AUTOMATED BCS DRUG CATEGORIZATION USING MACHINE LEARNING: AN EXPLAINABLE AI APPROACH FOR BCS CLASSIFICATION SYSTEM

Project Reference No.: *49S_BE_1530*

College : *PES University, Bengaluru*
Branch : *Department of Computer Science*
Guide(s) : *Dr Prajwala T R*
Dr Saniya Jawed
Student(s) : *Mr Hari Kiran K*
Mr Hari Shankar
Ms Krupa S
Mr Rhushya KC

Keywords:

Biopharmaceutics Classification System (BCS), Drug Classification, Molecular Descriptors, RDKit, LightGBM, Random Forest, XGBoost, CatBoost, Explainable Artificial Intelligence (XAI), SHAP (SHapley Additive exPlanations), Solubility Prediction (LogS), Molecular Fingerprints (ECFP4), Drug Discovery, Pharmacokinetics, Predictive Modelling

Introduction:

1. The Biopharmaceutics Classification System (BCS) is an important framework in pharmacology that classifies drug substances based on their solubility and permeability characteristics. Introduced in 1995, BCS has played a significant role in modern drug development, formulation strategies, and regulatory decision-making. The system provides a scientific basis for predicting the oral bioavailability of drugs, which is essential for understanding how effectively a drug is absorbed into the bloodstream after oral administration. Based on these physicochemical properties, drugs are classified into four categories: Class I (high solubility, high permeability), Class II (low solubility, high permeability), Class III (high solubility, low permeability), and Class IV (low solubility, low permeability). This classification helps researchers and regulatory authorities better understand the pharmacokinetic behaviour of drugs and supports the development of effective pharmaceutical formulations.
2. Despite its importance, traditional BCS classification methods rely on laboratory experiments that are often time-consuming, costly, and resource-intensive. These experimental procedures require specialised equipment and extensive testing, which makes them unsuitable for rapid screening during the early stages of drug discovery. With the increasing number of potential drug candidates being developed, there is a growing need for faster and more efficient computational approaches to predict BCS classes. Recent research has explored the use of machine learning techniques to analyse molecular descriptors and predict drug properties more efficiently.
3. In this project, a machine learning–based approach is proposed to automate the BCS classification process using molecular descriptors derived from drug structures. Techniques such as Random Forest, XGBoost, and CatBoost are used to build predictive models for classifying drugs into the four BCS categories. Additionally,

explainable artificial intelligence techniques like SHAP (SHapley Additive exPlanations) are incorporated to interpret model predictions and identify the key molecular features influencing classification. This approach aims to provide a faster, scalable, and more transparent method for predicting BCS classes, supporting pharmaceutical researchers in drug development and formulation design.

Objectives:

1. To develop a machine learning system for automated classification of drugs into BCS Classes I–IV.
2. This study aims to extract molecular descriptors and fingerprints from drug structures using RDKit.
3. The goal is to predict aqueous solubility (LogS) and perform BCS classification using models like Random Forest, XGBoost, and CatBoost.
4. Evaluate and compare different machine learning models to identify the best-performing approach.
5. We aim to incorporate Explainable AI techniques and create an interactive interface that facilitates real-time prediction and interpretation.

Methodology:

1. The Biopharmaceutics Classification System (BCS) is an important framework in pharmaceutical research that is used to categorise drug molecules based on their solubility and permeability. It helps researchers estimate oral drug absorption and design effective formulation strategies. By grouping drugs into four classes, BCS assists in understanding how drugs behave inside the human body. This classification plays a crucial role during drug development and regulatory evaluation.
2. Traditional laboratory-based methods used to determine BCS classes are highly accurate. However, these experimental procedures are slow, expensive, and resource-intensive. They also require specialised laboratory equipment and trained personnel. Because of these limitations, traditional methods are not suitable for rapid screening of large numbers of drug candidates in the early stages of drug discovery.
3. To address these challenges, this project proposes a comprehensive machine-learning-driven pipeline that automates the BCS classification process. The system integrates molecular descriptors, predictive modelling, and explainable artificial intelligence techniques to provide accurate and interpretable predictions.
4. As part of this pipeline, dataset preparation is carried out to ensure data quality and reliability. Two datasets are used: a solubility dataset containing approximately 9,943 drug compounds with SMILES and LogS values and a BCS-classified dataset consisting of 390 drugs with known class labels. The data is collected from publicly available sources and preprocessed by removing invalid or non-canonical SMILES representations and handling missing values using RDKit. The cleaned dataset is then split into training and testing sets for model development and evaluation.
5. The process begins with user-provided SMILES strings that represent the molecular structure of drug compounds. These SMILES strings are converted into detailed

physicochemical descriptors using the RDKit cheminformatics toolkit. These descriptors capture important molecular properties that influence drug behaviours.

6. Next, a regression model is used to predict aqueous solubility (LogS), which is a key factor in BCS classification. After estimating solubility, a comparative evaluation of three machine learning classification algorithms is performed. The algorithms used in this project include Random Forest, XGBoost, and CatBoost.
7. Among these models, CatBoost has shown promising preliminary results in early evaluations. It appears to offer better balance across the four BCS classes compared to the other models tested so far. A comprehensive evaluation is currently underway to confirm the most suitable model for the final pipeline.
8. To enhance transparency and interpretability, explainable AI techniques are planned to be incorporated into the system in the next phase. Specifically, SHAP (SHapley Additive exPlanations) is being explored to interpret both solubility predictions and classification decisions. The integration of SHAP analysis is currently in progress and will be completed as part of the remaining work.
9. Once SHAP analysis is fully integrated, the system is expected to reveal the influence of key physicochemical properties such as lipophilicity, polar surface area, and molecular weight on BCS class predictions. These insights will be validated and reported in the final phase of the project.
10. As part of the remaining work, a lightweight interactive interface is planned to be developed. This interface will allow users to input drug molecules, visualise model predictions, and explore the explanation results. The design and first prototype of this interface are not done yet.
11. Overall, this project is progressing towards demonstrating how machine learning and explainable AI can improve the speed, transparency, and scalability of BCS classification. The core pipeline has been established and is being refined, with the remaining modules focused on XAI integration and interface development currently in progress.

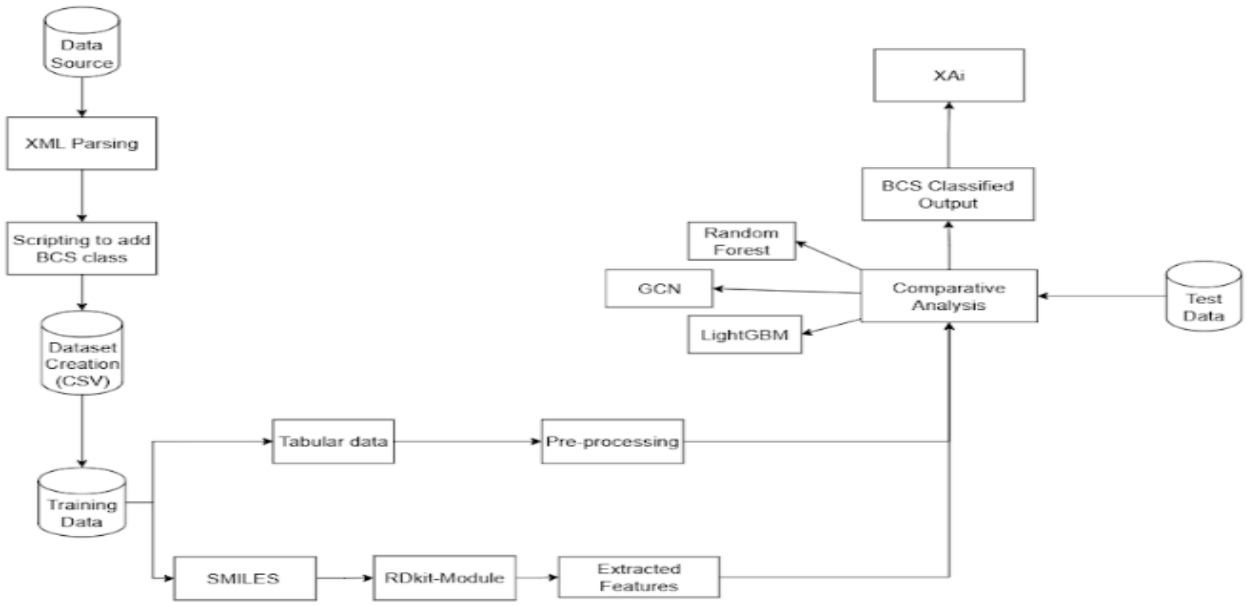

Figure 1: Proposed Architecture for Automated BCS Drug Classification

Result and Conclusion:

1. The proposed machine learning pipeline has been partially evaluated as part of the ongoing project work. So far, the evaluation has focused on the solubility prediction phase (LogS) and initial training of classification models for predicting BCS drug classes. Preliminary results indicate that the system can effectively model molecular relationships using molecular descriptors and fingerprints, though comprehensive evaluation is still underway.
2. For the solubility prediction stage, a comparative analysis was conducted using three models: LightGBM, Attentive FP, and Graph Convolutional Network. These models were trained using molecular descriptors and the ECFP4 fingerprint matrix to estimate the aqueous solubility (LogS) value. Among them, LightGBM produced the best performance with an R^2 score of 0.8512, indicating strong predictive capability. The Graph Convolutional Network model achieved an R^2 score of 0.708, as it relied mainly on molecular graph structures. The Attentive FP model performed slightly better than GCN with an R^2 score of 0.762, but its performance was still lower compared to LightGBM. Based on this comparative analysis, the LightGBM model was selected as the final model for predicting solubility values within the pipeline.
3. In the second phase, initial training and evaluation of classification models has been carried out. The models explored include Random Forest, XGBoost, and CatBoost, tested with different combinations of input features such as LogS, LogP, molecular descriptors, and ECFP4 fingerprints. Early observations suggest that models trained with both descriptors and ECFP4 fingerprints tend to outperform those relying only on basic physicochemical properties, though further testing is needed to confirm these trends.
4. Among the classifiers evaluated so far, CatBoost has shown a relatively better balance between precision and recall with a preliminary accuracy of approximately 0.75. XGBoost achieved similar accuracy of around 0.76 in initial tests but appeared to show some signs of overfitting for certain classes. Random Forest demonstrated stable performance but faced challenges with compounds near the boundary between Class II and Class IV. These observations are preliminary and further hyperparameter tuning and cross-validation are yet to be completed.
5. Preliminary confusion matrix analysis suggests that most misclassifications tend to occur between Class II and Class III drugs, likely due to their overlapping physicochemical characteristics. A detailed class-wise analysis and error analysis is planned as part of the remaining work to better understand these patterns.
6. At the current stage, approximately 80% of the project has been completed. The core pipeline including dataset preparation, molecular descriptor extraction, solubility prediction, and initial classifier training and comparison has been successfully implemented. The remaining work involves finalising the best classification model, integrating SHAP-based explanations, developing the interactive user interface, and conducting thorough validation and testing. These pending tasks form the final 20% of the project and are expected to be completed in the upcoming phase.

References:

[1] M. A. Ghanavati, S. Ahmadi, and S. Rohani, “A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: a comparative model and dataset analysis,” RSC Advances, 2024.
[2] M. Oja, S. Sild, and U. Maran, “Logistic classification models for pHpermeability profile: predicting permeability classes for the Biopharmaceutical Classification System,” ACS Publications, 2019.
[3] S. Y. K. Fong et al., “Establishing the pharmaceutical quality of Chinese herbal medicine: a provisional BCS classification,” ACS Publications, 2013.
[4] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, “Artificial intelligence to deep learning: machine intelligence approach for drug discovery,” Springer Nature, 2021.
[5] T. Hou, J. Wang, W. Zhang, W. Wang, and X. Xu, “Recent advances in computational prediction of drug absorption and permeability in drug discovery,” Current Medicinal Chemistry, 2006.
[6] D. Newby, A. A. Freitas, and T. Ghafourian, “Comparing multilabel classification methods for provisional biopharmaceutics class prediction,” Molecular Pharmaceutics, 2015.
[7] B. B. K. Reddy and A. Karunakar, “Biopharmaceutics Classification System: A regulatory approach,” Dissolution Technologies, 2011.
[8] F. Truzzi, C. Tibaldi, Y. Zhang, G. Dinelli, and E. D’Amen, “An overview on dietary polyphenols and their biopharmaceutical classification system (BCS),” International Journal of Molecular Sciences, 2021.
[9] Z. Wu, N. Wang, Z. Ye, H. Xu, G. Chan, and D. Ouyang, “FormulationBCS: A machine learning platform based on diverse molecular representations for BCS class prediction,” Molecular Pharmaceutics, vol. 22, pp. 330–342, 2025.

Future Scope:

The future scope of this project includes:

1. Enhance the model accuracy by training on larger and more diverse drug datasets.
2. Incorporate additional molecular descriptors and physicochemical features for improved prediction performance.
3. Extend the model to handle multi-class and borderline BCS classifications more effectively.
4. Integrate the system with Computer-Aided Drug Design (CADD) tools for real-time molecular optimisation.
5. Develop a user-friendly interface or web application for easier access by researchers and scientists.
6. Enable real-time screening of large chemical libraries for faster lead identification.
7. Improve model generalisation using advanced techniques such as deep learning and ensemble methods.
8. Validate the model with experimental (in-vitro/in-vivo) data for higher reliability.

9. Integrate the framework with pharmaceutical databases and decision-support systems.
10. Expand the system to support Quality by Design (QbD) approaches for early identification of critical quality attributes.
11. Explore its application in regulatory evaluation and formulation development.
12. Optimise computational efficiency for large-scale industrial use.