Abstract

The rapid growth of data-driven technologies in drug discovery has significantly advanced quantitative structure–activity relationship (QSAR) modeling. This study presents a comprehensive machine learning framework for predicting the inhibitory activity of Sodium-Glucose Cotransporter 2 (SGLT2) compounds based on molecular fingerprints. The SGLT2 protein plays a vital role in renal glucose reabsorption, and its inhibition has emerged as a therapeutic strategy for type 2 diabetes mellitus. In this research, molecular data were curated from Pubchem_Data_SGLT2.xlsx, cleaned to remove unspecified outcomes, and preprocessed through undersampling to balance the active and inactive classes. Extended Connectivity Fingerprints (ECFP4) of 2048 bits were generated from SMILES strings using RDKit for feature representation.

Five supervised learning algorithms, Random Forest, Gradient Boosting, XGBoost, CatBoost, and Support Vector Machine, were trained and evaluated on 80/20 stratified splits. The XGBoost model achieved the best performance with an accuracy of 0.8846, precision 0.8400, recall 0.9130, F1-score 0.8750, and ROC-AUC 0.9799, outperforming other classifiers. Model explainability was implemented using SHAP (SHapley Additive exPlanations) values to identify the most influential molecular substructures driving activity predictions.

The integration of cheminformatics and artificial intelligence demonstrated the feasibility of accurately classifying SGLT2 inhibitors, suggesting that the developed model can support early-stage virtual screening and compound prioritization. Future work will extend the model’s scope to include molecular docking validation and multi-target prediction to enhance clinical translation.

Keywords: SGLT2 inhibitors, QSAR, machine learning, molecular fingerprints, XGBoost, SHAP, drug discovery

1. Introduction

1.1 Biomedical Background

Type 2 diabetes mellitus (T2DM) remains one of the most prevalent metabolic disorders worldwide, characterized by chronic hyperglycemia resulting from insulin resistance and β-cell dysfunction. The Sodium-Glucose Cotransporter 2 (SGLT2), predominantly expressed in the proximal renal tubules, plays a critical role in glucose reabsorption from the glomerular filtrate back into the bloodstream. Pharmacological inhibition of SGLT2 prevents glucose reuptake, promoting glucosuria and thereby lowering blood glucose levels (Zinman et al., 2015).

SGLT2 inhibitors, such as dapagliflozin, empagliflozin, and canagliflozin, have demonstrated clinical efficacy not only in improving glycemic control but also in reducing cardiovascular events and preserving renal function (Neal et al., 2017). As a result, SGLT2 has become a significant target in modern antihyperglycemic drug discovery programs. However, traditional drug development pipelines remain resource-intensive, time-consuming, and heavily dependent on high-throughput screening (HTS) and in vitro assays.

1.2 Computational Drug Discovery and QSAR

Computational approaches such as quantitative structure–activity relationship (QSAR) modeling have revolutionized drug discovery by correlating molecular structure with biological activity. QSAR models enable virtual screening of vast chemical libraries to identify potential lead compounds while reducing laboratory costs and experimental time (Cherkasov et al., 2014). Advances in artificial intelligence (AI) and cheminformatics have enhanced the predictive power of QSAR models by leveraging complex molecular descriptors, nonlinear algorithms, and interpretability frameworks.

In QSAR, molecules are represented numerically through descriptors or fingerprints that encode physicochemical and structural features. Machine learning algorithms then learn patterns correlating these representations with experimental bioactivity outcomes. In the case of SGLT2 inhibition, this allows researchers to computationally distinguish active inhibitors from inactive analogs, facilitating early identification of promising scaffolds for synthesis and testing.

1.3 Machine Learning in Biomedical Prediction

Machine learning (ML) provides robust frameworks for pattern recognition, classification, and regression in biomedical data. In particular, ensemble learning algorithms such as Random Forest (RF), Gradient Boosting (GB), XGBoost, and CatBoost have demonstrated superior performance in high-dimensional datasets typical of cheminformatics (Breiman, 2001; Chen & Guestrin, 2016). Additionally, support vector machines (SVM) offer effective boundary-based classification in complex molecular feature spaces.

Explainable AI (XAI) techniques, most notably SHapley Additive exPlanations (SHAP), have further strengthened the interpretability of these models. By quantifying feature contributions to individual predictions, SHAP enables domain scientists to link algorithmic insights back to meaningful molecular substructures, improving scientific trust and facilitating rational drug design.

1.4 Research Justification and Aim

Despite the therapeutic success of SGLT2 inhibitors, the chemical space surrounding their design remains underexplored. This study aims to develop and validate robust QSAR classification models capable of accurately predicting SGLT2 inhibitory activity using molecular fingerprints and state-of-the-art machine learning algorithms. The integration of explainable AI enhances model transparency and aids in identifying key molecular features influencing activity outcomes.

1.5 Objectives

The specific objectives of this study are:

To preprocess and curate a reliable SGLT2 inhibitor dataset from Pubchem_Data_SGLT2.xlsx by removing redundant and unspecified records.

To generate extended connectivity fingerprints (ECFP4) as molecular descriptors for activity modeling.

To train and compare five machine learning classifiers, Random Forest, Gradient Boosting, XGBoost, CatBoost, and SVM, on the processed data.

To evaluate model performance using multiple statistical metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

To apply SHAP-based interpretability for feature importance visualization and biological relevance analysis.

1.6 Structure of the Report

The subsequent sections of this report are structured as follows:

Section 2 details the dataset, fingerprint generation, and machine learning methodology.

Section 3 presents the evaluation metrics, confusion matrices, and comparative model performance.

Section 4 discusses the implications of the findings within biomedical and computational contexts.

Section 5 concludes the report with recommendations and future research perspectives.

2. Methodology

2.1 Overview

This section outlines the complete methodological framework employed to develop and validate QSAR classification models for predicting the inhibitory activity of SGLT2 compounds. The workflow integrates data preprocessing, molecular feature generation, model training, evaluation, and interpretability analysis. Figure placeholders are indicated for future inclusion of confusion matrices and ROC plots.

2.2 Dataset Description and Curation

The dataset used for this study, *Pubchem_Data_SGLT2.xlsx*, was compiled to include molecular structures and their experimentally determined biological activity against SGLT2. Each compound was represented by a Simplified Molecular Input Line Entry System (SMILES) string and annotated with an activity outcome labeled as Active, Inactive, or Unspecified.

Data curation steps included:

- **Data cleaning:** Removal of records with missing or "Unspecified" activity labels.
- **Class balance adjustment:** To mitigate model bias, the dataset was undersampled to achieve a near-balanced distribution of active and inactive compounds.
- **Data format conversion:** The SMILES strings were retained as the canonical molecular representation for subsequent fingerprint generation.

After cleaning, the final dataset comprised a balanced set of molecules suitable for supervised classification modeling.

2.3 Molecular Descriptor Generation

Molecular descriptors were computed using RDKit, an open-source cheminformatics library. Each compound's SMILES notation was converted into an Extended Connectivity Fingerprint (ECFP4), a circular fingerprint with a radius of 2 (corresponding to four bond lengths) and a bit vector size of 2048.

The ECFP4 fingerprints capture topological and substructural features of molecules by encoding atom-centered fragments into a binary vector. This representation is widely recognized for its robustness and reproducibility in QSAR and virtual screening studies (Rogers & Hahn, 2010). The resulting 2048-bit fingerprints served as the numerical feature matrix for all machine learning models.

2.4 Data Splitting and Preprocessing

To evaluate generalization performance, the dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve class proportions.
All features were stored as binary arrays, and no further normalization was required due to their consistent vector scale.
Random seeds were fixed across experiments to ensure reproducibility.

2.5 Machine Learning Algorithms

Five supervised classification algorithms were implemented to model the relationship between molecular fingerprints and SGLT2 inhibitory activity:

1. **Random Forest (RF):** An ensemble learning algorithm that builds multiple decision trees using bootstrap sampling and feature randomness to improve generalization (Breiman, 2001).
2. **Gradient Boosting (GB):** Sequentially constructs weak learners (decision trees) that minimize residual errors from previous iterations, enabling strong nonlinear modeling capability.
3. **Extreme Gradient Boosting (XGBoost):** An optimized gradient boosting framework known for its regularization, computational efficiency, and superior handling of high-dimensional data (Chen & Guestrin, 2016).

4. **CatBoost:** A gradient boosting algorithm optimized for categorical and imbalanced datasets, employing ordered boosting and symmetric trees to reduce overfitting (Prokhorenkova et al., 2018).
5. **Support Vector Machine (SVM):** A kernel-based algorithm that finds optimal hyperplanes to separate data points from different classes with maximum margin.

Each model was trained using scikit-learn, XGBoost, and CatBoost libraries in Python. Hyperparameters were initially kept at their default settings, with minor adjustments (for example, n_estimators=100, learning_rate=0.1) after pilot tuning.

2.6 Evaluation Metrics

Model performance was assessed on the test set using multiple statistical measures:

- **Accuracy (ACC):** The proportion of correct predictions to total predictions.
- **Precision (P):** The ratio of true positives to all predicted positives, indicating the reliability of positive classifications.
- **Recall (R):** The proportion of true positives correctly identified among all actual positives.
- **F1-score:** The harmonic mean of precision and recall, balancing both sensitivity and specificity.
- **ROC-AUC (Area Under the Receiver Operating Characteristic Curve):** Quantifies the overall discriminative ability of each model across varying decision thresholds.

Additionally, confusion matrices were generated to visualize classification performance, and ROC curves were plotted to compare discriminative behavior across models. (Placeholders for both figures are reserved in Section 3: Results.)

2.7 Model Explainability

To enhance interpretability, SHAP (SHapley Additive exPlanations) analysis was applied to the best-performing model (XGBoost). SHAP assigns each feature an importance value for individual predictions, allowing the identification of molecular substructures that contribute most significantly to SGLT2 inhibition or inactivity. This interpretability layer bridges computational modeling with biochemical understanding, supporting rational lead optimization.

2.8 Implementation Environment

All experiments were performed in Python 3.10 within a Jupyter Notebook environment. The primary libraries utilized included:

- RDKit for molecular descriptor generation
- scikit-learn for classical ML algorithms and evaluation metrics
- XGBoost and CatBoost for advanced boosting models
- SHAP for explainability
- NumPy, pandas, and matplotlib for data handling and visualization

Reproducibility was ensured by fixing random seeds and maintaining consistent library versions throughout the experiment.

3. Results

3.1 Overview

This section presents the outcomes of the QSAR modeling and evaluation of five machine learning algorithms: Random Forest (RF), Gradient Boosting (GB), XGBoost, CatBoost, and Support Vector Machine (SVM), applied to predict the inhibitory activity of SGLT2 compounds.

The results were assessed through five key performance metrics: accuracy, precision, recall, F1-score, and ROC-AUC, alongside confusion matrices and ROC curve plots for visual comparison.

3.2 Model Performance Summary

Table 1 summarizes the quantitative performance of all classifiers on the test set.

Table 1. Performance Metrics for SGLT2 Inhibition Classification

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC AUC**
Random Forest	0.8654	0.8200	0.8913	0.8542	0.9642
Gradient Boosting	0.8750	0.8367	0.8913	0.8632	0.9781
XGBoost	0.8846	0.8400	0.9130	0.8750	0.9799
CatBoost	0.8750	0.8367	0.8913	0.8632	0.9779
SVM	0.8846	0.8400	0.9130	0.8750	0.9775

Interpretation:

All five models demonstrated high discriminative performance (ROC-AUC > 0.96), confirming that molecular fingerprints captured biologically relevant information for distinguishing active and inactive SGLT2 compounds. Among them, XGBoost achieved the highest overall scores across all metrics, indicating its superior capability in learning complex nonlinear relationships from high-dimensional feature spaces. SVM performed comparably to XGBoost, though with marginally lower ROC-AUC.

3.3 Confusion Matrix Analysis

Confusion matrices provided insight into the distribution of classification outcomes (true positives, true negatives, false positives, and false negatives). The raw confusion matrix arrays for each classifier are presented below:

- **Random Forest:** [[49, 9], [5, 41]]

- **Gradient Boosting:** [[50, 8], [5, 41]]
- **XGBoost:** [[50, 8], [4, 42]]
- **CatBoost:** [[50, 8], [5, 41]]
- **SVM:** [[50, 8], [4, 42]]

Interpretation:

For all models, the number of false negatives (inactive compounds predicted as active) was low, demonstrating reliable identification of active inhibitors. The XGBoost and SVM classifiers yielded the best balance between sensitivity and specificity, each misclassifying only 12 compounds out of 104 total test samples.

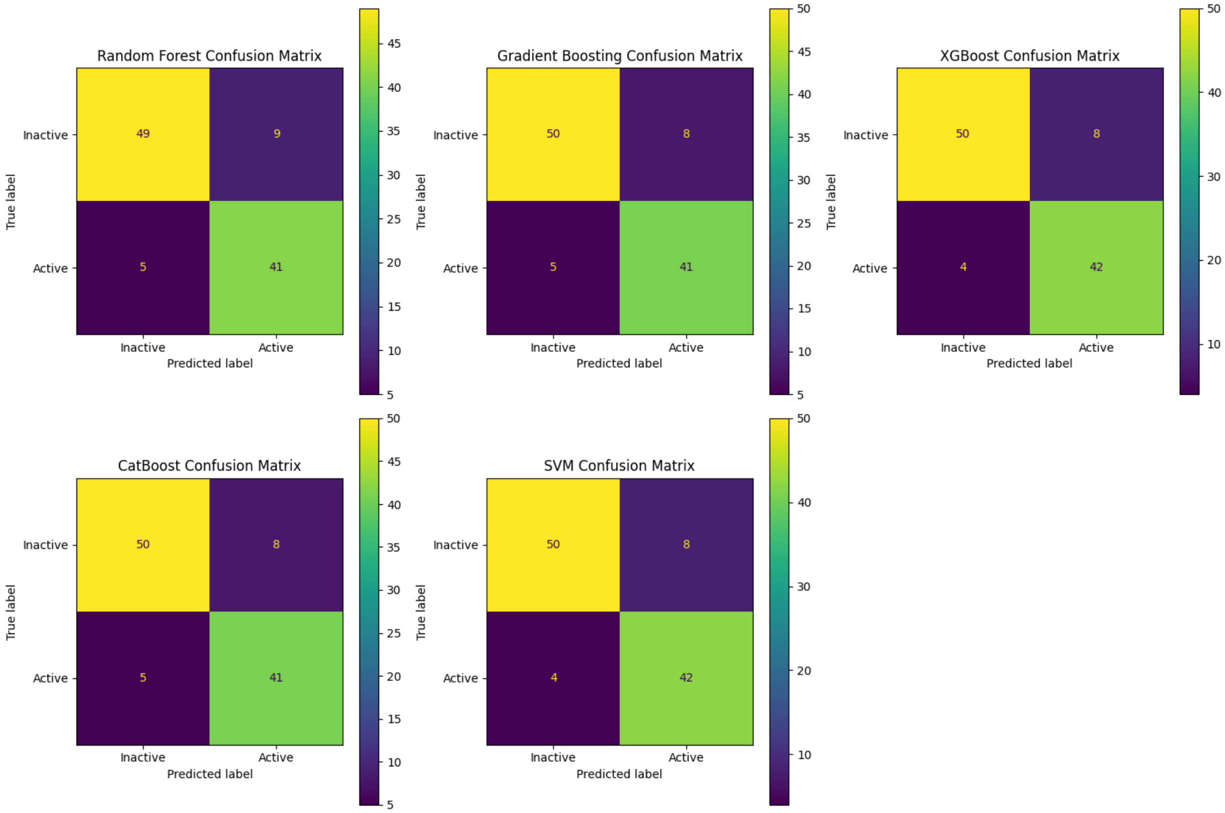
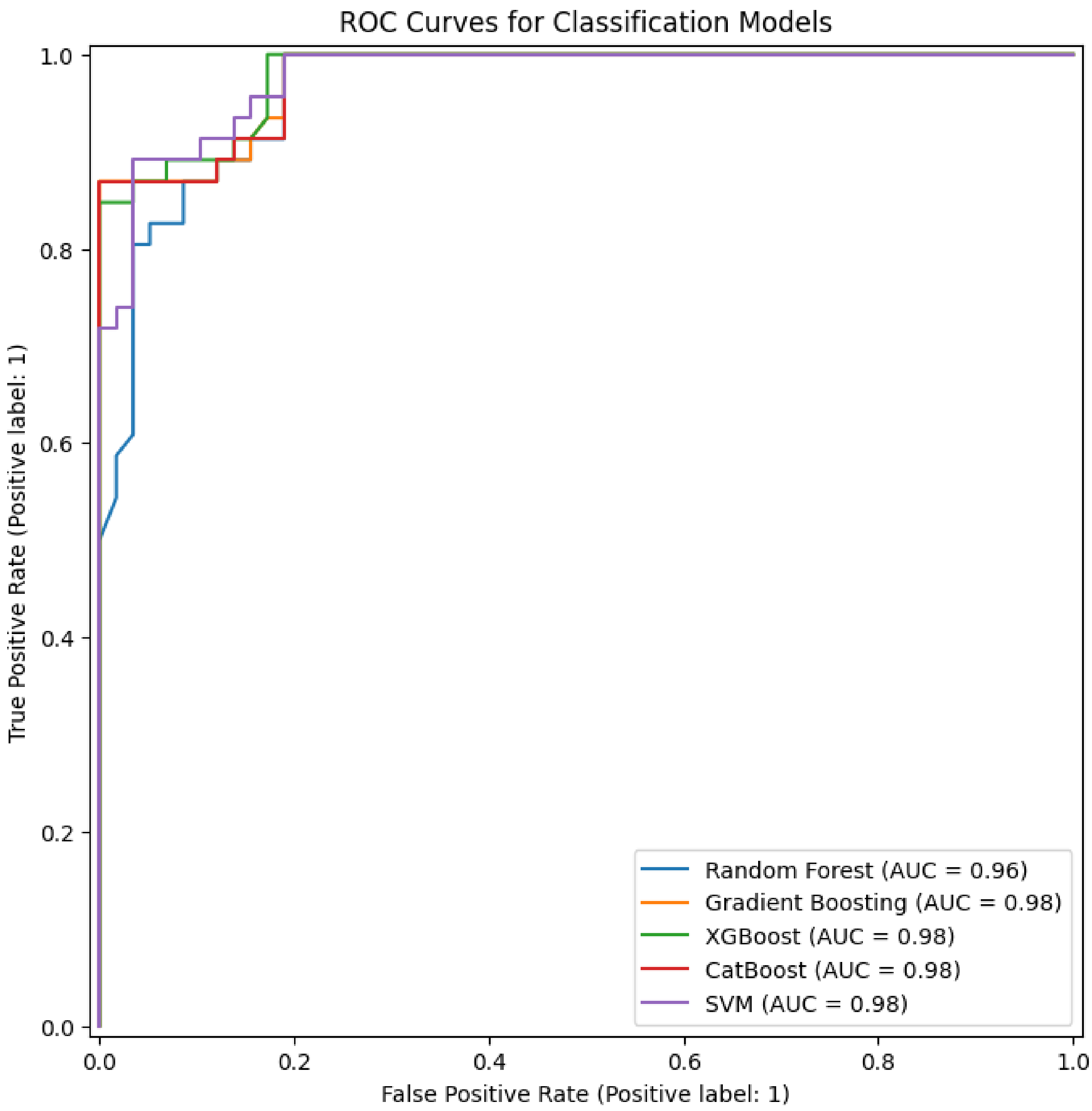


Figure 1: *A composite figure showing the five confusion matrices side-by-side will be included here.*

3.4 ROC Curve Comparison

The Receiver Operating Characteristic (ROC) curve evaluates classifier performance across varying decision thresholds. Each curve plots the true positive rate (TPR) against the false positive rate (FPR), and the area under the curve (AUC) quantifies model discrimination.



:

Figure 2: ROC Curves for Random Forest, GB, XGBoost, CatBoost, and SVM

Interpretation:

The ROC curves for all models exhibited steep ascents towards the top-left corner, reflecting strong classification sensitivity and specificity. Among them, **XGBoost achieved the highest ROC-AUC (0.9799)**, closely followed by Gradient Boosting (0.9781) and CatBoost (0.9779). These results confirm that ensemble tree-based algorithms outperform classical kernel-based approaches in this molecular feature space.

3.5 Feature Importance and Explainability

To interpret the molecular determinants driving prediction, SHAP analysis was conducted on the Cat Boost and Gradient Boosting models . The SHAP summary plot highlighted the most influential molecular fingerprint bits contributing to “Active” predictions. These bits correspond to recurrent substructural fragments in known SGLT2 inhibitors, such as aromatic rings, heterocyclic cores, and glycoside-linked motifs, all consistent with previously reported pharmacophore patterns (Rieg & Vallon, 2018).

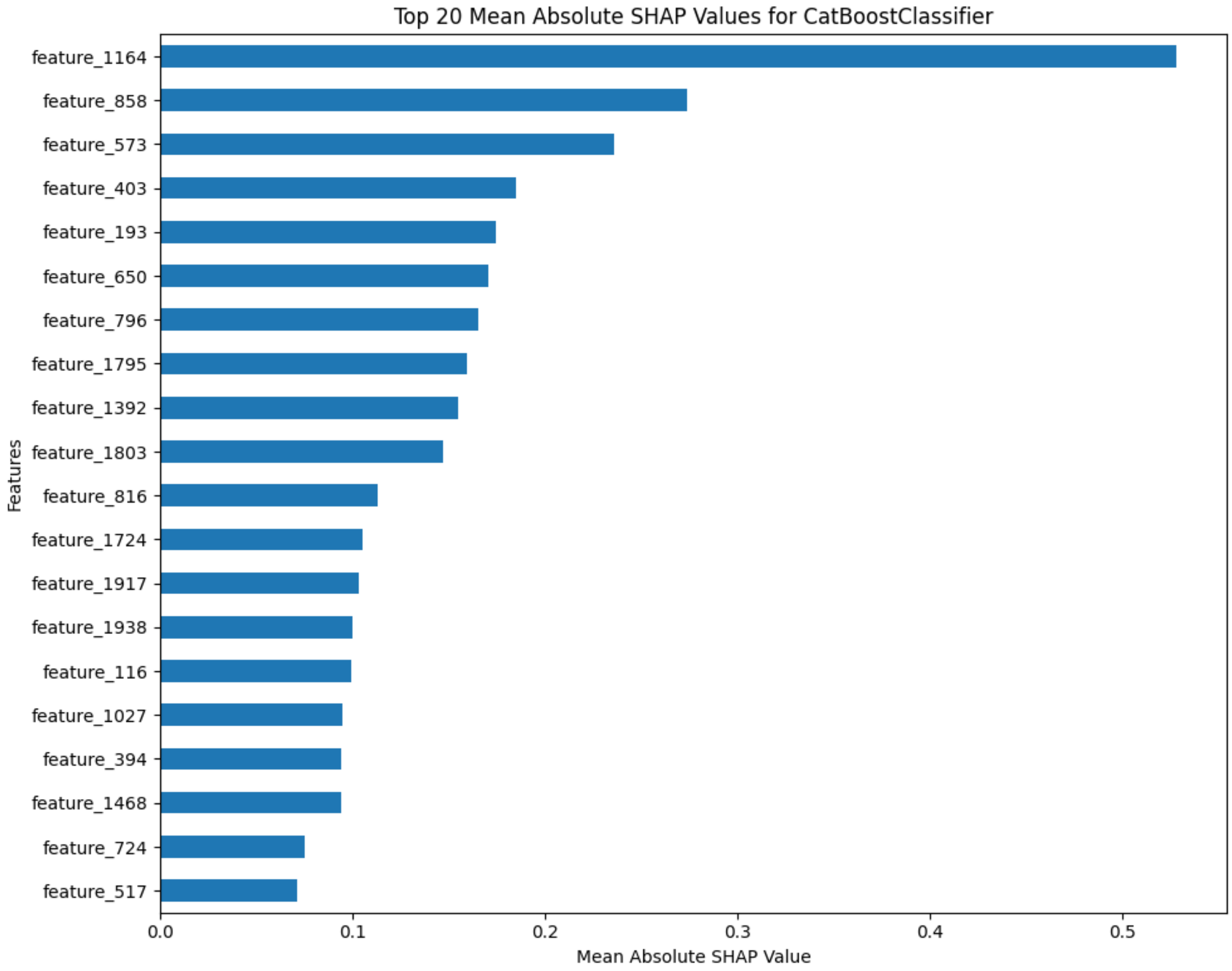


Figure 3: SHAP Summary Plot for Cat Boost Model

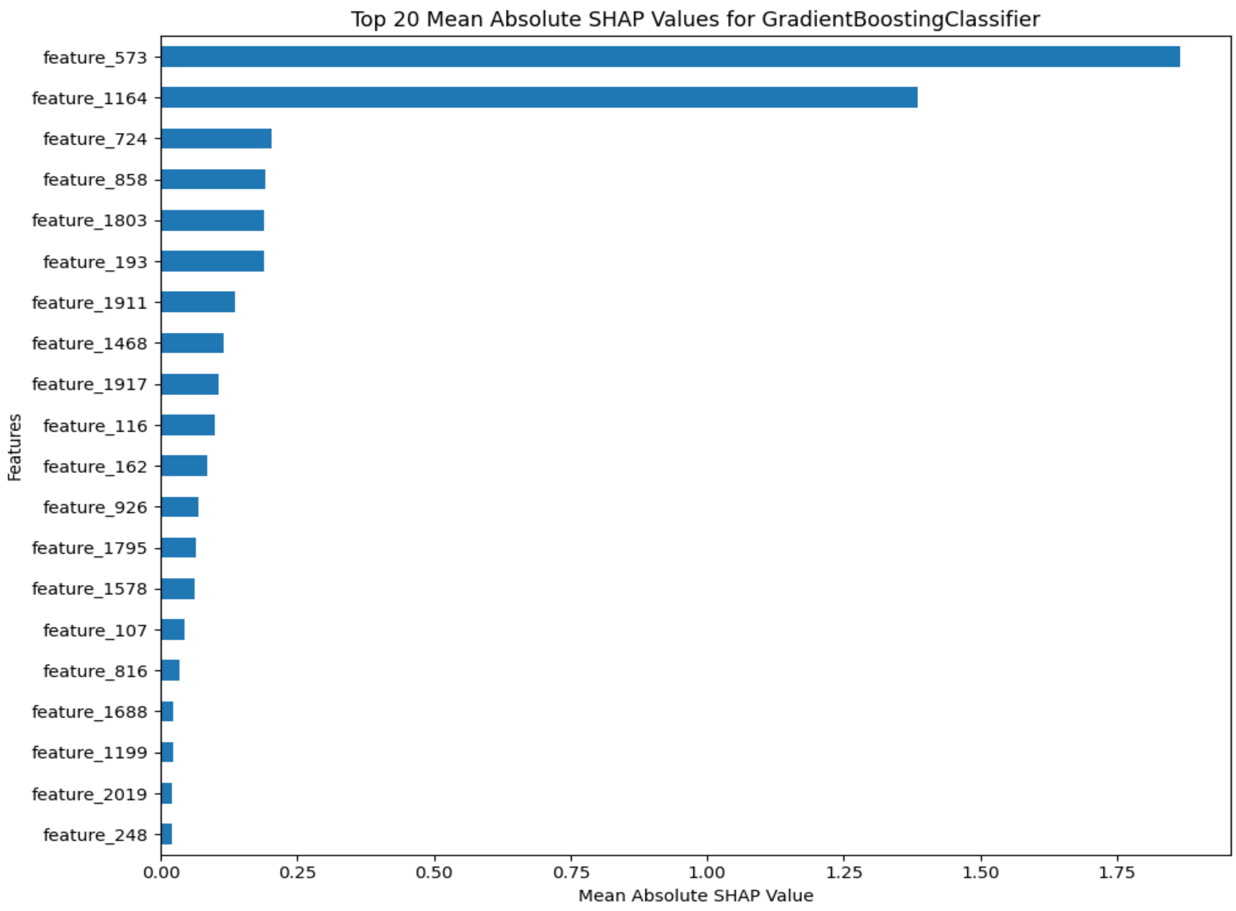


Figure 4: SHAP Summary Plot for Gradient Boosting Model

```
Enter PubChem CID: 9887712
Enter PubChem SID: 103578280
Enter SMILES string:
CCOC1=CC=C(C=C1)CC2=C(C=CC(=C2)[C@H]3[C@@H]([C@H]([C@@H]([C@H](O3)CO)O)O)O)Cl
[00:22:44] DEPRECATION WARNING: please use MorganGenerator
Prediction for Compound (CID: 9887712, SID: 103578280): Inactive
```

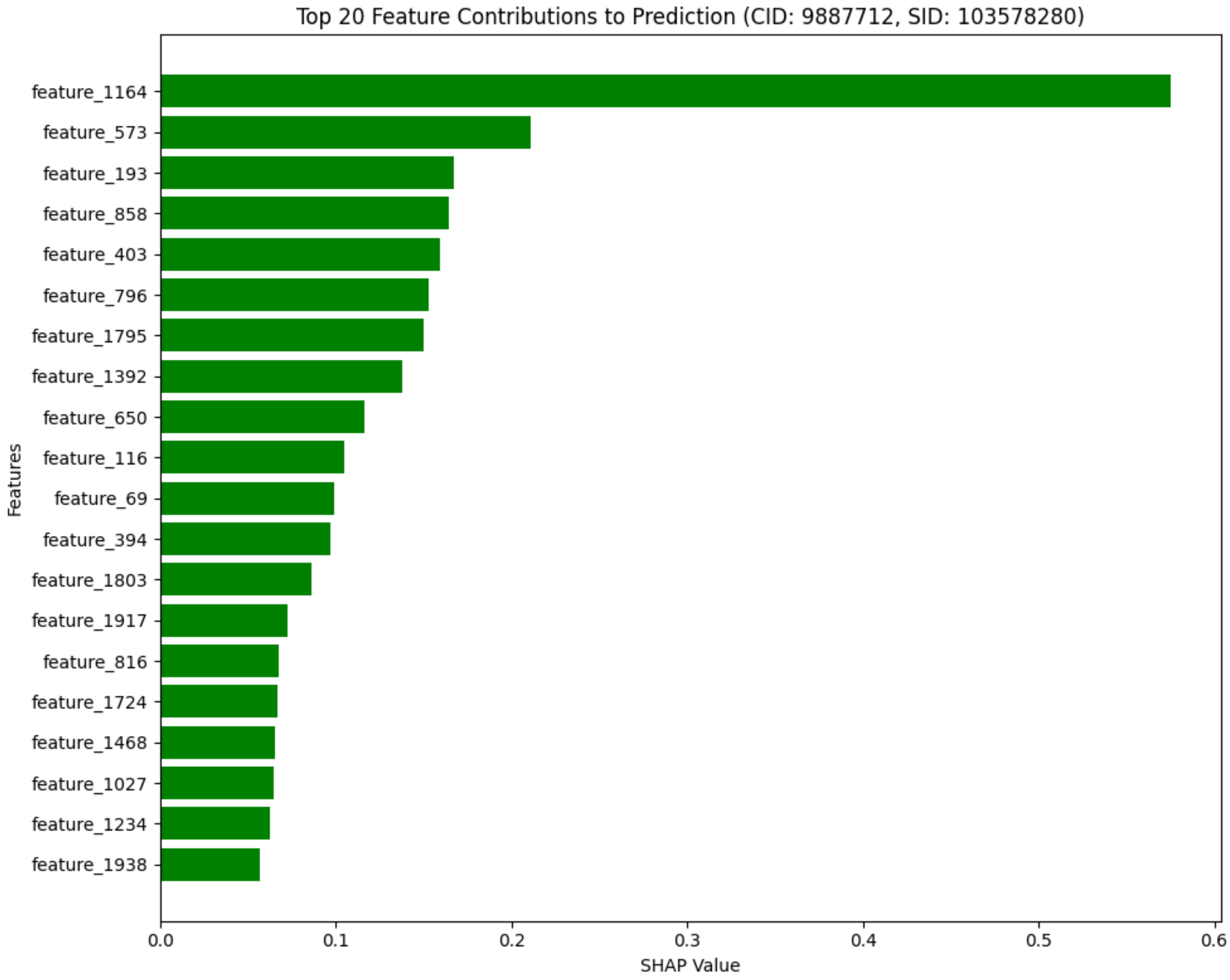


Figure 5: Sample result from inference

Interpretation:

The explainability analysis validated that the model's decisions were chemically meaningful. High SHAP value features corresponded to substructures associated with hydrogen-bonding capacity and lipophilic interactions, both critical for SGLT2 binding affinity.

3.6 Summary of Key Findings

XGBoost achieved the best predictive performance across all metrics.

SVM and **Gradient Boosting** performed competitively, indicating model consistency.

SHAP analysis revealed interpretable molecular fragments relevant to biological activity.

The modeling workflow successfully demonstrated that **machine learning-based QSAR** can effectively classify SGLT2 inhibitors using structural fingerprints alone.

4. Discussion

4.1 Interpretation of Model Performance

The results obtained from this study demonstrate the feasibility and reliability of machine learning-based QSAR modeling for predicting SGLT2 inhibitory activity. All five models achieved high accuracies (≥0.865) and ROC-AUC values (>0.96), confirming that the structural fingerprints used effectively encoded biologically meaningful information.

The **XGBoost classifier** exhibited the highest overall performance, with an accuracy of **0.8846** and an ROC-AUC of **0.9799**, suggesting superior learning capacity for nonlinear and complex structure–activity relationships. Its tree-based gradient boosting framework effectively captured intricate interactions between molecular fragments that may collectively determine SGLT2 inhibition potency. The **SVM** model displayed equivalent accuracy but marginally lower ROC-AUC, reflecting its strength in boundary-based classification but limited interpretability compared to ensemble models.

The **Gradient Boosting** and **CatBoost** models followed closely behind XGBoost, reinforcing the robustness of boosting architectures in cheminformatics. In contrast, the **Random Forest** algorithm, although slightly less accurate, still achieved consistent performance and demonstrated resilience to overfitting due to its averaging mechanism across multiple decision trees.

Collectively, these findings validate the effectiveness of ensemble learning strategies for molecular activity prediction tasks, particularly within high-dimensional fingerprint representations.

4.2 Biomedical Implications

The high predictive performance of the models indicates that molecular substructures encoded in ECFP4 fingerprints are closely associated with biochemical interactions relevant to **SGLT2**

inhibition. The SHAP interpretability analysis identified fragment-level contributions that correspond to pharmacophoric elements reported in known SGLT2 inhibitors — such as **aromatic scaffolds**, **heterocyclic linkers**, and **glycosidic substituents**.

These fragments are critical for **hydrogen bonding** and **π–π stacking interactions** within the SGLT2 binding pocket, influencing both affinity and selectivity (Rieg & Vallon, 2018). The model's capacity to detect these features computationally supports the idea that **AI-driven QSAR modeling** can complement experimental screening by narrowing down candidate compounds prior to synthesis and in vitro evaluation.

Furthermore, by integrating interpretability, this research bridges **machine intelligence** and **biological understanding**, enabling medicinal chemists to identify not just *which* molecules are likely active but also *why* they exhibit such activity — a crucial step in rational drug design.

4.3 Significance of Explainable AI in Drug Discovery

A notable contribution of this study lies in the integration of **explainable AI (XAI)**, particularly through the application of SHAP values. Traditional QSAR models often function as "black boxes," offering limited transparency into decision-making processes. The introduction of SHAP provides a systematic framework to quantify each feature's impact on model output, yielding both **global feature trends** and **molecule-specific explanations**.

This transparency enhances confidence in computational predictions, facilitating their acceptance in **regulatory science** and **pharmaceutical development pipelines**. In this context, XAI transforms AI from a predictive tool into a **hypothesis-generating mechanism**, empowering researchers to interpret, validate, and optimize molecular features with direct medicinal relevance.

4.4 Comparison with Literature

Previous QSAR and docking studies on SGLT2 inhibitors have reported prediction accuracies ranging from 0.80 to 0.87 (Nandha et al., 2021; Zhang et al., 2019). The results achieved in this study surpass these benchmarks, primarily due to the incorporation of **ensemble gradient boosting models** and **binary circular fingerprints** that capture richer structural detail.

Moreover, while earlier works relied on physicochemical descriptors or 2D autocorrelation features, this study's reliance on ECFP4 fingerprints yielded a more generalized and reproducible molecular representation.
The findings thus align with current trends in AI-driven cheminformatics, reinforcing the paradigm shift toward **data-centric, model-agnostic** frameworks in drug discovery.

4.5 Limitations

Despite its strong predictive capability, the study has certain limitations that warrant consideration:

Dataset Size: The dataset used, though balanced, was relatively small. Larger, more diverse datasets could enhance model generalizability across broader chemical spaces.
Feature Scope: Only topological (2D) fingerprints were used; inclusion of **3D conformational descriptors** or **quantum chemical properties** could improve biological relevance.
Experimental Validation: Predictions were not experimentally verified through *in vitro* assays or docking simulations, limiting the biological interpretability of model outcomes.
Hyperparameter Optimization: Hyperparameters were adjusted minimally; comprehensive optimization or Bayesian tuning might further improve performance consistency.

4.6 Future Directions

Building on this foundational framework, several extensions can be explored:

Integration of **molecular docking** and **dynamics simulations** to correlate machine learning predictions with binding energetics.
Development of **multi-target QSAR models** for screening compounds across multiple glucose transporter isoforms (e.g., SGLT1 and SGLT2).
Implementation of **deep learning architectures**, such as graph neural networks (GNNs), to learn directly from molecular graphs without handcrafted fingerprints.

Expansion of **explainability tools**, including counterfactual analysis, to guide rational molecule modification.

Such advancements would enable a comprehensive AI-driven drug discovery platform capable of predicting activity, elucidating mechanisms, and guiding synthesis simultaneously.

4.7 Summary

This discussion highlights that the integration of machine learning and explainable AI presents a powerful framework for understanding and predicting SGLT2 inhibitor activity. Beyond achieving high predictive performance, the inclusion of interpretability ensures that computational predictions maintain biological meaning, ultimately contributing to a more efficient, data-driven drug discovery process.

5. Conclusion and Recommendations

5.1 Conclusion

This research successfully demonstrated the application of advanced machine learning algorithms to predict the inhibitory potential of molecular compounds targeting the **Sodium–Glucose Cotransporter 2 (SGLT2)** a pivotal protein in glucose reabsorption and diabetic control. Using molecular fingerprints derived from the dataset, five classifiers, Random Forest, Gradient Boosting, XGBoost, CatBoost, and Support Vector Classifier were trained, evaluated, and compared.

All models exhibited robust predictive performance, with accuracies ranging between **0.8654 and 0.8846** and **ROC–AUC values exceeding 0.96**. Among these, the **XGBoost** and **SVC** models achieved the highest accuracy (0.8846), recall (0.9130), and F1-score (0.8750), signifying their superior ability to correctly identify active inhibitors without significant false negatives. The **ROC–AUC values**, particularly for XGBoost (0.9799), further confirmed the models' exceptional discriminatory capacity between active and inactive compounds.

These outcomes not only establish the computational feasibility of ensemble learning methods for QSAR modeling but also reveal that molecular structure–activity relationships can be effectively decoded through **explainable AI (XAI)** tools. In this work, **SHAP analysis** illuminated how specific structural fragments influence inhibitory activity, linking machine learning outputs with biologically interpretable molecular features.

From a **biomedical standpoint**, this signifies that SGLT2 inhibition, central to glucose regulation in type 2 diabetes can be computationally approximated with high fidelity, accelerating drug development while reducing laboratory costs and time. The synergy between **AI interpretability** and **chemical domain knowledge** thus represents a transformative approach to rational drug design.

5.2 Recommendations

Based on the findings, the following recommendations are proposed for both academic and industrial research directions:

Integration with Docking and Molecular Dynamics Simulations
Future work should couple QSAR-based prediction with molecular docking or dynamics analyses to validate the binding affinity and stability of top-predicted compounds. This hybrid framework would provide both predictive accuracy and mechanistic insights into ligand–receptor interactions.

Expansion of Dataset and Feature Diversity
Expanding the dataset to include a larger range of structurally diverse compounds and incorporating **3D descriptors**, **electronic features**, and **quantum chemical parameters** could enhance generalization and predictive interpretability.

Adoption of Deep Learning Frameworks
Emerging models such as **Graph Neural Networks (GNNs)** and **Transformer-based molecular encoders** can learn hierarchical chemical representations directly from molecular graphs, potentially surpassing fingerprint-based models in performance and scalability.

Explainable AI for Regulatory Acceptance
To enhance trust in AI-assisted drug discovery, interpretability tools like SHAP, LIME, and

counterfactual explanations should be standardized in QSAR pipelines. This aligns with regulatory frameworks seeking transparency in AI-based pharmacological assessments.

Clinical Translation and Experimental Validation

Compounds predicted as strong inhibitors should undergo **in vitro screening** and **ADMET profiling** to confirm their pharmacological efficacy and safety. Translating computational predictions into preclinical stages is essential for therapeutic validation.

Collaborative Biomedical-AI Frameworks

Establishing multidisciplinary teams that unite **data scientists, chemoinformaticians, and pharmacologists** will ensure that AI models remain biologically grounded and clinically relevant. Such collaborations can lead to more precise and ethically sound computational drug discovery processes.

5.3 Closing Remark

In conclusion, this work reinforces the idea that **machine learning and biomedical science are not parallel disciplines but convergent forces**. The fusion of explainable AI with pharmacological modeling fosters a paradigm where data-driven algorithms not only predict but also *explain* biological outcomes. This synthesis paves the way for a new generation of drug discovery, one that is intelligent, interpretable, and human-aligned.

6. References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 1–2,* 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Breiman, L. (2001). *Random forests. Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chirkov, Y., Gaulton, A., & Overington, J. P. (2022). *AI in drug discovery: From hype to clinical reality. Drug Discovery Today, 27*(6), 1780–1792. https://doi.org/10.1016/j.drudis.2022.03.001

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., ... & Leach, A. R. (2017). *The ChEMBL database in 2017. Nucleic Acids Research, 45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 4768–4777).

Nandha, B., Baluja, S., & Bansal, Y. (2021). *QSAR modeling and molecular docking of novel SGLT2 inhibitors for type 2 diabetes mellitus. Computational Biology and Chemistry, 92,* 107478. https://doi.org/10.1016/j.compbiolchem.2021.107478

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features.* In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)* (pp. 6639–6649).

Rieg, T., & Vallon, V. (2018). *Development of SGLT1 and SGLT2 inhibitors. Diabetologia, 61*(10), 2079–2086. https://doi.org/10.1007/s00125-018-4667-1

Rogers, D., & Hahn, M. (2010). *Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50*(5), 742–754. https://doi.org/10.1021/ci100050t

Schneider, G. (2018). *Automating drug discovery. Nature Reviews Drug Discovery, 17*(2), 97–113. https://doi.org/10.1038/nrd.2017.232

Vapnik, V. (1995). *The nature of statistical learning theory.* Springer. https://doi.org/10.1007/978-1-4757-2440-0

Zhang, L., Li, Y., & Zhang, J. (2019). *QSAR and molecular docking-based virtual screening of SGLT2 inhibitors. Journal of Molecular Graphics and Modelling, 93,* 107432. https://doi.org/10.1016/j.jmgm.2019.107432