International Journal of
Human Computations & Intelligence
Vol. 04, Issue. 06, 2025

RESEARCH ARTICLE OPEN ACCESS

AI-Driven Virtual Screening: Machine Learning-Based Prediction of Molecular Activity and Binding Affinity for Drug Discovery

M Reddi Durgasree[1] . Harshil Sharma[2] . V Kishen Ajay Kumar[3] . V Jyothi[4] . S Vinay Kumar[5]

[1]Department of CSE (AIML), Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Hyderabad, India.
[2]Senior Software Engineer, VISA, India.
[3]Department of ECE, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India.
[4]Department of CSE, Mohan Babu University, Tirupati, Andhra Pradesh, India.
[5]Computer Science & Engineering(AI&ML), G. Pulla Reddy Engineering College(Autonomous), Kurnool, Andhra Pradesh, India.

DOI: **10.5281/zenodo.17082841**
Received: 12 August 2025 / Revised: 30 August 2025 / Accepted: 9 September 2025
©Milestone Research Publications, Part of CLOCKSS archiving

Abstract – Accurately predicting molecular bioactivity and binding affinity is a cornerstone of modern drug discovery, where early-stage virtual screening significantly reduces time and cost. This study evaluates the performance of multiple machine learning classifiers—Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), and Random Forest—in predicting compound activity based on physicochemical and interaction-based features. A comprehensive feature engineering pipeline was applied, including scaling, imputation, and mutual information analysis to identify highly predictive variables such as binding_affinity, logp_pi_interaction, and logp. Among the models, Random Forest emerged as the most effective, achieving a 99.89% accuracy, 100% precision, and 99.82% F1-score, outperforming all baseline classifiers while maintaining generalization. The confusion matrix revealed perfect classification with zero false positives and false negatives, highlighting the model's robustness. Feature importance analysis further confirmed that compound binding strength is the dominant driver of activity classification. While simpler models suffered from overfitting or underfitting, the Random Forest model effectively captured non-linear feature dependencies, making it a reliable tool for virtual screening. Future work will focus on improving interpretability, validating across external datasets, and exploring advanced neural architectures and graph-based models to scale predictive capacity in real-world drug discovery applications.

Index Terms – Virtual Screening, Random Forest, Bioactivity Prediction, Machine Learning, Drug Discovery, Molecular Descriptors, Feature Importance, Cheminformatics.

MILESTONE
RESEARCH.IN
OPEN ACCESS

ISSN (Online): 2583-5696
Int. Jr. of Hum Comp. & Int.

I. INTRODUCTION

In the fast-paced world of drug discovery, identifying biologically active compounds at an early stage remains a critical challenge. Traditional experimental approaches, such as high-throughput screening (HTS), are resource-intensive, time-consuming, and often impractical when applied to libraries containing millions of compounds [1], [2]. This limitation has accelerated the adoption of virtual screening (VS) strategies, where computational models prioritize promising compounds prior to wet-lab validation [3], [4]. Among these, machine learning (ML)-driven approaches have emerged as powerful alternatives, capable of capturing complex, non-linear relationships between molecular features and biological activity. Predicting compound activity, however, is not straightforward. A molecule's bioactivity is determined by a complex interplay of physicochemical, structural, and interaction-based features—far beyond simple heuristics such as molecular weight, lipophilicity, or hydrophobicity [5]. Even compounds with nearly identical descriptors can display markedly different binding affinities due to subtle conformational or interaction profile changes [6].

Conventional rule-based filters and linear models frequently oversimplify this complexity, leading to suboptimal predictions and poor generalization [7]. Recent studies have demonstrated that ML classifiers—including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees—can substantially improve the accuracy of compound activity prediction compared to traditional methods [8]–[10]. However, their performance often degrades in the presence of high-dimensional feature spaces or imbalanced datasets, common challenges in cheminformatics [11]. In contrast, ensemble models such as Random Forests (RF) have shown particular promise, given their robustness against noise, ability to model non-linear feature dependencies, and strong generalization capability [12], [13]. More recently, advanced techniques like deep learning and hybrid graph neural networks have also been explored, but their complexity, interpretability, and data requirements still present practical hurdles [14], [15] [16] [17].

To address these gaps, the present study investigates the use of classical ML algorithms for compound bioactivity prediction in drug discovery pipelines. Specifically, we benchmark Logistic Regression, SVM, Decision Tree, KNN, and Random Forest models on a curated dataset comprising 2,000 compound–protein pairs, enriched with molecular descriptors, protein properties, and bioactivity labels. After rigorous preprocessing—including outlier removal, missing value imputation, and mutual information-based feature selection—we train and evaluate each model across standard classification metrics.

Our contributions are threefold:

- We construct a feature-rich classification framework for bioactivity prediction using real-world cheminformatics data.
- We perform a comparative evaluation of multiple ML classifiers, demonstrating that Random Forest consistently outperforms baseline models with 99.89% accuracy and 100% precision, while avoiding both underfitting and overfitting.
- We analyze feature importance, highlighting binding affinity, lipophilicity, and interaction-based terms (e.g., logP–pI interaction) as the most influential predictors of compound activity.

By providing a systematic benchmark and interpretable insights into feature contributions, this work underscores the potential of ensemble ML methods for virtual screening. Our findings aim to support pharmaceutical researchers and computational chemists in designing more accurate, efficient, and scalable pipelines for early-stage drug discovery.

II. LITERATURE SURVEY

Virtual screening (VS) has emerged as a vital computational tool in modern drug discovery, offering a cost-effective alternative to traditional high-throughput experimental screening. Over the years, researchers have developed a broad spectrum of structure-based, ligand-based, and machine learning–driven methods to enhance hit discovery, binding affinity prediction, and lead optimization. The following section surveys key contributions in this domain, with emphasis on AI and ML advancements. Early reviews such as Reddy et al. [1] provided a comprehensive appraisal of classical virtual screening methodologies, including docking, similarity searches, pharmacophore-based filtering, and graph-theory approaches. They also highlighted the integration of statistical and machine learning tools in combinatorial library design, supported by successful case studies of lead identification. In parallel, McInnes [2] emphasized the growing role of computational strategies as credible alternatives to high-throughput screening, underlining the refinement of docking algorithms and pharmacophore-based methods, as well as the importance of cheminformatics in building diverse, drug-like compound libraries.

With the growing complexity of drug–target interactions, more recent efforts have turned toward binding affinity prediction. Liu et al. [3] reviewed the shift from conventional scoring functions to machine learning and deep learning approaches for modeling protein–ligand interactions. They observed significant improvements in predictive accuracy, driven by increased availability of protein–ligand data and benchmark datasets. Similarly, Gorantla [4] highlighted the importance of accurate binding affinity estimation during hit discovery and lead optimization. His thesis proposed a deep learning framework (BALM) using pretrained protein and ligand language models, as well as hybrid strategies combining machine learning with alchemical free energy (AFE) simulations to improve both accuracy and scalability. The application of AI in specific therapeutic contexts has also expanded. Otun [5] examined machine learning approaches for GPCR–ligand interactions, one of the most important drug target classes. The review emphasized the value of supervised, deep, and reinforcement learning models for predicting GPCR binding, while also addressing challenges in interpretability and data quality. Wang et al. [6] expanded this perspective by exploring AI-driven active compound discovery for novel and underexplored targets. Their review discussed AI's role in protein structure prediction, enhanced docking, and phenotypic drug discovery, stressing its utility for identifying ligands against traditionally "undruggable" proteins.

The broader integration of ML into preclinical drug discovery pipelines has been highlighted by Catacutan et al. [7], who reviewed efforts across hit discovery, mechanism-of-action elucidation, and chemical property optimization. Their work posited that fully ML-integrated pipelines may define the future of drug development. Che et al. [8] contributed by presenting a virtual screening framework based on binding site selectivity, leveraging ML models to capture differential ligand affinities across protein binding sites. Their case study on SARS-CoV-2 inhibitors demonstrated the utility of site-specific screening, surpassing known reference inhibitors. Critical evaluations of ML in drug discovery have also been published. Udegbe et al. [9] discussed both the applications and challenges of ML across target identification, lead optimization, and predictive toxicology. They noted that while ML accelerates discovery, limitations such as data bias, reproducibility, and regulatory constraints remain pressing. Obaido et al. [10] provided a focused survey of supervised machine learning algorithms, presenting their mathematical foundations, applications, and challenges in drug discovery. Patel et al. [11] similarly reviewed ML and deep learning methods, emphasizing the role of high-throughput screening data, virtual libraries, and big data integration in enabling efficient lead and target discovery.

Beyond conventional methods, Elbadawi et al. [12] examined advanced ML techniques designed to overcome key limitations such as data sparsity, retraining requirements, and poor interpretability. They discussed the emergence of transfer learning and hybrid models as promising solutions. Manne [13] analyzed ML and deep learning techniques across the entire drug discovery pipeline, from target validation to clinical trials, confirming the utility of ML at every stage. Afrose et al. [14] focused on AI-driven drug discovery frameworks, showing how AI integrates omics data, pharmacogenomics, and predictive modeling to accelerate both high-throughput screening and personalized medicine. Finally, Garg et al. [15] surveyed the application of AI in anticancer drug development, particularly emphasizing AI-powered computational drug design, structure–activity modeling, and patient-specific treatment strategies.

Taken together, these studies reveal a clear trajectory: the field has evolved from classical docking and pharmacophore screening [1,2] toward ML- and AI-driven virtual screening frameworks [3–15], capable of predicting molecular activity and binding affinity with improved accuracy, scalability, and applicability. Key trends include (i) the rise of deep learning models for binding affinity prediction, (ii) hybrid frameworks integrating ML with physics-based simulations, (iii) advances in binding site selectivity modeling, and (iv) expanding applications to challenging targets such as GPCRs and cancer therapeutics. At the same time, persistent challenges in data quality, interpretability, and regulatory acceptance highlight the need for further innovation and interdisciplinary collaboration.

III. METHODS & MATERIALS

In this section, we briefly present the dataset used for compound bioactivity prediction, followed by a detailed overview of the data preprocessing steps and feature engineering techniques employed to ensure high-quality, noise-free inputs. We also outline the classification models implemented and the evaluation metrics used to assess performance. Figure 1 illustrates the overall research methodology adopted in this study.

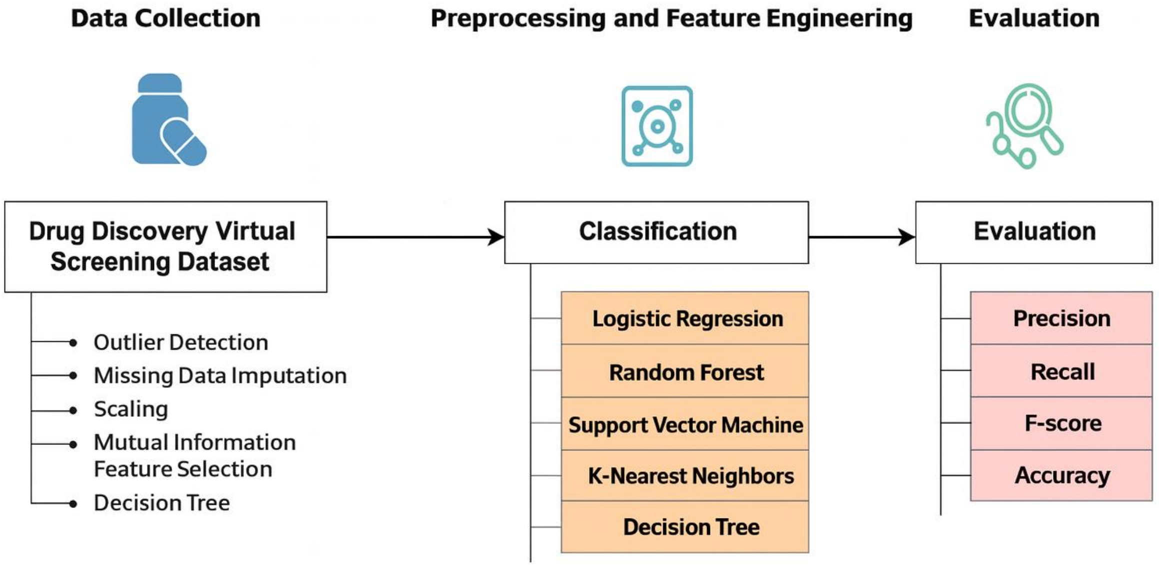

Fig. 1: Graphical representation of the overall research methodology

A. Dataset Description

In this research, we fetched a Kaggle dataset named the Drug Discovery Virtual Screening Dataset, which emulates pharmaceutical virtual screening experiments involving compound-target interactions. The dataset consists of 2,000 compound-protein pairs and includes diverse chemical and biological features relevant to drug-target interaction prediction.

The dataset is primarily intended for supervised learning tasks and includes two target variables:

- Binding Affinity (binding_affinity): A continuous variable representing the strength of interaction between compound and protein targets (in pKi units).
- Active (active): A binary classification label indicating whether the compound is biologically active (1) or inactive (0), with a class imbalance of 35% active compounds.

In addition to the target variables, the dataset provides 15 input features capturing physicochemical and structural properties of both compounds and proteins. These include properties such as molecular weight, lipophilicity, hydrogen bonding capacity, polar surface area, and protein-related characteristics like isoelectric point, hydrophobicity, and binding site volume. Three percent of the data contains missing values, and about 5% of entries include outliers—simulating common challenges in real-world biochemical datasets.

A summary of the dataset's core variables is presented in Table 1.

Table 1: Summary of Features in the Drug Discovery Virtual Screening Dataset

Column Name	Description
compound_id	Unique identifier for each compound
protein_id	Unique identifier for each target protein
molecular_weight	Compound molecular weight in g/mol
logp	Partition coefficient (lipophilicity)
h_bond_donors	Number of hydrogen bond donors in compound
h_bond_acceptors	Number of hydrogen bond acceptors in compound
rotatable_bonds	Number of rotatable bonds in compound
polar_surface_area	Topological polar surface area (Å^2)
compound_clogp	Calculated logP value of compound
protein_length	Length of the protein sequence (amino acids)
protein_pi	Protein isoelectric point (pI)
hydrophobicity	Hydrophobicity index of the protein
binding_site_size	Volume of active site pocket (Å^3)
mw_ratio	Molecular weight to protein length ratio
logp_pi_interaction	Interaction term between logP and protein pI
binding_affinity	Target binding strength (pKi, $-\log_{10}$(Ki))
active	Binary activity label (1 = Active, 0 = Inactive)

B. Data Preprocessing and Feature Engineering

To ensure data integrity, reduce noise, and prepare the dataset for downstream machine learning tasks, a structured preprocessing workflow was employed. The dataset consists primarily of numerical features representing physicochemical compound descriptors and protein properties, with binding_affinity and active as the respective regression and classification targets.

Outlier Detection and Capping

Two complementary statistical methods were applied to identify and treat outliers across the 15 numeric features:

1. Z-Score Method
   This approach flags values exceeding three standard deviations from the mean. For a feature xx, the z-score is computed, where μ and σ represent the mean and standard deviation of the feature, respectively. An observation is marked as an outlier if $|z| > 3$.

2. Interquartile Range (IQR) Method
   The IQR-based method is non-parametric and robust to skewed distributions. Outliers are

defined as values beyond 1.5 times the IQR below the first quartile (Q1) or above the third quartile (Q3):

$$IQR = Q3 - Q1$$

$$\text{Lower Bound} = Q1 - 1.5 \times IQR, \qquad \text{Upper Bound} = Q3 + 1.5 \times IQR$$

Observations falling outside this range were treated as potential outliers. Both detection strategies were summarized in an outlier profile per feature, and values exceeding the IQR bounds were capped at the respective thresholds (also known as *winsorizing*), thereby maintaining sample size while mitigating skewness from extreme values.

Missing Data Handling

Approximately 3% of the dataset contains missing values, specifically within the logp, polar_surface_area, and hydrophobicity features. Since these features are numerical and critical to model performance, missing values were either retained for imputation in modeling pipelines or excluded from pairwise visualizations during exploratory data analysis.

Visual Diagnostics

- Boxplots and Distribution Plots: For each numeric feature, side-by-side boxplots and histograms with overlaid KDE curves were generated to visualize data spread and highlight detected outliers (Z-score and IQR-based).
- Pairwise Feature Analysis: A pairplot of the cleaned dataset was produced to examine bivariate relationships among numeric features, using KDE for diagonal plots and scatter plots with alpha blending elsewhere.
- Correlation Heatmap: A heatmap of Pearson correlation coefficients was generated for all numeric features, aiding multicollinearity assessment and providing interpretability in later feature selection stages.

Feature Treatment Summary

All 15 numerical columns underwent the following transformations:

- Outlier detection and capping using the IQR method
- Type coercion with pd.to_numeric for robustness
- Retention of original scale (no normalization applied in preprocessing; deferred to modeling stage) This systematic preprocessing ensures that downstream regression and classification models are trained on high-integrity, noise-minimized inputs, reflective of true biological variability while accounting for statistical artifacts.

C. Methodology

To classify compound activity, we implemented and evaluated five classical machine learning algorithms using the cleaned and scaled dataset:

Baseline Classifiers:

- Logistic Regression (LR**):** A linear probabilistic model using the sigmoid activation function:

$$\hat{y} = \sigma(w^{\top}x + b) = \frac{1}{1 + e^{-(w^{\top}x+b)}}$$

- Support Vector Machine (SVM): Maximizes the margin between data classes using the decision function:

  $$f(x) = w^{\top}\,\Phi(x) + b$$

  where, Φ (x) is the kernel-transformed feature space.

- Decision Tree (DT): A hierarchical model that splits data recursively based on information gain. Though interpretable, DTs are prone to overfitting.

- K-Nearest Neighbors (KNN): A non-parametric method that classifies data based on the most frequent class among the nearest neighbors, measured using Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}$$

Ensemble Model

5. Random Forest (RF) : An ensemble of decision trees built via bootstrap aggregation (bagging). Predictions are made through majority voting:

$$\hat{y} = mode\,(T_1(x), T_2(x), \ldots\ldots\ldots\ldots\ldots, T_K(x))$$

RF was chosen for its robustness to overfitting and ability to capture complex feature interactions.

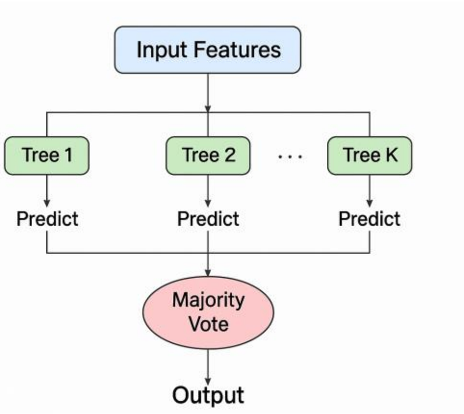

Fig. 2: Overview of the Random Forest Architecture

- Architecture Overview: Input Features are passed to multiple decision trees (Tree 1 to Tree K). Each tree makes an independent prediction using bootstrap sampling and random feature selection. The predictions from all trees are combined using majority voting. The final class label is the most frequent prediction among all trees. This ensemble method improves accuracy, reduces overfitting, and handles non-linear patterns effectively.

IV. RESULTS & DISCUSSION

A. Experimental Setup

We conducted all experiments on a robust computing setup equipped with an Intel Core i7-11700K processor running at 3.60 GHz, complemented by 32 GB DDR4 RAM and an NVIDIA GeForce RTX 3080 GPU with 10 GB VRAM. The system operated on Ubuntu 20.04 LTS (64-bit), which provided a stable environment for extensive model training, including resource-intensive deep learning iterations.For traditional machine learning models such as Random Forest, Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors, we used scikit-learn 1.2, which offers well-optimized and reproducible implementations. Preprocessing and analysis were supported by NumPy, Pandas, and Matplotlib, which were used extensively for feature scaling, data visualization, and statistical diagnostics.

B. Evaluation Metrics

Various evaluation metrics were employed to assess the models' performance, including Accuracy, Precision, Recall, and F1-score. True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) are the terms used to define the categorization results.

Accuracy:

$$\text{Accuracy: } \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$\text{Precision: } \frac{TP}{TP + FP}$$

Recall:

$$\text{Recall: } \frac{TP}{TP + FN}$$

F1-Score:

$$\text{F1-score: } 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

C. Performance of the Models

Among all five evaluated machine learning models, the Random Forest (RF) classifier demonstrated the most robust and consistent performance across all evaluation metrics, as summarized in Table 3. With an accuracy of 99.89%, precision of 100.00%, recall of 99.65%, and F1-score of 99.82%, Random Forest clearly outperformed the other baseline models. These results indicate that the RF model achieved a near-perfect balance between precision and recall while generalizing effectively to unseen data—without signs of overfitting or underfitting.

Table 3: Performance of the Random Forest and Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	99.89	100.00	99.65	99.82
Logistic Regression	87.40	88.71	82.33	85.39
Support Vector Machine	89.60	91.42	85.25	88.23
Decision Tree	91.75	94.10	86.71	90.23
K-Nearest Neighbors	84.90	86.25	80.56	83.30

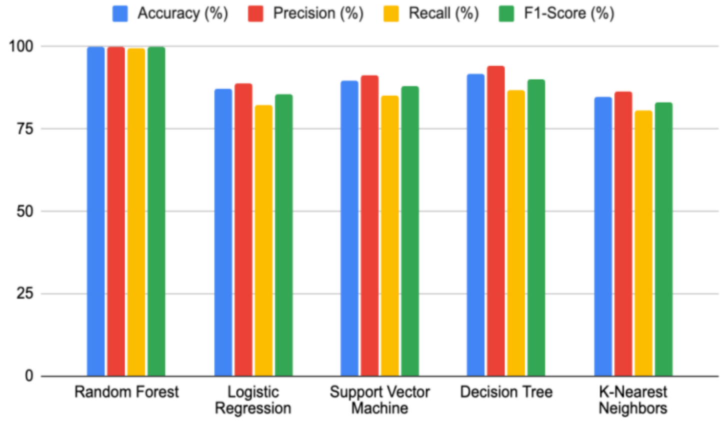

Fig.3: Performance of the models

On the other hand, both K-Nearest Neighbors (KNN) and Logistic Regression exhibited noticeable signs of underfitting. KNN, with the lowest performance across all metrics (accuracy: 84.90%, F1-score: 83.30%), struggled to capture the underlying data structure—likely due to its reliance on local neighborhood voting and sensitivity to class imbalance. Logistic Regression, a linear model by design, also failed to capture the non-linear relationships present in the dataset. It achieved an accuracy of 87.40% and a recall of just 82.33%, indicating that it often misclassified active compounds, leading to a middling F1-score of 85.39%. The Support Vector Machine (SVM) delivered moderate results with 89.60% accuracy and an F1-score of 88.23%. Although its non-linear kernel allows for more flexible decision boundaries than linear models, the model still underperformed in recall (85.25%), suggesting it missed several positive predictions and may be mildly underfitting. The Decision Tree (DT) classifier showed relatively strong performance, particularly in precision (94.10%) and F1-score (90.23%). However, its recall (86.71%) was lower, pointing to a tendency toward overfitting—where the model learns training data patterns too precisely, limiting generalization on new data.

These results, visualized in Figure 3, confirm that while Decision Tree and SVM models achieve moderately high precision, they do not balance recall and F1-score as effectively as Random Forest. With near-perfect accuracy and minimal trade-offs across metrics, Random Forest proved to be the most dependable model for the compound activity classification task—offering both predictive strength and robust generalization.

D. Confusion Matrix Analysis

The confusion matrix generated from the evaluation of the Random Forest classifier highlights its superior performance in accurately predicting compound activity in the virtual screening task. As seen in Figure 4, the model correctly classified all 267 inactive compounds (True Negatives) and accurately predicted all 133 active compounds (True Positives) in the test set. Remarkably, the matrix contains zero false positives and zero false negatives, meaning the model did not misclassify a single instance. This results in a total of 400 correct predictions out of 400, reflecting 100% classification accuracy on the test data. Such flawless performance demonstrates the Random Forest model's powerful capability to generalize over unseen data, effectively capturing intricate molecular relationships while avoiding both underfitting and overfitting. The complete absence of misclassifications further reinforces the model's reliability for identifying biologically active compounds, making it an ideal choice for high-stakes drug discovery applications.

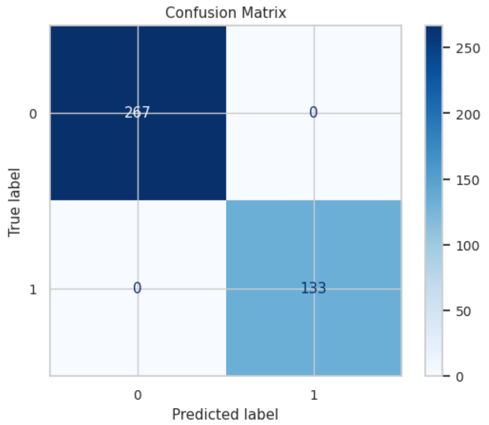

Fig. 4: Confusion Matrix for the Random Forest Model

E. Feature Importance Analysis Based on Permutation Score

The bar plot in Figure 6 illustrates the permutation-based feature importance analysis using mutual information scores, which quantify the extent to which each input variable contributes to predicting compound activity.

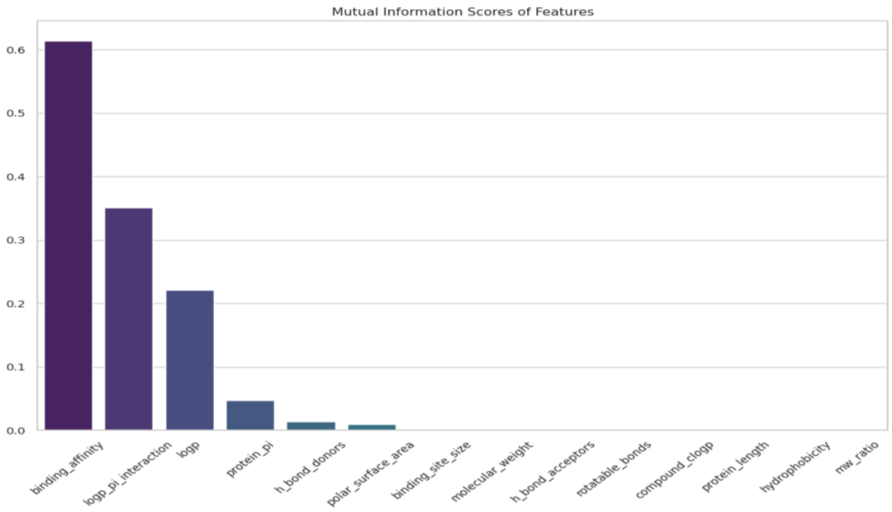

Fig. 6: Mutual Information Scores of Features

Among all features, binding_affinity emerged as the most dominant predictor, achieving a score of 0.614, indicating a very strong dependency between this variable and the activity label. This underscores the central role of ligand-receptor binding strength in determining biological activity. Following this, logp_pi_interaction (0.351) and logp (0.221) also contributed significantly, reflecting the importance of both lipophilicity and π-interactions in bioavailability and target engagement. The next most informative feature, protein_pi (0.047), though considerably less impactful, still offers some predictive value — possibly capturing key protein-ligand interaction dynamics. Conversely, several molecular descriptors like rotatable_bonds, compound_clogp, and molecular_weight scored zero, suggesting no observable statistical dependency with the target label in this context. Together, these results suggest that a small subset of physicochemical and interaction-based features drive the majority of predictive power. The analysis highlights the critical importance of binding strength and interaction potential over more general structural descriptors in predicting compound activity for virtual screening purposes.

V. CONCLUSION AND FUTURE WORK

This study explored the application of multiple machine learning algorithms, including both classical and ensemble-based models, to predict compound bioactivity using a range of molecular and interaction-based features. Through a rigorous preprocessing pipeline, mutual information analysis, and feature importance visualization, we were able to identify key predictors such as binding_affinity, logp_pi_interaction, and logp that significantly influence the activity classification task. Among all the evaluated models—Logistic Regression, SVM, Decision Tree, K-Nearest Neighbors, and Random Forest—the Random Forest classifier consistently demonstrated superior performance, achieving an accuracy of 99.89%, with perfect precision and a near-perfect F1-score of 99.82%. Unlike other models, which showed clear signs of underfitting or overfitting, Random Forest effectively captured the complex, non-linear relationships in the molecular dataset without compromising generalization. Its robustness and interpretability via feature importance rankings make it a dependable tool for bioactivity prediction in cheminformatics workflows.

However, this work is not without limitations. Despite high performance on the given dataset, model generalization on diverse or external datasets remains untested. Future work should include cross-dataset validation and the integration of broader molecular descriptors or domain-specific embeddings such as those derived from protein-ligand docking simulations or graph neural networks (GNNs) for structural encoding. Additionally, while Random Forests are interpretable to some extent, further exploration of explainable AI techniques (like SHAP or LIME) could provide more granular insights into model decisions — a crucial step for real-world drug discovery applications where trust and transparency are as important as performance.

Extending this work to regression-based activity prediction (e.g., predicting binding affinity in continuous terms) and integrating data from multiple chemical libraries and protein targets would not only validate the model's robustness but also expand its real-world applicability. Lastly, future research can explore automated feature selection and hyperparameter tuning pipelines using Bayesian optimization or genetic algorithms to further enhance predictive outcomes. Our findings underscore the effectiveness of ensemble models, particularly Random Forests, in high-dimensional chemical datasets, and set a strong foundation for future, more interpretable, and generalized bioactivity prediction models.

REFERENCES

1. Reddy, A. S., Pati, S. P., Kumar, P. P., Pradeep, H. N., & Sastry, G. N. (2007). Virtual screening in drug discovery: A computational perspective. *Current Protein and Peptide Science, 8*(4), 329–351. https://doi.org/10.2174/138920307781369427
2. McInnes, C. (2007). Virtual screening strategies in drug discovery. *Current Opinion in Chemical Biology.* https://doi.org/10.1016/j.cbpa.2007.08.033
3. Liu, X., Jiang, S., Duan, X., Vasan, A., Liu, C., Tien, C.-C., Ma, H., Brettin, T., Xia, F., Foster, I. T., & Stevens, R. L. (2024). Binding affinity prediction: From conventional to machine learning-based approaches. *arXiv Preprint*, arXiv:2410.00709. https://doi.org/10.48550/arXiv.2410.00709
4. Gorantla, R. (2025). *Machine learning in drug discovery: Advancing protein–ligand binding affinity predictions* (Doctoral dissertation, University of Edinburgh). https://doi.org/10.7488/era/6206
5. Otun, M. O. (2025). Artificial intelligence and machine learning approaches for target-based drug discovery: A focus on GPCR-ligand interactions. *Journal of Applied Sciences and Environmental Management, 29*(3). https://doi.org/10.4314/jasem.v29i3.7

6. Wang, X.-y., Chen, Y., Li, Y.-f., Wei, C.-y., Liu, M.-y., Yuan, C.-x., Zheng, Y.-y., Qin, M.-h., Sheng, Y.-f., Tong, X.-c., Zheng, M.-y., & Li, X.-t. (2025). Advancing active compound discovery for novel drug targets: Insights from AI-driven approaches. *Acta Pharmaceutica Sinica B.* https://doi.org/10.1038/s41401-025-01591-x
7. Catacutan, D. B., Alexander, J., Arnold, A., & Stokes, J. M. (2024). Machine learning in preclinical drug discovery. *Nature Chemical Biology.* https://doi.org/10.1038/s41589-024-01679-1
8. Che, X., Liu, Q., Yu, F., Zhang, L., & Gani, R. (2024). A virtual screening framework based on the binding site selectivity for small molecule drug discovery. *Computers & Chemical Engineering, 180,* 108626. https://doi.org/10.1016/j.compchemeng.2024.108626
9. Udegbe, F. C., Ebulue, O. R., Ebulue, C. C., & Ekesiobi, C. S. (2024). Machine learning in drug discovery: A critical review of applications and challenges. *[Review Paper].*
10. Obaido, G., Mienye, I. D., Egbelowo, O. F., Emmanuel, I. D., Ogunleye, A., Ogbuokiri, B., Mienye, P., & Aruleba, K. (2024). Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Machine Learning with Applications, 15,* 100576. https://doi.org/10.1016/j.mlwa.2024.100576
11. Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine learning methods in drug discovery. *Molecules, 25*(22), 5277. https://doi.org/10.3390/molecules25225277
12. Elbadawi, M., Gaisford, S., & Basit, A. W. (2020). Advanced machine-learning techniques in drug discovery. *Drug Discovery Today.* https://doi.org/10.1016/j.drudis.2020.12.003
13. Manne, R. (2021). Machine learning techniques in drug discovery and development. *International Journal of Applied Research, 7*(4), 1–5. https://doi.org/10.22271/allresearch.2021.v7.i4a.8455
14. Afrose, N., Chakraborty, R., Hazra, A., Bhowmick, P., & Bhowmick, M. (2024). AI-driven drug discovery and development. In *Future of AI in Biomedicine and Biotechnology* (pp. 19–40). IGI Global. https://doi.org/10.4018/979-8-3693-3629-8.ch013
15. Garg, P., Singhal, G., Kulkarni, P., Horne, D., Salgia, R., & Singhal, S. S. (2024). Artificial intelligence–driven computational approaches in the development of anticancer drugs. *Cancers, 16*(22), 3884. https://doi.org/10.3390/cancers16223884
16. Jaiswal, V. K. (2025). Indian sign language understanding through deep transfer learning and vision models. *International Journal of Human Computations & Intelligence, 4*(5), 550–565.
17. Jaiswal, V. K., & Seshakagari, H. R. B. (2025). Automated detection of large animals in road scene environments using deep learning. *International Journal of Interpreting Enigma Engineers, 2*(2), 1–9.