scientific reports

OPEN

# ACLPred: an explainable machine learning and tree-based ensemble model for anticancer ligand prediction

Arvind Kumar Yadav & Jun-Mo Kim

**Several small molecules have been approved for cancer treatment, but the continuously growing cancer cases have further encouraged the identification of new anticancer drug compounds. Experimental methods are costly and time-consuming, thus rapid and cost-effective alternative method is much required. The effective identification of anticancer compounds using machine learning (ML) offers a promising solution, reducing both time and cost. In this study, small molecules with known inhibitory activities, both anticancer and non-anticancer were considered to train classification models. Molecular descriptors were calculated, and multistep feature selection was applied to identify significant features. Multiple ML algorithms were employed to build classification models and evaluated their performance using independent test and external datasets. The tree-based ensemble model, particularly Light Gradient Boosting Machine (LGBM), achieved the highest prediction accuracy of 90.33%, with an area under the receiver operating characteristic curve (AUROC) of 97.31%. Consequently, LGBM model was implemented in our proposed method, ACLPred. The ACLPred demonstrated superior prediction accuracy with good generalizability compared to existing methods. SHapley Additive exPlanations (SHAP) analysis provided model interpretability and revealed that topological features made major contributions to decision-making. ACLPred is available as an open-source, user-friendly graphical interface at https://github.com/ArvindYadav7/ACLPred for the screening of potential anticancer compounds.**

**Keywords** Anticancer ligand, Cancer, Ensemble machine learning, Multistep feature selection

Cancer continues to be a major cause of morbidity and mortality worldwide, with an estimated 20 million new cases and approximately 10 million deaths annually[1,2]. The global cancer burden is rising because of factors, such as lifestyle changes, an aging population, environmental pollution, and increased exposure to carcinogens. Understanding cancer pathogenesis is difficult because of its complexity, which arises from tumor heterogeneity, epigenetic changes, genetic mutations, and dynamic interactions within the tumor microenvironment[3]. Despite advances in targeted therapy, precision medicine, and early detection, the development of novel anticancer drugs remains a high priority. Consequently, researchers have focused on identifying small anticancer molecules with the desired efficacy to enhance existing cancer therapies[4].

Small anticancer molecules are crucial for cancer treatment because they can precisely target cancer cells while minimizing toxicity to healthy tissue. These compounds are promising candidates for targeted therapy, as they frequently inhibit key proteins and signaling pathways involved in cancer progression[5]. Their small size facilitates improved cellular penetration, enhanced bioavailability, and easier chemical modification to optimize drug efficacy[6]. High-throughput screening of therapeutic molecules from large chemical libraries remains the most suitable method for identifying novel drug candidates[7]. However, conventional approaches, such as high-throughput screening and structure-based drug design, face limitations in efficiently identifying active compounds[8,9]. Furthermore, experimental methods for discovering anticancer molecules are expensive, time-consuming, and labor-intensive. Recent technological advancements have enabled the development of computational methods that streamline and accelerate the discovery of novel anticancer agents[10]. With the increasing availability of chemical libraries, biological datasets, and computational resources, artificial intelligence and machine learning (ML) have emerged as transformative tools in small-molecule cancer drug development[11,12]. Publicly available chemical compounds can be used to develop effective anticancer agents[13–15].

Functional Genomics & Bioinformatics Laboratory, Department of Animal Science and Technology, Chung-Ang University, Anseong 17546, Gyeonggi-do, Republic of Korea. ✉email: junmokim@cau.ac.kr

These techniques allow the rapid identification of small chemical compounds with high success rates in preclinical and clinical development.

In contrast to traditional methods, ML algorithms can learn from large chemical datasets and prioritize high-potential compounds with notable precision, significantly reducing both time and cost[16,17]. ML-based approaches have achieved considerable success in drug repurposing[18,19]identifying target-specific anticancer molecules[20,21]and accurately predicting compound toxicity[22,23]. Integrating chemical information with ML enables researchers to efficiently screen millions of small molecules and prioritize those with the greatest therapeutic potential. Computational methods have also been developed to predict the activity of small anticancer compounds[24,25]often using cancer cell line data to forecast tissue-specific responses. Li and Huang developed the web server CDRUG[26] to predict the anticancer activities of chemical compounds. It uses a weighted similarity score between query and active compounds and outperforms other baseline models, achieving an area under the curve of 0.87. Al-Jarf et al. introduced pdCSM[27]which utilizes graph-based signatures to predict anticancer properties of small molecules and achieved an area under the curve of 0.94 with 86% prediction accuracy. Recently, Balaji et al. proposed the ML-based method MLASM[28]which also screens small molecules for anticancer potential. This approach employed the Light Gradient Boosting Machine (LGBM) algorithm with molecular descriptors and achieved an accuracy of 79% on independent test data. However, despite these advancements, more robust and precise models are still needed to improve prediction accuracy and enhance anticancer compound screening.

In this study, we introduced an improved ML-based method for the prediction of anticancer small molecules using upgraded feature descriptors with a multistep feature selection strategy. After model optimization, we implemented an efficient LGBM-based anticancer ligand-prediction method called ACLPred. We followed the explainable ML technique to quantify the important descriptors that affect model prediction. The detailed construction workflow of ACLPred is shown in Fig. 1. Tenfold cross-validation and independent evaluation demonstrated that the proposed model achieved satisfactory performance for anticancer ligand identification. Moreover, it outperformed existing methods with improved prediction accuracy on independent test datasets. Thus, the successful implementation of this model will help improve the performance of ACLPred and offer a robust method for screening potential anticancer molecules from a large pool of compound databases.

## Materials and methods

### Data collection and processing

The selection of appropriate, fine, and accurately categorized datasets is crucial for the development of effective ML models. Here, we used the datasets curated by Balaji et al. for training and testing the MLASM method[28]retrieved from the PubChem BioAssay database[29]. After preprocessing, 5000 active and 5000 inactive anticancer small molecules were selected. A simplified molecular input line entry system (SMILES) was used to collect all of the molecules[30]. The Tanimoto coefficient ($T_c$)[31] was computed to measure the similarity among the molecules based on their fingerprints using the DataStructs Python package. It is defined as:

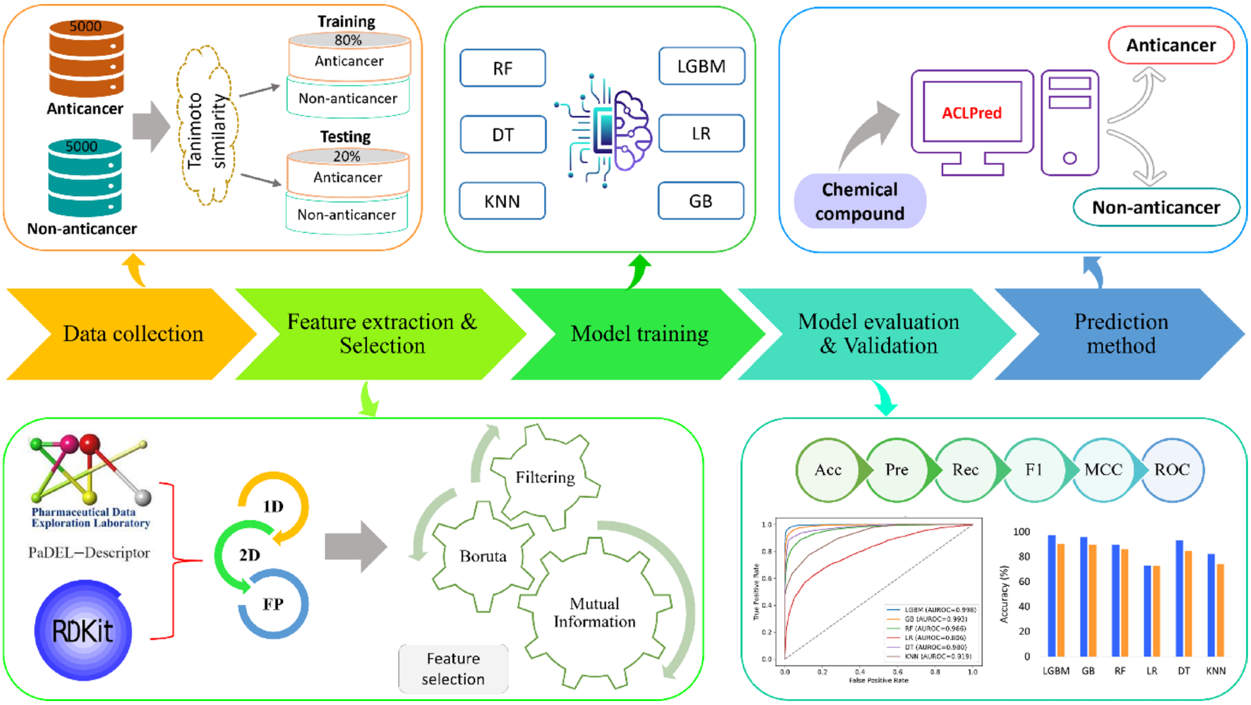


**Fig. 1**. Workflow methodology for proposed ACLPred. *RF* random forest, *LGBM* Light Gradient Boosting Machine, *DT* decision tree, *LR* logistic regression, *KNN* knearest neighbors, *GB* gradient boosting, *MCC* Matthews correlation coefficient, *ROC* receiver operating characteristic curve.

$$T_c(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Where $A$ and $B$ are molecular fingerprints (bit vectors).$|A \cap B|$ is the number of common bits in both fingerprints.$|A|$ and $|B|$ are the total numbers of bits for each fingerprint. $T_c$ ranges from zero to one, and a higher coefficient indicates a greater degree of structural similarity. Molecules with a coefficient of $> 0.85$ were excluded to filter out highly similar molecules. A total of 4706 active and 4867 inactive compounds were identified. Finally, a balanced dataset of 4706 active and 4706 inactive compounds was constructed.

## Feature calculation

An effective numerical representation of molecules is crucial for the development of potential predictive models. Molecular descriptors and fingerprints have been demonstrated to be useful for predicting the properties of various small molecules[22,32,33]. Various types of descriptors were calculated for the chemical molecules represented as SMILES strings using the PaDELPy[34] and RDKit[35] libraries in Python. In total, 1446 one-dimensional (1D) and two-dimensional (2D) descriptors and 881 molecular fingerprints (FP) with a bit vector size of 2048 were calculated using PaDELPy. RDKit produced a total of 210 molecular descriptors. We combined these three categories of descriptors and removed duplicate descriptors; the feature extraction process yielded 2536 descriptors. Before model training, we examined the descriptor dataset for missing (NaN) and infinite values. Descriptors with missing values were substituted with zero, based on the assumption that the absence of a descriptor may indicate a lack of the corresponding molecular feature. Infinite values were replaced with the mean of the corresponding column, as these were considered likely to result from division or transformation artifact. Thus, these enriched and diverse feature sets present broad molecular properties of the compounds which can be utilized to build a reliable model.

## Feature selection

A dataset with many inappropriate features can reduce the model's performance, leading to overfitting, slower computation, and poor prediction. Our dataset comprised 2536 features for 9412 molecules. To select the most relevant features from the dataset, we used a feature-selection process to make the model more accurate and efficient. Additionally, this process helps with dimensionality reduction, faster computation, and the prevention of overfitting. A Python environment was used to execute all the feature selection techniques[36].

### *Variance and correlation filter*

First, we used the variance threshold to filter out low-variance features from the dataset because features with very low variance do not contribute useful information to the model. The variance was calculated using the following formula and features' variance < 0.05 were dropped from the dataset.

$$Var(X_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

where $X_j$ represents feature column $j$, $x_{ij}$ is the value of feature $j$ for sample $i$, $\bar{x}_j$ is the mean of feature $j$, and $n$ is the number of samples.

Furthermore, we used a correlation threshold of 0.85 to eliminate the strongly correlated features. Correlation between features was calculated using the following Pearson correlation coefficient formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) - (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

Where $x$, $y$ represent feature vectors, $\bar{x}$, $\bar{y}$ represents means of vectors $x$ and $y$, respectively, and $r_{xy}$ represents the pearson correlation coefficient between $x$ and $y$.

These relevant filtrations were used with the VarianceThreshold technique available in the scikit-learn package[37] to ensure that only the important features remained. These processes revealed 1313 features for all molecules.

### *Boruta algorithm*

Boruta is a powerful random forest (RF) classifier-based feature selection algorithm that can identify statistically important features in high-dimensional datasets[38]. The Boruta algorithm evaluates the importance of each original feature in proportion to its matching shadow feature to separate important features from those that are not. It computes importance Z-scores for each original feature $j$ as follows:

$$Z_j = \frac{\text{Importance}_j - \mu_{shadow}}{\sigma_{shadow}}$$

where $\text{Importance}_j$ represents the importance score of original features $j$, and $\mu_{\text{shadow}}$ and $\alpha_{\text{shadow}}$ are represents mean and standard deviation of importance scores of shadow features.

The algorithm chooses a feature set that is highly appropriate for the dependent variable ( $Z_j \gg$ max( $Z_{\text{shadow}}$)), instead of selecting the smallest feature set for which a particular model is most suitable. This method produced 431 important features with strong predictive power for anticancer small molecules.

*Mutual information*
A mutual information feature selection technique was used to select a set of descriptors that could capture most information regarding the target variable[39]. It describes the relation between each molecular descriptor and the anticancer label in terms of uncertainty, i.e., entropy. When the two variables are independent, the mutual information value is zero, and a higher value indicates greater dependency. Therefore, the goal was to select features with a higher score for mutual information regarding the anticancer target variable. The mutual information score was calculated using a stratified 10-fold cross-validation with a threshold value of 0.025 for each feature in the training dataset. Finally, this process resulted in a total of 330 highly relevant features. The mutual information between a feature *X* and target *Y* is defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot\log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right)$$

Where $X$ is a feature, $Y$ is the target variable,$p(x,y)$ is the join probability of $x$ and $y$ occurring together, and $p(x), p(y)$ are marginal probabilities of $x$ and $y$.

Thus, the rigorous feature selection process produced a set of significant features (Table 1) for anticancer prediction. This approach enhances the interpretability of the models, decreases the risk of overfitting, and enhances their anticancer predictive capabilities.

## Model training

Using a random state of 42, we split the dataset at a ratio of 80:20 for training and testing the models to predict the anticancer molecules. To offer a baseline assessment of the model's generalization abilities throughout the complete dataset, a random split was selected. Training dataset (80% data) was subjected to 10-fold cross-validation for model building and tuning hyperparameters. The test set (20% data) was kept entirely separate and used only for final performance evaluation. Within the training set, six popular ML algorithms, namely decision tree (DT)[40], RF[41], gradient boosting (GB)[42], LGBM[43], logistic regression (LR)[44], and k-nearest neighbors[45], were used for model training. The DT is an effective approach that generates a tree-like structure by dividing data according to feature values to make decisions[40]. An ensemble learning technique called RF builds several DTs and makes predictions by calculating the mean of the decisions of each tree[41]. GB and LGBM are tree-based ensemble techniques, particularly GB algorithms, which generate a strong classifier by combining multiple weak classifiers[46]. To maximize the prediction accuracy, GB creates models in a sequential manner, where each new model focuses on fixing the errors of the earlier models to improve the prediction accuracy[42]. LGBM is an improved and highly efficient GB framework optimized for speed and memory consumption[43]. LR models a binary dependent variable and calculates the likelihood of a specific class using a logistic function[44]. K-nearest neighbors is a supervised algorithm that makes predictions on how close a data point is to its closest neighbors within a training set[45]. All algorithms for model training were implemented using scikit-learn in the Python programming language[37]. Appropriate hyperparameter tuning (Supplementary Table S1) was employed for all the ML models using a grid search approach to determine the optimal model. Subsequently, a 20% independent test dataset was used to evaluate the transferability of the trained models.

## Model evaluation

Several metrics including as accuracy, precision, recall, F1 score, Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUROC), were used to assess the model's performance. Accuracy is an important parameter that represents the proportion of correct predictions of a model based on true positive and true negative predictions. Precision represents how many of the positive predictions made by the model are actually positive. Recall measures the model's ability to distinguish between actual positive predictions. The MCC evaluates the quality of binary classifications. A higher MCC value indicates a better model. The F1 score measures the balance of precision and recall and is helpful when minimizing both false positives and negatives. AUROC evaluates the discriminative capability of the model between classes. This is particularly helpful in assessing various models and their adaptability to various conditions. A higher AUROC value represents a better classification performance across various thresholds. As 10-fold cross-validation was applied during model training, the final reported results represent the average performance across all 10 folds. Further, to evaluate whether performance differences between models were statistically significant, we

| Step | Features retained | Features removed | Criteria (threshold) |
|---|---|---|---|
| Initial descriptors | 2536 | – | All computed molecular descriptors |
| Variance & correlation filtering | 1313 | 1223 | Variance filter (0.05), and correlation filter (0.85) |
| Boruta algorithm | 431 | 882 | Importance-based selection using random forests |
| Mutual information | 330 | 101 | Top features by mutual information score (0.025) |

**Table 1**. Summary of the feature selection process.

conducted a pairwise statistical t-test on 10-fold cross-validation results of model accuracy. A p-value < 0.05 was considered statistically significant.

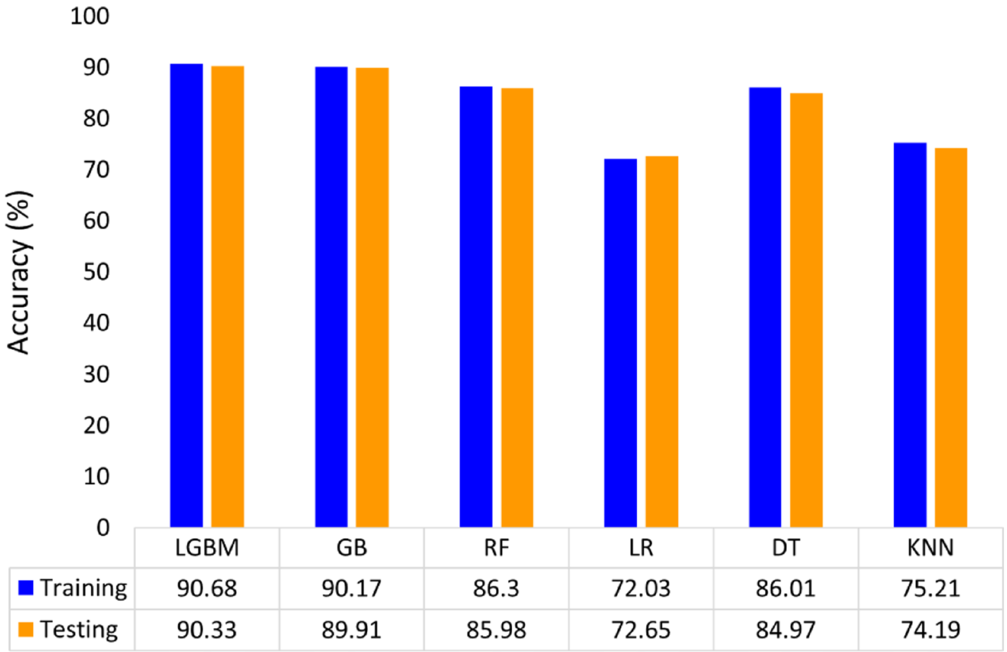
## Model prediction explanation

ML models are used to classify chemical compounds based on their properties, such as biological activities, pharmacokinetics, and molecular structure, helping researchers accelerate drug candidates[47]. To make these predictions, ML models use large and diverse datasets. However, it can be difficult to recognize the logic behind ML models' predictions, which are sometimes supposed of as opaque and black-box techniques. Understanding the underlying causes of accurate ML prediction is of great interest, particularly in pharmaceutical research[48,49]. Explainable ML techniques, such as SHapley Additive explanations (SHAP)[50], offer novel opportunities to uncover the opaqueness of black-box models and offer a more understandable and transparent decision-making process[51,52]. In this study, the SHAP Python package was used to calculate SHAP values and generate a plot.

# Results

## Performance evaluation of models

We utilized 2537 1D, 2D, and FP chemical descriptors as our processed data and applied a multi-technique feature selection strategy to identify the significantly contributing features. The final selected features were input into six different ML algorithms for model training to predict anticancer compounds. To improve model performance, hyperparameter optimization was performed using GridSearchCV on the training dataset with 10-fold cross-validation techniques. All reported metrics are averaged over the 10 folds used in cross-validation. The tree-based ensemble algorithm, LGBM, outperformed the other models and attained the maximum accuracy of 90.68% on the training dataset and 90.33% on the test dataset (Fig. 2). The comprehensive performance evaluation of the classification algorithms, including AUROC, accuracy, precision, recall, F1 score, and MCC are presented in Table 2. For each metric, the highest score achieved by any model is highlighted in bold. It was observed that LGBM performed best in terms of all evaluated metrics. The lowest performance was obtained with the LR model, with prediction accuracies of 72.03% and 72.65% for the training and test datasets, respectively. The tree-based ensemble model LGBM also achieved the highest AUROC (97.73% for training and 97.31% for testing, respectively. The area under precision-recall curve for test set was 97.6% (Fig. 3A,B). This shows that LGBM efficiently classified the dataset between anticancer and non-anticancer drugs with better precision. Furthermore, the model performance results demonstrated that tree-based algorithms, such as LGBM, GB, DT, and RF, performed well. A paired t-test was performed to statistically compare the performance of each classifier, and the obtained p-values are visualized as a heatmap in Supplementary Fig. S1. Our top performing LGBM model exhibited statistically significant difference compared to all other models ($p < 0.001$), except for GB ($p = 0.0645$). Similarly, RF and DT showed no significant difference ($p = 0.4955$), while all other model comparison were significant ($p = 0.001$). This analysis highlights that model selection substantially affects performance outcomes.



| | LGBM | GB | RF | LR | DT | KNN |
|---|---|---|---|---|---|---|
| ■ Training | 90.68 | 90.17 | 86.3 | 72.03 | 86.01 | 75.21 |
| ■ Testing | 90.33 | 89.91 | 85.98 | 72.65 | 84.97 | 74.19 |

**Fig. 2**. Accuracy comparison of the performance of the models on training and test datasets. (*RF* random forest, *LGBM* Light Gradient Boosting Machine, *DT* decision tree, *LR* logistic regression, *KNN* k-nearest neighbors, *GB*, gradient boosting).

| | Training | | | | | | Testing | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | AUROC | Acc | Pre | Recall | F1 | MCC | AUROC | Acc | Pre | Recall | F1 | MCC |
| LGBM | **97.73** | **90.68** | **91.55** | **89.57** | **90.54** | **81.37** | **97.31** | **90.33** | **92.03** | **88.31** | **90.13** | **80.73** |
| GB | 96.90 | 90.17 | 90.94 | 89.20 | 90.05 | 80.34 | 96.83 | 89.91 | 91.49 | 87.99 | 89.71 | 79.88 |
| RF | 94.02 | 86.30 | 88.28 | 83.69 | 85.91 | 72.68 | 93.97 | 85.98 | 87.74 | 83.63 | 85.64 | 72.04 |
| LR | 79.44 | 72.03 | 73.63 | 68.68 | 71.03 | 44.19 | 81.38 | 72.65 | 73.93 | 69.93 | 71.87 | 45.37 |
| DT | 90.57 | 86.01 | 87.74 | 83.71 | 85.66 | 72.10 | 90.15 | 84.97 | 87.82 | 81.19 | 84.37 | 70.14 |
| KNN | 85.76 | 75.21 | 78.23 | 69.83 | 73.76 | 50.73 | 85.84 | 74.19 | 77.05 | 68.86 | 72.73 | 48.65 |

**Table 2**. Performance comparison of various models on both the training (average performance of 10-fold cross-validation) and test datasets, shown in percentage. Top performance values of each metric are highlighted in bold. Metrics: *AUROC* area under the receiver operating characteristic curve, *Acc* accuracy, *Pre* precision, F1 = F1 score, *MCC* Matthews correlation coefficient. *RF* random forest, *LGBM* Light Gradient Boosting Machine, *DT* decision tree, *LR* logistic regression, *KNN* k-nearest neighbors, *GB* gradient boosting.

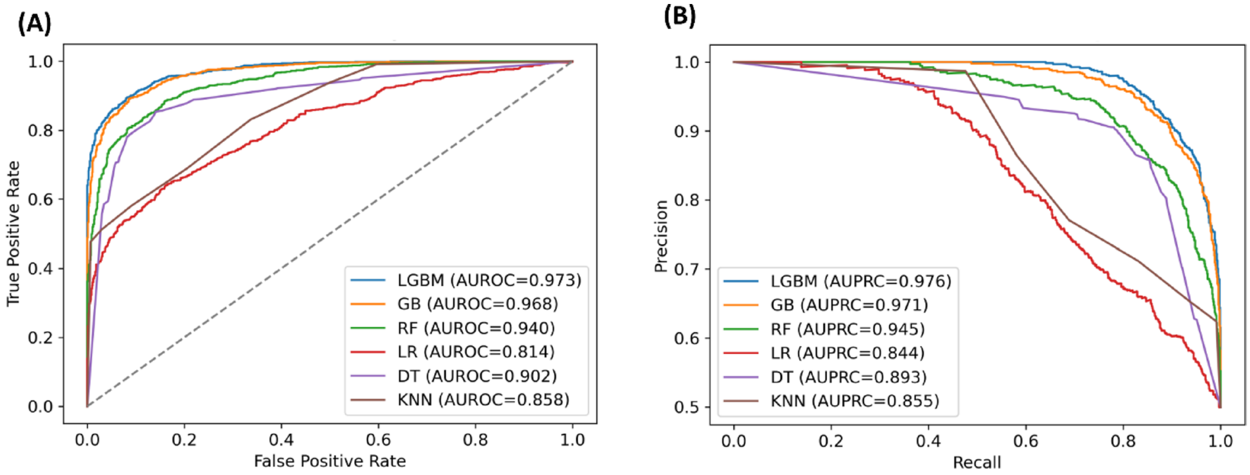


**Fig. 3**. Performance evaluation with area under the receiver operating characteristic curve (AUROC) and area under precision-recall curves (AUPRC) for different machine learning models. (**A**) AUROC and (**B**) AUPRC for the test data. (*RF* random forest, *LGBM* Light Gradient Boosting Machine, *DT* decision tree, *LR* logistic regression, *KNN* k-nearest neighbors, *GB* gradient boosting).

## Comparison of our model with the existing model

To demonstrate the performance and efficiency of ACLPred, we compared its performance with that of recently developed tree-based ensemble method and other existing methods. We compared the training and testing performances of MLASM and ACLPred because both were developed using the same dataset. Compared to MLASM, ACLPred showed superior performance in terms of all measured evaluation metrics for both datasets (Fig. 4). With the accuracy gain of 13.68% and 11.33% for the training and testing datasets, respectively, ACLPred demonstrated a notable improvement. The improvements in AUROC were 12.73% and 9.31% for the training and testing datasets, respectively. Additionally, we compared the performance of our proposed method with that of other existing cell line-specific anticancer prediction methods. However, these methods use different datasets to predict tissue-specific anticancer compounds. The prediction performances of all methods are listed in Table 3.

## SHAP analysis for feature interpretation

After evaluating the effectiveness of the LGBM, we conducted a model interpretation analysis. SHAP was used to predict the most significant features influencing the model's capacity for anticancer ligand prediction. SHAP offers comprehensible and straightforward perceptions into the decision-making processes of complex models by rating the significance of variables based on their impact on the predictions. The most significant molecular descriptors that aid in the prediction of anticancer ligand in the LGBM model were arranged according to their SHAP values and are depicted on the vertical axis of the SHAP summary plot (Fig. 5). The horizontal axis represents the SHAP values, which indicate the influence of each descriptor on the output distribution of the model. Among the top 20 descriptors shown in Fig. 5, IPC (information for polynomial coefficients-based information theory) appears as the top influential descriptor. A higher positive SHAP value of IPC demonstrates the increased probability of a chemical compound being identified as an anticancer agent.

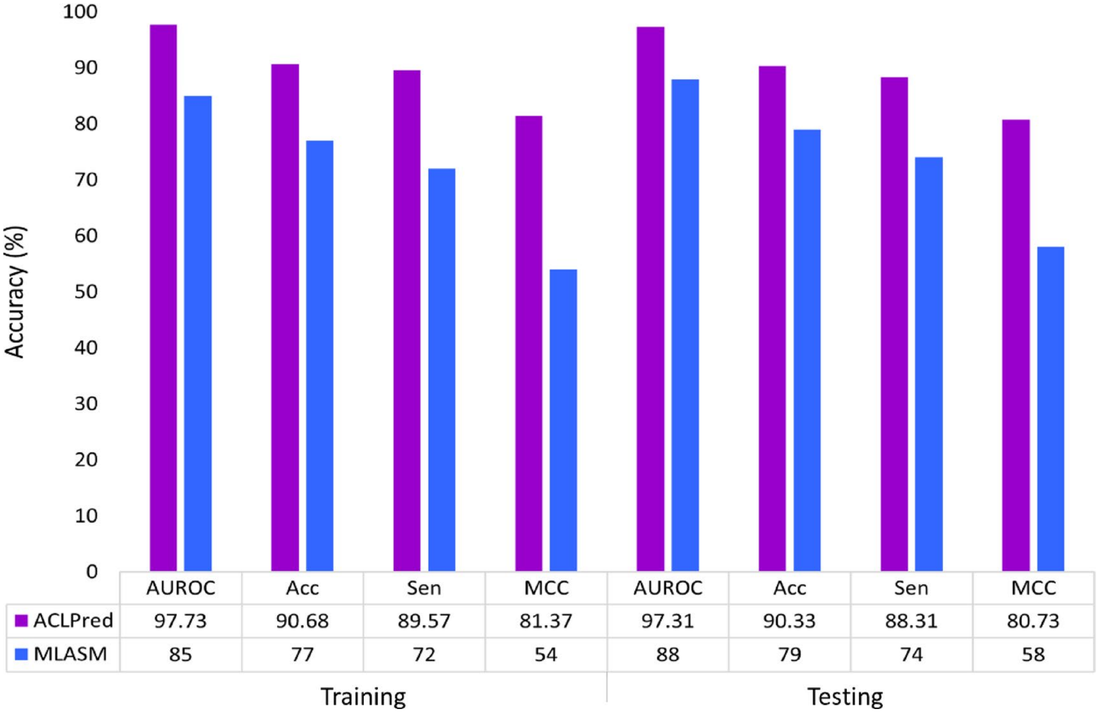


**Fig. 4**. Performance comparison of ACLPred with the previously published method MLASM using the same dataset. (*AUROC* area under the receiver operating characteristic curve, *Acc* accuracy, *MCC* Matthews correlation coefficient).

| Method | AUC | Accuracy | Sensitivity | MCC |
|---|---|---|---|---|
| ACLPred | 97 | 90 | 88 | 80 |
| pdCSM-cancer | 94 | 86 | 84 | 72 |
| MLASM | 88 | 79 | 74 | 58 |
| CDRUG | 87 | * | 81 | * |

**Table 3**. Performance of ACLPred along with the existing methods on their independent test dataset. *Accuracy and MCC predictive scores were not reported by CDRUG. *AUC* area under the curve, *MCC* Matthews correlation coefficient.

A summary of the top 20 important descriptors predicted by SHAP, with higher contributions toward anticancer ligand prediction, is shown in Table 4. Among the top descriptors, 12 were from PaDEL, and the remaining eight were from RDKit. The highest correlation with the anticancer class was observed with the topological descriptor ipc, which quantifies the structural complexity of the molecules. Of the 20 descriptors, six (ipc, MolLogP, R_TpiPCTPC, VSA_Estate10, AATS8v, and SpMAD_Dt) were highly correlated with the positive class means toward anticancer prediction (Fig. 5). This shows that these descriptors are crucial for pushing the model toward the prediction of anticancer compounds.

## External validation and benchmarking

We further tested this using a blind dataset to verify the robustness and generalizability of ACLPred. Balaji et al. performed external validation of the existing MLASM method using a very small dataset ($n = 10$). This small sample size may not be sufficient to draw robust assumptions about the effectiveness and generalizability of the model across diverse scenarios. To overcome this limitation, we emphasize the necessity of a larger and more comprehensive dataset for rigorous evaluation of the model. Therefore, we collected Food and Drug Administration (FDA)-approved cancer drugs from the Anticancer Fund database (https://www.anticancerfund.org/)[53] as a blind dataset. A total of 180 FDA-approved cancer drugs available with SMILES data were considered in our final dataset. The dataset was compared to the training and testing datasets to ensure that no compounds were present in the main dataset. Of the 180 FDA-approved cancer drugs, 162 (90%) were predicted to be active anticancer compounds (Supplementary Table S2). Furthermore, benchmarking our proposed method with existing methods is crucial for assessing its performance. However, again, we faced limitations in our comparisons with the existing model on a blind dataset, as MLASM has no publicly available source code or implementation details. This restriction indicates that computational research must be more open and accessible

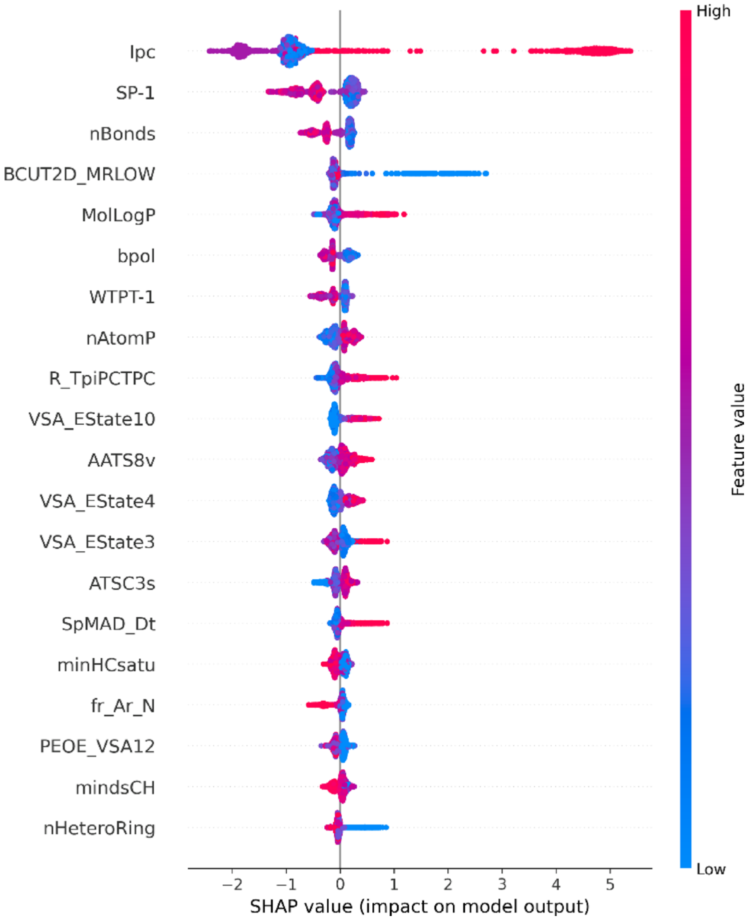


**Fig. 5**. Feature interpretation with SHapley Additive explanations (SHAP) analysis for the light gradient boosting machine model shows the impact of each feature on anticancer molecule predictions. Features are ranked by SHAP values based on their contribution to the model. Positive SHAP values increase the likelihood of anticancer prediction, whereas negative values decrease it.

to facilitate comprehensive benchmarking and reproducibility. Other existing methods, such as pdCSM-cancer and CDRUG, predict activity against cancer cell lines; therefore, direct comparisons with our method are not feasible. Consequently, although we report strong performance, caution should be exercised when interpreting our results relative to prior work.

## Discussion

The prevalence of cancer is rising worldwide; hence, the development of new anticancer drugs with high therapeutic potential is essential for cancer treatment. The discovery of new drug candidates that can inhibit the function of target proteins includes extensive screening and experimental analyses of large chemical libraries[54]. However, the identification of new anticancer drugs from this high-throughput data era is labor-intensive, costly, and time-consuming[55]. Therefore, a combined strategy incorporating both experimental and computational methods is extremely important. Computationally, it is possible to identify molecules with chemical structures similar to those of the active lead compounds. Moreover, the availability of a large chemical library encourages the use of cutting-edge technologies, such as ML, for the fast and efficient prediction of potential lead compounds[56,57].

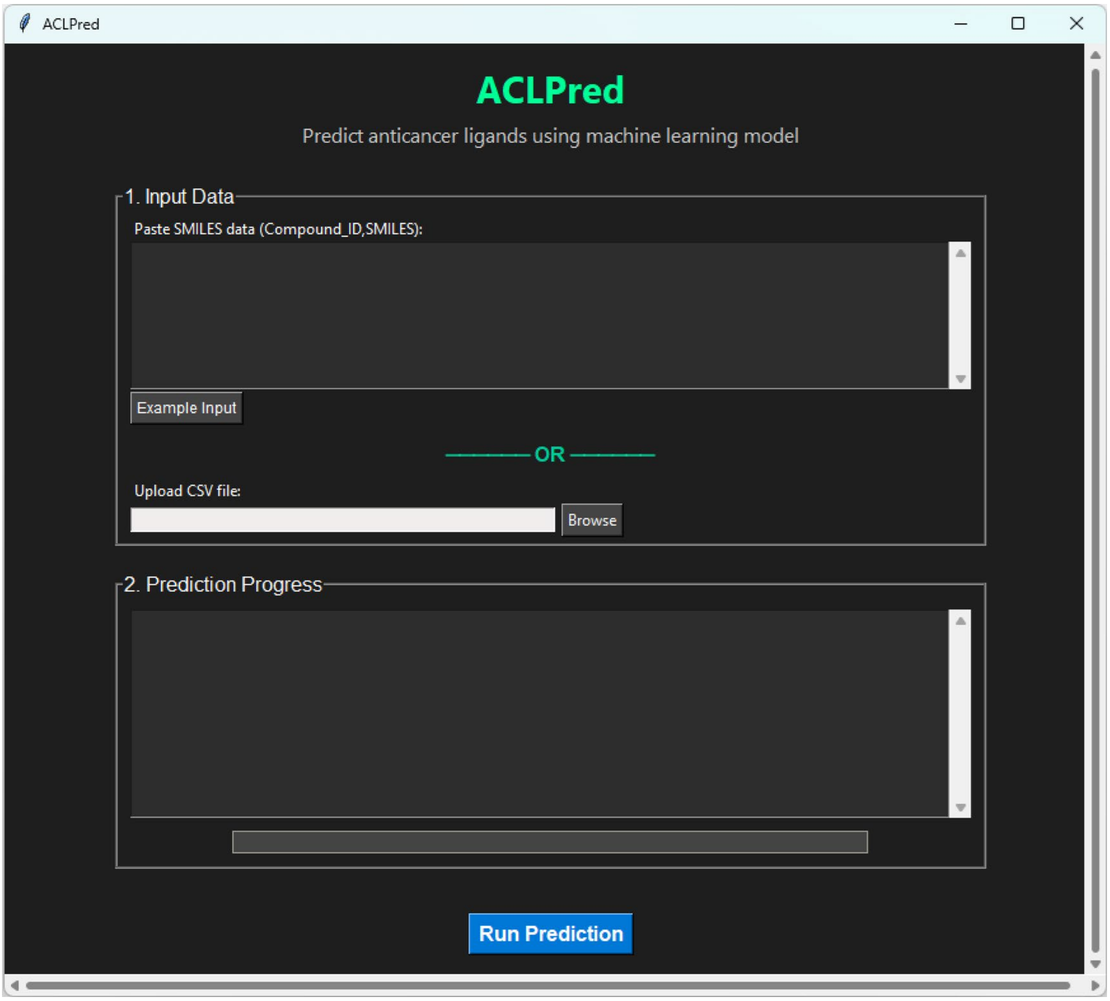
In this study, we developed an ML-based screening method for small-molecule anticancer compounds. We used 5000 active (anticancer) and 5000 inactive (non-anticancer) compounds based on their activities extracted from the PubChem BioAssay database used by Balaji et al. Redundant and structurally similar compounds with a $T_c$ score of > 0.85 were removed. Finally, a balanced dataset of 4706 active and inactive compounds was fed for descriptor calculations using the Python package implementations of PaDEL and RDKit. Feature selection was performed to identify relevant descriptors for developing robust and accurate models. The performance of ML models can be enhanced by using feature selection techniques before model construction[58]. Additionally, multistep feature selection techniques achieve high prediction accuracy with more clinical interpretability[22,59]. Therefore, we applied multistep feature selection techniques to determine the optimum number of top-ranked features. The final dataset contained 330 descriptors, of which 275, six, and 49 were from PaDEL, FP, and RDKit, respectively. The dataset was subjected to model training and testing using various ML algorithms. Among the

| Descriptor name | Tool | Descriptor class | Description |
|---|---|---|---|
| ipc | RDKit | Topological | The amount of information included in the coefficients of the characteristic polynomial of a hydrogen-suppressed molecular graph's adjacency matrix. |
| SP-1 | PaDEL | Chi Path | Based on a simple path of length one in the molecule's molecular graph. |
| nBonds | PaDEL | Bond count | Number of bonds. |
| BCUT2D_MRLOW | RDKit | BCUT | Calculates the lowest atomic contribution to molar refractivity (MR). |
| MolLogP | RDKit | Molecular property | Wildman–Crippen LogP value. |
| bpol | PaDEL | BPol | Total atomic polarizabilities of all bonded atoms (including implicit hydrogens) with respect to their absolute disparities. |
| WTPT-1 | PaDEL | Weighted path | Sums the weights of atom pairs connected by a single bond in the molecule. |
| nAtomP | PaDEL | Largest Pi system | Number of atoms in the largest π-system. |
| R_TpiPCTPC | PaDEL | Path count | The proportion of total path count (up to order 10) to total conventional bond order (up to order 10). |
| VSA_Estate10 | RDKit | MOE-type | Represents the compound's electrical state and its propensity to give or receive electrons. |
| AATS8v | PaDEL | Autocorrelation | Weighted by van der Waals volumes, the average Broto–Moreau autocorrelation (lag 8) follows. |
| VSA_Estate4 | RDKit | MOE-type | Calculates Estate values for specific atom types. |
| VSA_Estate3 | RDKit | MOE-type | Calculates Estate values for different atom types. |
| ATSC3s | PaDEL | Autocorrelation | Weighted by intrinsic state, Centered Broto–Moreau autocorrelation (lag 3). |
| SpMAD_Dt | PaDEL | Detour matrix | Spectral mean absolute deviation from the detour matrix. |
| minHCsatu | PaDEL | Constitutional | Minimum number of hydrogen atoms attached to saturated atoms. |
| fr_Ar_N | RDKit | Fragment-based | Counts the number of aromatic nitrogen atoms in a molecule. |
| PEOE_VSA12 | RDKit | MOE-type | Represents total van der Waals surface area within a defined range. |
| mindsCH | PaDEL | Electrotopological state atom type | E-State minimum for = CH – atom types. |
| nHeteroRing | PaDEL | Ring count | Number of rings with heteroatoms (halogens, N, O, P, or S). |

**Table 4**. Summary of the top 20 contributing descriptors predicted through SHapley additive explanations analysis. *MOE* Molecular Operating Environment, *BPol* Bond Polarizability, *BCUT* Burden Eigenvalue.

six ML algorithms used, tree-based algorithms are the most common because ensemble-based tree methods are more efficient and robust for prediction[46]. The model was trained using a 10-fold cross-validation technique, and its performance was assessed using a test dataset. The outcomes showed that the tree-based methods outperformed the others (Table 2). The maximum performance was achieved by the tree-based ensemble method, LGBM, with an AUROC of 97.31%, an accuracy of 90.33%, an F1 score of 90.13%, and a precision of 92.03% on the independent test dataset. AUROC and area under precision-recall curves (AUPRC) were drawn to visually represent the predictive proficiency of our models (Fig. 3A,B). The AUROC and AUPRC values obtained by the LGBM model were 97.31% and 97.60% on the testing datasets, respectively, highlighting the robustness of our top-performing model. Furthermore, statistical analysis confirmed that our top performing LGBM model differed significantly ($p < 0.001$) with other models except GB. These finding emphasizing importance of model selection and supporting the robustness of its predictive capability (Supplementary Fig. S1). A comparative study suggested that GB algorithms are more efficient in predicting molecular properties[60]. Therefore, we decided to implement a tree-based ensemble method, LGBM, in our proposed ACLPred method. LGBM has been applied in various prediction methods because of its high prediction accuracy, capacity to minimize overfitting problems, and fast computing time[61–63]. The existing method MLASM[28] also achieved the maximum performance with LGBM algorithm using the same dataset. Our method achieved superior performance compared to MLASM employing the same algorithm (Fig. 4). It achieved improved prediction accuracies of 13.68% and 11.33% for the training and testing datasets, respectively. This improvement can be largely attributed to the data processing and implementation of a rigorous multistep feature selection strategy. It allowed to identify the most informative descriptors while reducing the overfitting and noise. It is also important to note that our feature vectors are 330-dimensional, whereas MLASM uses 510-dimentional, which lowers the computing cost. These findings highlight the significance of careful feature selection to develop a robust and generalizable ML model for chemical screening applications. We further evaluated the ACLPred using an external validation dataset, confirming its robustness capability across unseen data. FDA-approved drugs were considered validation and ACLPred positively predicted 90% drugs as anticancer compounds (Supplementary Table S2), which further highlighted the robustness of the proposed method.

Beyond model accuracy, the broader acceptance of computational methods in pharmaceutical research depends heavily on the interpretability of predictions[64]. A model explainability analysis using SHAP was performed to improve trust in and adoption of the model. SHAP can pinpoint the factors that affect a model's decision and facilitate a deeper understanding of the predictive mechanisms behind compound activity models[52]. In our LGBM model, the topology-based descriptor ipc was identified as the top influencing descriptor (Fig. 5). It computes the frequency distribution of atoms and their connectivity in the molecule and quantifies the complexity of the molecular graph[65]. Other highly correlated descriptors in the positive class, such as MolLogP, R_TpiPCTPC, VSA_Estate10, AATS8v, and SpMAD_Dt, exhibit diverse physicochemical and topological properties relevant to anticancer activity[66,67]. MolLogP determines the lipophilicity of a compound and is a key physicochemical factor that affects membrane permeability, absorption, and bioavailability[68]. R_TpiPCTPC is



ACLPred

ACLPred

Predict anticancer ligands using machine learning model

1. Input Data

Paste SMILES data (Compound_ID,SMILES):

Example Input

OR

Upload CSV file:

Browse

2. Prediction Progress

Run Prediction

**Fig. 6**. Graphical user interface (GUI)-based prediction platform of ACLPred.

associated with the topology and size of the molecule and captures bond order and path information that reveals chemical properties[69]. VSA_Estate10 represents the van der Waals surface area and electronic state information of the compound, which is crucial for understanding ligand-receptor interactions[70]. The AATS8v descriptor calculates atomic van der Waals forces at a specific lag, capturing distribution patterns related to steric effects and molecular shape[71]. SpMAD_Dt is a spectral movement descriptor that describes atom-atom distances within a molecule and reflects overall molecular topology[72]. The top-ranked features are those that our model finds most important when making predictions. It is not necessarily that such features are the most prevalent in the dataset; rather, they are the most predictive of biological activity according to the patterns the model has learned. Therefore, the highlighted decision-making descriptors have biological significance and play a major role in the model's decision to classify a compound as anticancer or non-anticancer. Thus, the aforementioned analysis suggests that ACLPred shows substantial improvements and is more practically applicable than existing methods.

## Standalone method

To provide a fast and efficient method for anticancer small-molecule prediction, we have provided a GUI-based Python package available at https://github.com/ArvindYadav7/ACLPred. The GUI framework of ACLPred is developed using the tkinter library of Python and looks like Fig. 6. This freely accessible, standalone ACLPred provides flexibility to users. It allows users to manually input or upload an input CSV file containing one or multiple SMILES strings for prediction. The output result stored in a CSV file indicates whether the given compounds have predicted as 'active' (anticancer) or 'inactive' (non-anticancer) with their prediction probability score. This flexible, standalone method facilitates the screening of larger datasets and enhances its practical utility. Comprehensive user guidelines on using the application can be found on the GitHub link.

## Conclusion

In this study, six ML algorithms were used to build robust and efficient predictive models of anticancer compounds. Many features, such as molecular descriptors and fingerprints, were utilized for model training using different feature selection techniques. The models were evaluated using various performance metrics, and the tree-based ensemble method (LGBM) outperformed the others. The results also suggest that the multistep feature selection technique effectively reduced data dimensionality while improving the model's prediction performance. FDA-approved cancer drugs were used in the external dataset to evaluate the reliability and generalizability of the model. Furthermore, an explainability analysis of the LGBM revealed the important molecular characteristics driving its anticancer properties. Finally, by implementing our tree-based ensemble method, LGBM, we developed ACLPred for the rapid and efficient prediction of anticancer compounds. Moreover, ACLPred outperformed existing methods across all evaluation metrics, and its availability as a standalone, GUI-based tool facilitates the screening of anticancer compounds. We believe ACLPred will be a beneficial resource for identifying novel and potential anticancer compounds. While the present study provides promising in silico predictions, further investigations such as toxicity evaluation, ADMET profiling, and in vivo validation are essential to establish the therapeutic relevance of the identified compounds.

## Data availability

The code and dataset are available at https://github.com/ArvindYadav7/ACLPred.

Received: 10 June 2025; Accepted: 18 August 2025
Published online: 25 August 2025

## References

1. Organization, W. H. O. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer (2025).
2. Yadav, A. K. & Singh, T. R. Computational approach for assessing the involvement of SMYD2 protein in human cancers using TCGA data. *J. Genetic Eng. Biotechnol.* **21**, 122 (2023).
3. AlJarf, R., Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. PiscesCSM: prediction of anticancer synergistic drug combinations. *J. Cheminform.* **16**, 81 (2024).
4. Manica, M. et al. Toward explainable anticancer compound sensitivity prediction via multimodal Attention-Based convolutional encoders. *Mol. Pharm.* **16**, 4797–4806 (2019).
5. Zhong, L. et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Sig Transduct. Target. Ther.* **6**, 1–48 (2021).
6. Liu, B., Zhou, H., Tan, L., Siu, K. T. H. & Guan, X. Y. Exploring treatment options in cancer: tumor treatment strategies. *Sig Transduct. Target. Ther.* **9**, 1–44 (2024).
7. Martin, R. L., Heifetz, A., Bodkin, M. J. & Townsend-Nicholson, A. High-Throughput Structure-Based drug design (HT-SBDD) using drug docking, fragment molecular orbital calculations, and molecular dynamic techniques. *Methods Mol. Biol.* **2716**, 293–306 (2024).
8. Pala, D. & Clark, D. E. Caught between a ROCK and a hard place: current challenges in structure-based drug design. *Drug Discovery Today*. **29**, 104106 (2024).
9. Batool, M., Ahmad, B. & Choi, S. A. Structure-Based drug discovery paradigm. *Int. J. Mol. Sci.* **20**, 2783 (2019).
10. Lin, X., Li, X. & Lin, X. A. Review on applications of computational methods in drug screening and design. *Molecules* **25**, 1375 (2020).
11. Duo, L. & Liu,Yu, R. Jianfeng, tang, Bencan & and hirst, J. D. Artificial intelligence for small molecule anticancer drug discovery. *Expert Opin. Drug Discov.* **19**, 933–948 (2024).
12. Sayers, E. W. et al. Database resources of the National center for biotechnology information in 2025. *Nucleic Acids Res.* **53**, D20–D29 (2025).
13. Zheng, S. et al. Machine learning–enabled virtual screening indicates the anti-tuberculosis activity of aldoxorubicin and Quarfloxin with verification by molecular docking, molecular dynamics simulations, and biological evaluations. *Brief. Bioinform.* **26**, bbae696 (2025).
14. Dai, W., Li, L. & Guo, D. Integrating bioassay data for improved prediction of drug-target interaction. *Biophys. Chem.* **266**, 106455 (2020).
15. Schapin, N., Majewski, M., Varela-Rial, A., Arroniz, C. & Fabritiis, G. D. Machine learning small molecule properties in drug discovery. *Artif. Intell. Chem.* **1**, 100020 (2023).
16. Paul, D. et al. Artificial intelligence in drug discovery and development. *Drug Discov Today*. **26**, 80–93 (2021).
17. Zhang, K. et al. Artificial intelligence in drug development. *Nat. Med.* **31**, 45–59 (2025).
18. Cai, L. et al. Machine learning for drug repositioning: recent advances and challenges. *Curr. Res. Chem. Biology*. **3**, 100042 (2023).
19. Urbina, F., Puhl, A. C. & Ekins, S. Recent advances in drug repurposing using machine learning. *Curr. Opin. Chem. Biol.* **65**, 74–84 (2021).
20. Kumar, R., Chaudhary, K., Singla, D., Gautam, A. & Raghava, G. P. S. Designing of promiscuous inhibitors against pancreatic cancer cell lines. *Sci. Rep.* **4**, 4668 (2014).
21. He, S. et al. Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells. *Front Pharmacol* **12**, (2021).
22. Goel, M., Amawate, A., Singh, A., Bagler, G. & ToxinPredictor Computational models to predict the toxicity of molecules. *Chemosphere* **370**, 143900 (2025).
23. Setiya, A., Jani, V., Sonavane, U. & Joshi, R. MolToxPred: small molecule toxicity prediction using machine learning approach. *RSC Adv.* **14**, 4201–4220 (2024).
24. Menden, M. P. et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLOS ONE*. **8**, e61318 (2013).
25. Singh, H. et al. Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer*. **16**, 77 (2016).
26. Li, G. H. & Huang, J. F. CDRUG: a web server for predicting anticancer activity of chemical compounds. *Bioinformatics* **28**, 3334–3335 (2012).
27. Al-Jarf, R., de Sá, A. G. C., Pires, D. E. V. & Ascher, D. B. pdCSM-cancer: using Graph-Based signatures to identify small molecules with anticancer properties. *J. Chem. Inf. Model.* **61**, 3314–3322 (2021).
28. Balaji, P. D., Selvam, S., Sohn, H. & Madhavan, T. MLASM: machine learning based prediction of anticancer small molecules. *Mol. Divers.* **28**, 2153–2161 (2024).
29. Wang, Y. et al. PubChem's bioassay database. *Nucleic Acids Res.* **40**, D400–D412 (2012).
30. Weininger, D. SMILES, a chemical Language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

31. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
32. Carracedo-Reboredo, P. et al. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **19**, 4538–4558 (2021).
33. Galushka, M. et al. Prediction of chemical compounds properties using a deep learning model. *Neural Comput. Applic.* **33**, 13345–13366 (2021).
34. Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32**, 1466–1474 (2011).
35. RDKit Open-Source Cheminformatics Software. https://www.rdkit.org/.
36. Sanner, M. F. Python: a programming Language for software integration and development. *J. Mol. Graph Model.* **17**, 57–61 (1999).
37. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
38. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
39. Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18** (Suppl 2), S231–240 (2002).
40. SONG, Y. & LU, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry*. **27**, 130–135 (2015).
41. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).
42. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals Stat.* **29**, 1189–1232 (2001).
43. Ke, G. et al. Curran associates, Inc.,. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* **30**, (2017).
44. Kleinbaum, D. G. & Klein, M. Introduction to logistic regression. In *Logistic Regression: A Self-Learning Text* (eds. Kleinbaum, D. G. & Klein, M.) 1–39. https://doi.org/10.1007/978-1-4419-1742-3_1 (Springer, 2010).
45. Kramer, O. K-Nearest neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors* (ed. Kramer, O.) 13–23 (Springer, 2013). https://doi.org/10.1007/978-3-642-38652-7_2.
46. Mahajan, P., Uddin, S., Hajati, F. & Moni, M. A. Ensemble learning for disease prediction: A review. *Healthc. (Basel)*. **11**, 1808 (2023).
47. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. *J. Comput. Aided Mol. Des.* **34**, 1013–1026 (2020).
48. Ribeiro, M. T., Singh, S. & Guestrin, C. *'Why Should I Trust You?': Explaining the Predictions of Any Classifier*. https://doi.org/10.48550/arXiv.1602.04938 (2016).
49. Polishchuk, P. Interpretation of quantitative Structure–Activity relationship models: past, present, and future. *J. Chem. Inf. Model.* **57**, 2618–2639 (2017).
50. Lundberg, S. & Lee, S. I. *A Unified Approach to Interpreting Model Predictions*. https://doi.org/10.48550/arXiv.1705.07874 (2017).
51. Karim, M. R. et al. Explainable AI for bioinformatics: methods, tools and applications. *Brief. Bioinform*. **24**, bbad236 (2023).
52. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. *J. Med. Chem.* **63**, 8761–8777 (2020).
53. Pantziarka, P., Capistrano, I., De Potter, R., Vandeborne, A., Bouche, G. & L. & An open access database of licensed cancer drugs. *Front. Pharmacol.* **12**, 627574 (2021).
54. Yadav, A. K., Singh, T. R. & and Novel inhibitors design through structural investigations and simulation studies for human PKMTs (SMYD2) involved in cancer. *Mol. Simul.* **47**, 1149–1158 (2021).
55. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov*. **18**, 463–477 (2019).
56. Rehman, A. U. et al. Role of artificial intelligence in revolutionizing drug discovery. *Fundamental Res.* https://doi.org/10.1016/j.fmre.2024.04.021 (2024).
57. Singh, S., Gupta, H., Sharma, P. & Sahi, S. Advances in artificial intelligence (AI)-assisted approaches in drug screening. *Artif. Intell. Chem.* **2**, 100039 (2024).
58. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A review of feature selection methods for machine Learning-Based disease risk prediction. *Front. Bioinform.* **2**, 927312 (2022).
59. Wang, H. et al. An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records. *BMC Med. Inf. Decis. Mak.* **25**, 84 (2025).
60. Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L. & Sieber, S. A. Practical guidelines for the use of gradient boosting for molecular property prediction. *J. Cheminform.* **15**, 73 (2023).
61. Shaker, B. et al. LightBBB: computational prediction model of blood–brain-barrier penetration based on LightGBM. *Bioinformatics* **37**, 1135–1139 (2021).
62. Zhang, J., Mucs, D., Norinder, U., Svensson, F. & LightGBM An effective and scalable algorithm for prediction of chemical Toxicity–Application to the Tox21 and mutagenicity data sets. *J. Chem. Inf. Model.* **59**, 4150–4158 (2019).
63. Zhang, C., Lei, X. & Liu, L. Predicting Metabolite–Disease associations based on LightGBM model. *Front. Genet.* **12**, 660275 (2021).
64. Kırboğa, K. K., Abbasi, S. & Küçüksille, E. U. Explainability and white box in drug discovery. *Chem. Biol. Drug Des.* **102**, 217–233 (2023).
65. Nolte, T. M., Peijnenburg, W. J. G. M., Hendriks, A., Jan & van de Meent, D. Quantitative structure-activity relationships for green algae growth Inhibition by polymer particles. *Chemosphere* **179**, 49–56 (2017).
66. Guo, H. et al. *Tailoring Chemical Molecular Representation to Specific Tasks via Text Prompts*. https://doi.org/10.48550/arXiv.2401.11403 (2024).
67. Bertato, L., Chirico, N. & Papa, E. QSAR models for the prediction of dietary biomagnification factor in fish. *Toxics* **11**, 209 (2023).
68. Morak-Młodawska, B., Jeleń, M., Martula, E. & Korlacki, R. Study of lipophilicity and ADME properties of 1,9-Diazaphenothiazines with anticancer action. *Int. J. Mol. Sci.* **24**, 6970 (2023).
69. Chen, T. & Manz, T. A. Bond orders of the diatomic molecules. *RSC Adv.* **9**, 17072–17092.
70. Du, X. et al. Insights into Protein–Ligand interactions: mechanisms, models, and methods. *Int. J. Mol. Sci.* **17**, 144 (2016).
71. Escayola, S., Bahri-Laleh, N. & Poater, A. % V Bur index and steric maps: from predictive catalysis to machine learning. *Chem. Soc. Rev.* **53**, 853–882 (2024).
72. Kehrein, J., Bunker, A., Luxenhofer, R. & POxload Machine learning estimates drug loadings of polymeric micelles. *Mol. Pharm.* **21**, 3356–3374 (2024).

## Acknowledgements

We thank the National Research Foundation of Korea (NRF), Ministry of Education for providing us with financial support through grants.

## Author contributions

J.M.K. conceived the study. A.K.Y. carried out all the experiments and data analysis. A.K.Y. and J.M.K. participated in the overall design and coordination of the study. The first draft of the manuscript was prepared by A.K.Y. Both authors read and approved the final manuscript.

## Funding

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [grant number: RS-2018-NR031061].

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-16575-4.

**Correspondence** and requests for materials should be addressed to J.-M.K.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.