx.2.1.73

5 The Insight, P. Blood Brain Barrier Technologies Market Global Analysis by 2028. *The Insight Partners* (2021).

6 Meng, F., Xi, Y., Huang, J. & Ayers, P. W. A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Sci Data* **8**, 289 (2021). https://doi.org/10.1038/s41597-021-01069-5