Overcoming Class Imbalance in BBB Permeability Prediction Using Machine Learning and Explainable AI Techniques

Manuel Acquistapace[1], **Stefano Billeter**[2]
[1,2]SUPSI, Viganello , Switzerland
manuel.acquistapace@student.supsi.ch[1]
stefano.billeter@student.supsi.ch[2]

ABSTRACT

The classification of compounds based on their ability to penetrate the blood-brain barrier (BBB) is a critical task in neuropharmacology, influencing the development of effective central nervous system (CNS) therapeutics. This study employs machine learning techniques to predict BBB penetration, addressing the challenge of efficient CNS drug delivery. We developed a predictive model using a dataset composed of chemical properties and BBB permeability data, utilizing algorithms such as random forests and neural networks. Our model demonstrated robust performance, achieving a high accuracy rate, which suggests its potential utility in early-stage drug discovery. The implications of these findings are significant, providing a computationally efficient tool to predict BBB penetration and enhance CNS drug development. This study builds on the paper [1], which also explored uncertainty estimation parameters influencing BBB permeability, by integrating these parameters into a predictive machine learning framework.

Keywords: Blood-brain barrier penetration, BBBp prediction through, classification ML

Link to our GitHub repo is here.

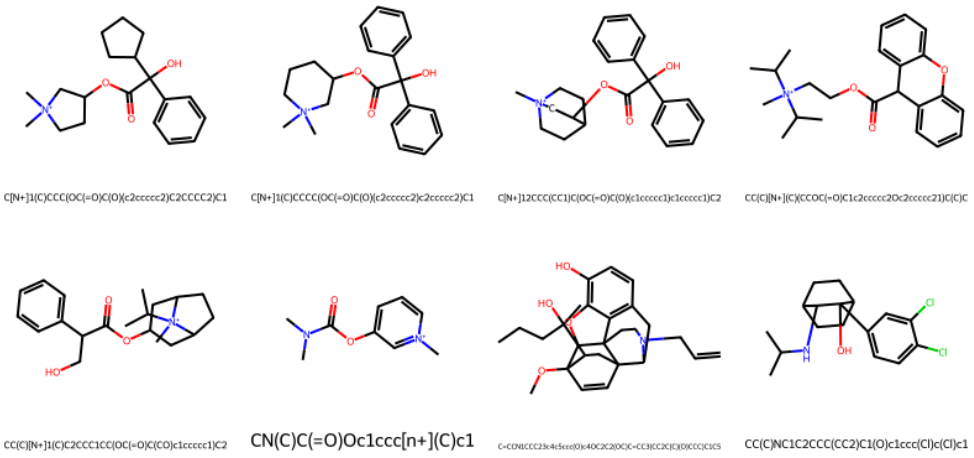

Fig 1. Eight molecules present in our dataset

I. INTRODUCTION

The treatment of central nervous system (CNS) disorders is increasingly critical in an aging global population, where CNS diseases are now only surpassed by cardiovascular conditions in prevalence. Despite the pressing need, the development of CNS-targeted drugs remains fraught with challenges, evidenced by a notably low clinical trial success rate of about 8%—significantly less than that for cardiovascular medications, which stands at 20% [2]. Moreover, the fight against cancer, with a clinical success rate for new drugs around 5% [3], is further complicated by the prevalence of brain metastases in approximately 20% of cancer patients, who often face dismal prognoses despite receiving whole-brain

radiation therapy. Treatments capable of crossing the blood-brain barrier (BBB) such as second-generation kinase inhibitors, are crucial for improving outcomes in such cases [4, 5].

The BBB itself is a protective structure formed by the endothelial cells of brain capillaries. It effectively controls the exchange of molecules between the blood and the CNS while protecting the brain from toxic substances and maintaining neurological homeostasis [6]. However, the BBB also significantly impedes drug delivery to the brain; it is estimated that 98% of small molecules do not penetrate the BBB [7]. The permeability of the BBB is influenced by various mechanisms, including passive diffusion for small lipophilic compounds and specialized transport processes for hydrophilic and larger molecules. The design of effective CNS drugs requires not only high activity and low toxicity but also optimized physicochemical properties to ensure sufficient brain exposure [8].

II. METHODOLOGY

For this study, we utilized two datasets as described in the reference paper [1]. The datasets were distinctly split into one for training and validation, and another dedicated for testing. Each dataset comprises SMILES (Simplified Molecular Input Line Entry System) representations of compounds alongside their respective classes for prediction purposes. It is noteworthy that the class distribution within these datasets is imbalanced, a characteristic that will be visually represented in the subsequent plot. The distribution appeared to be the same for both the datasets.

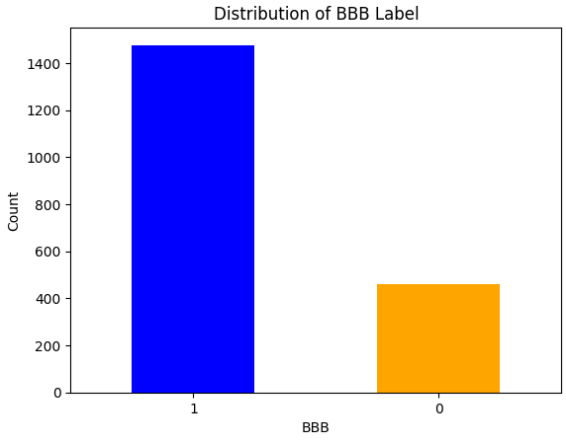

Fig 2. Class imbalance in training dataset

Prior to feature extraction, an initial preprocessing step was carried out to canonicalize the SMILES strings, ensuring uniformity across the datasets. Additionally, any duplicate entries were identified and removed to maintain the integrity and uniqueness of the dataset.

Following preprocessing, we employed the RDKit library along with other analytical tools to generate a comprehensive set of over 5,000 features per compound. These features include, but are not limited to, standard chemical descriptors such as LogP and molecular weight. We also utilized several widely recognized chemical fingerprints, including Morgan and PubChem fingerprints, which provided a diverse and rich feature set, thereby facilitating a more robust machine learning analysis.

After the generation of features, we addressed the class imbalance inherent in the datasets. Utilizing the imbalanced-learn library, we applied random under-sampling and over-sampling techniques. This process resulted in three distinct datasets for comparative analysis: Original Dataset (with the initial class imbalance), Oversampled Dataset (the number of instances in the underrepresented class was increased to match that of the more prevalent class), Undersampled Dataset this dataset involved reducing the size of the predominant class by randomly removing samples, aligning its size with that of the minority class).

In the course of our study, we explored various techniques to address class imbalance, including the implementation of Synthetic Minority Over-sampling Technique (SMOTE), as explained in [9] and Adaptive Synthetic Sampling (ADASYN) as per the paper [10]. However, our application of these techniques resulted in substantial challenges, notably the extensive presence of NaNs in the synthetic data generated, which made us discard these options.

These steps were essential for evaluating how variations in class distribution affect the performance of our predictive models.

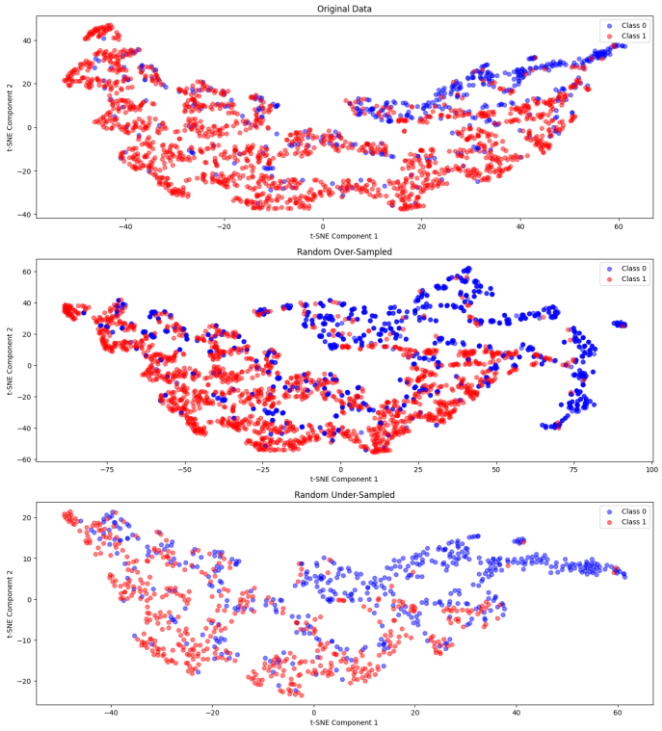

Fig 3. t-SNE representations after tackling class imbalance

III. Models - baseline

To rigorously assess the predictive capabilities of various machine learning models for BBB penetration classification, we employed a stratified K-fold cross-validation approach with five folds. This method was applied to three differently prepared datasets: original, oversampled, and undersampled, allowing each model to be trained and validated across these varied data environments to ensure robustness and generalizability.

For final testing, the models were evaluated on a separate test dataset, which was augmented with generated features but maintained the original class imbalance to realistically simulate prediction scenarios. The test dataset was split using a scaffold splitting technique, as seen in [11], ensuring that molecules with similar structures were grouped together. This provides a stringent test of the model's ability to generalize to new, chemically related compounds.

Prior to final evaluation, all models underwent hyperparameter tuning to optimize performance. The tuning process was tailored for each model to effectively explore the parameter space and enhance predictive accuracy.

We incorporated a diverse range of machine learning models in our study. Logistic Regression was used for its straightforward probabilistic framework suitable for binary classification. Decision Tree and Random Forest were selected for their robustness in handling feature-rich data, with the latter providing an ensemble approach to improve prediction stability. Gradient Boosting was implemented to sequentially build models, each correcting errors from the previous, enhancing the overall prediction accuracy. k-Nearest Neighbors was chosen for its effectiveness in classifying new cases based on similarity measures, and Naive Bayes was utilized for its efficiency in probabilistic classification under strong feature independence assumptions.

This diverse array of models enabled a comprehensive analysis of the datasets, providing insights into which methods are most effective in handling the complexities associated with BBB penetration classification.

Subsequently, we focused on further leveraging our results by training a neural network specifically on the oversampled dataset, which demonstrated superior outcomes in preliminary analyses. The success of this neural network model not only enhances our predictive accuracy but also paves the way for deeper insights through explainable AI (XAI) techniques. By implementing tools such as counterfactual explanations and LIME (Local Interpretable Model-agnostic Explanations), both facilitated by the use of the exmol library, we provide a more transparent and understandable model. This integration of XAI not only enriches our understanding of model decisions but also empowers users by allowing them to input a SMILES string and receive not just a prediction but also intuitive, detailed explanations of the model's reasoning process. This approach aims to bridge the gap between advanced machine learning techniques and practical, user-friendly applications, enhancing the utility and accessibility of our predictive models.

IV. Results

	Imbalanced		Undersampled		Oversampled	
	Acc.	AUC	Acc.	AUC	Acc.	AUC
LogReg	0.8686	0.91	0.7962	0.90	0.8400	0.90
Dec.Tr.	0.7314	0.66	0.7676	0.82	0.7981	0.78
RF	0.8857	0.92	0.8419	0.93	0.8952	0.93
GB	0.8857	0.91	0.8286	0.93	0.8610	0.91
KNN	0.8362	0.87	0.7390	0.84	0.7600	0.83
NB	0.6857	0.60	0.6933	0.69	0.7142	0.66

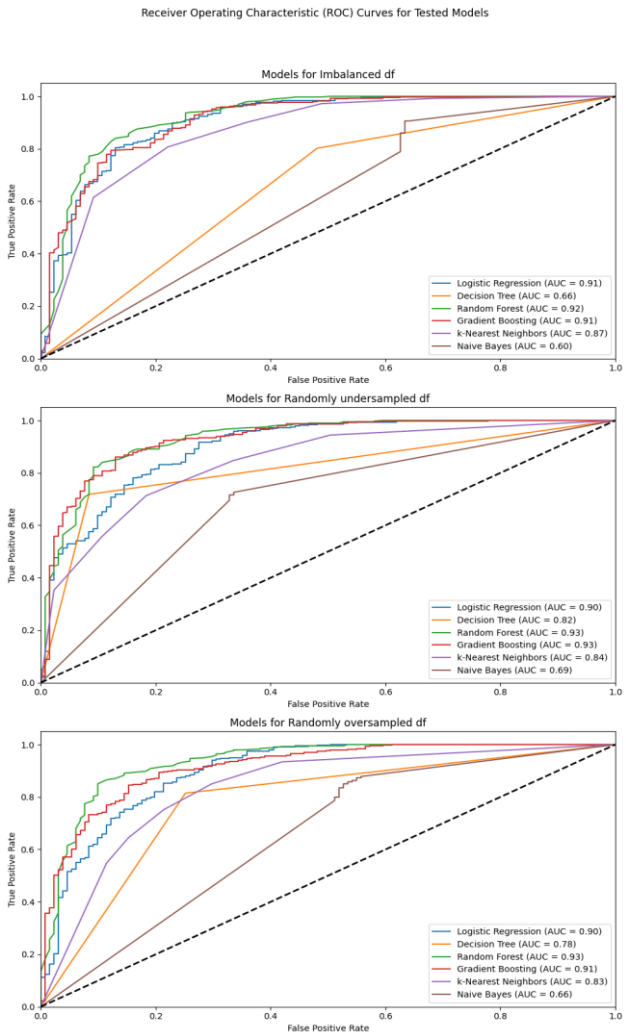

Fig 4. AUROC for baseline models

In our comparative analysis of various machine learning models across differently sampled datasets, we observed a notable performance variance. The Random Forest model, particularly when trained on the oversampled dataset, exhibited superior performance both in terms of accuracy (0.8952) and AUC (0.93), aligning it as the most effective model among those tested. This model's robustness is further highlighted by its impressive validation results on the oversampled dataset, where it achieved an average accuracy of 0.9695, indicating its strong predictive capabilities.

The Logistic Regression model, while generally stable across different sampling methods, achieved its best accuracy on the imbalanced dataset (0.8686), but saw reduced performance on the undersampled dataset. Similarly, Decision Trees showed an improvement in AUC when trained on the undersampled dataset (0.82), suggesting better handling of minority class instances, yet they did not perform as well as Random Forest.

Gradient Boosting and k-Nearest Neighbors demonstrated consistent results but did not reach the performance highs of the Random Forest model. Naive Bayes lagged behind the other models, particularly struggling on the imbalanced dataset with the lowest AUC (0.60), which slightly improved in the oversampled scenario.

The consistently high performance of the Random Forest model on the oversampled dataset suggests that this combination effectively counters the inherent class imbalance present in the original data. This method appears to mitigate overfitting—a common issue in machine learning where models might perform well on training data but poorly on unseen data. The scaffold splitting technique used in testing likely contributed to a rigorous assessment of generalizability across all models, emphasizing the Random Forest's capability to maintain accuracy and robustness in diverse conditions.

Furthermore, as aforementioned, we trained a neural network, as seen in [12].

The neural network trained on the oversampled dataset exhibited consistent improvement across multiple performance metrics during training. The AUROC (Area Under the Receiver Operating Characteristic Curve) increased notably from 0.83659 in the first epoch to 0.90741 by the tenth epoch, indicating enhanced model discrimination capability over time. Similarly, the AUPRC (Area Under the Precision-Recall Curve) improved, reflecting better precision and recall balance, peaking at 0.94430 in the final epoch. The F1 score, combining precision and recall, also showed progressive growth, reaching 0.93286 by the end of training. These metrics show that the model effectively learned from the oversampled dataset without apparent overfitting, as demonstrated by consistent gains in validation metrics without a corresponding increase in loss, which trended downward across epochs.

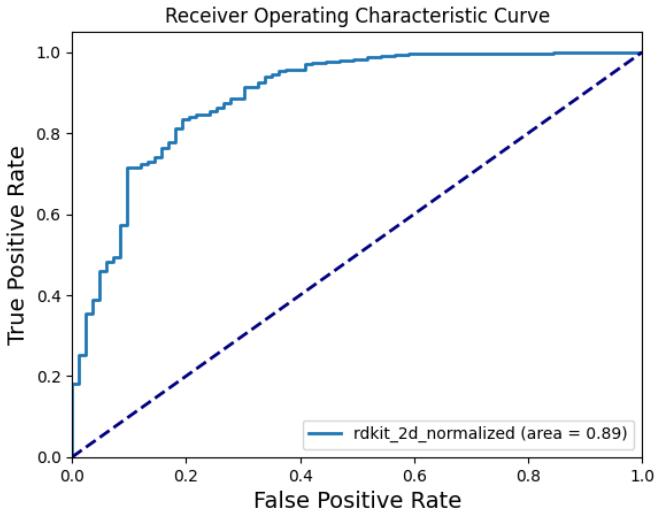

Fig 5. AUROC for NN

V. XAI

In the concluding chapter of our study, we delve into the application of Explainable Artificial Intelligence (XAI) techniques to enhance the interpretability and usability of our neural network model. By employing the exmol library, as used in [13] we integrated both LIME (Local Interpretable Model-agnostic Explanations) and counterfactual explanations into our analysis framework. These methods provide insight into how the model makes its predictions, offering users a transparent view of the predictive process.

To demonstrate the practical application of these XAI techniques, we have provided a fully functional codebase on our GitHub repository, the link to which is available in the introduction of this paper. This repository includes scripts that allow users to input SMILES strings of molecules to receive predictions along with comprehensive explanations generated by LIME and counterfactual analysis.

As part of our testing phase, we applied these XAI methods to three specific molecules from our test set. This allowed us to not only validate the model's predictive accuracy but also to showcase the detailed explanation capabilities provided by exmol. These case studies highlight the potential of our model to serve both as a robust predictive tool and a medium for gaining deeper insights into the molecular features that influence BBB penetration, thereby making our approach particularly valuable for researchers and practitioners in the field of drug discovery.

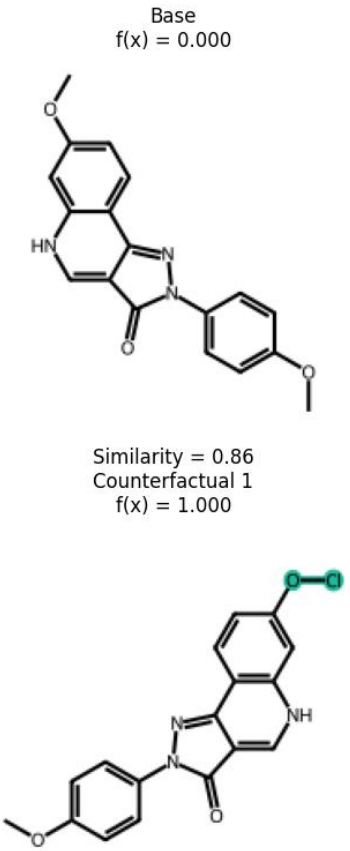

Fig 6. Example of prediction and counterfactual

VI. Conclusions

This study demonstrated the effective use of machine learning to predict blood-brain barrier (BBB) penetration, essential for CNS drug development. Among various models tested, the Random Forest model trained on an oversampled dataset proved most effective, addressing data imbalances and enhancing predictive accuracy. Additionally, the integration of explainable AI techniques, particularly through exmol for counterfactual explanations and LIME for local interpretability, provided crucial insights into molecular interactions influencing BBB penetration. These findings support the further application of advanced machine learning methods in drug discovery and underscore the potential of combining machine learning with cheminformatics to improve therapeutic development.

In our study, we chose not to incorporate uncertainty estimation methods such as those utilized in the referenced paper [1], which effectively employed techniques like entropy and Monte Carlo dropout to enhance prediction reliability. These methods enabled the GROVER-BBBp model to distinguish BBB+ from BBB− compounds with an accuracy of over 99% when predictions with high confidence

(uncertainty score < 0.1) were considered. Despite the compelling performance enhancements reported through the use of uncertainty estimation, our project prioritized a streamlined approach focused on traditional machine learning algorithms. This decision was influenced by our aim to maintain computational efficiency and model simplicity, avoiding the complexity and computational demands associated with advanced uncertainty measures. This allowed us to focus on maximizing direct performance metrics and interpretability without the additional layer of uncertainty analysis. Incorporating such techniques in future projects could potentially enhance model robustness and provide deeper insights into the confidence levels of the predictions, based on the promising results demonstrated in the referenced paper.

VII. References

[1] Tong, X., Wang, D., Ding, X., Tan, X., Ren, Q., Chen, G., Rong, Y., Xu, T., Huang, J., Jiang, H. and Zheng, M., 2022. Blood–brain barrier penetration prediction enhanced by uncertainty estimation. Journal of Cheminformatics, 14(1),p.44. DOI: https://doi.org/10.1186/s13321-022-00619-2

[2] Di L, Rong H, Feng B (2013) Demystifying brain penetration in central nervous system drug discovery. J Med Chem 56:2–12 DOI: 10.1021/jm301297f

[3] Colclough N, Chen K, Johnstrom P, Strittmatter N, Yan Y, Wrigley GL, Schou M, Goodwin R, Varnas K, Adua SJ et al (2021) Preclinical comparison of the blood-brain barrier permeability of osimertinib with other EGFR TKIs. Clin Cancer Res 27:189–201 DOI: 10.1158/1078-0432.CCR-19-1871

[4] Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3:711–716 DOI: 10.1038/nrd1470

[5] Brown PD, Ahluwalia MS, Khan OH, Asher AL, Wefel JS, Gondi V (2017) Whole-brain radiotherapy for brain metastases: evolution or revolution? J Clin Oncol 36:483–491DOI: 10.1200/JCO.2017.75.9589

[6] Patel NC (2020) Methods to optimize CNS exposure of drug candidates. Bioorg Med Chem Lett 30:127503 DOI: 10.1016/j.bmcl.2020.127503

[7] Gabathuler R (2010) Approaches to transport therapeutic drugs across the blood-brain barrier to treat brain diseases. Neurobiol Dis 37:48–57 DOI: 10.1016/j.nbd.2009.07.028

[8] Yu H, Yu Z, Jiang W, Hong L (2014) Lead compound optimization strategy (4)—improving blood-brain barrier permeability through structural modification. Acta Pharm Sin 49:789–799 https://pubmed.ncbi.nlm.nih.gov/25212022/

[9] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357. DOI: https://doi.org/10.1613/jair.953

[10] He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). Ieee. DOI: 10.1109/IJCNN.2008.4633969

[11] Sakiyama, H., Fukuda, M. and Okuno, T., 2021. Prediction of blood-brain barrier penetration (bbbp) based on molecular descriptors of the free-form and in-blood-form datasets. Molecules, 26(24), p.7428. DOI: https://doi.org/10.3390/molecules26247428

[12] Alsenan, S., Al-Turaiki, I. and Hafez, A., 2020. A recurrent neural network model to predict blood–brain barrier permeability. Computational Biology and Chemistry, 89, p.107377. DOI: https://doi.org/10.1016/j.compbiolchem.2020.107377

[13] Fradkin, P., Young, A., Atanackovic, L., Frey, B., Lee, L.J. and Wang, B., 2022. A graph neural network approach for molecule carcinogenicity prediction. Bioinformatics, 38(Supplement_1), pp.i84-i91.DOI: 10.1093/bioinformatics/btac266