Unveiling Machine Learning Algorithms for predicting Drug Activity against Lung Cancer Cell Lines

Kanagasabapathy Gokulakrishnan[1], Krishnamoorthy Hema Nandini Rajendran[2], Veerappapillai Shanthi[2], Pachaiappan Jayakrishnan[3] and Karuppasamy Ramanathan[2]*

1. Department of Software and Systems Engineering, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA
2. Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA
3. Department of Micro and Nanoelectronics, School of Electronics Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA
*kramanathan@vit.ac.in

Abstract

Lung cancer remains a significant global health concern, posing a substantial burden on both patients and healthcare systems. As a result, there is an urgent need for innovative therapeutic interventions to manage lung cancer more effectively. In this study, we developed classification models using machine learning algorithms to predict drug responses in lung cancer cell lines. A diverse dataset was retrieved, consisting of 692 active and 1,071 inactive compounds tested against five major lung cancer cell lines: CaLu-06, HCC-78, NCI-H322, NCI-H358 and NCI-H522. Drug-like properties of these compounds were generated and employed as descriptors for model development.

The proposed method utilised techniques such as z-score, correlation analysis, recursive feature elimination with cross-validation and SMOTE to pre-process the data and identify key features. Further, hyperparameter optimisation was conducted using Optuna to fine-tune model parameters and enhance performance. The results revealed that Random Forest reached an accuracy of 0.80 and an AUC of 0.85. This positions it as the best model, with significant implications for drug discovery and personalised lung cancer therapies. The implementation materials alongside python code are accessible freely at https://github.com/Gokulakrish13/Machine-Learning-Classifiers-for-Predicting-Active Molecules-Against-Lung-Cancer-Cells.git.

Keywords: Lung cancer, Drug Response, GDSC, Cell lines, Machine Learning Models.

Introduction

Lung cancer (LC), characterised by a poor overall 5-year survival rate, is one of the leading causes of cancer-related deaths globally.[8] Despite extensive efforts to develop enhanced therapeutic strategies including chemotherapies, targeted therapies and immunotherapies, only modest survival rates and clinical outcomes have been observed.[17,19] Moreover, the drug resistance associated with existing molecules further worsens the situation, clearly highlighting the necessity for novel therapeutics. Interestingly, lung cancer treatment could be significantly improved by exploring new and innovative sources of bioactive compounds. These naturally occurring molecules have the potential to act as anti-cancer agents, offering a promising avenue for more effective therapies.

In recent years, phenotypic-based drug screening has garnered significant attention due to the substantial number of drugs identified and approved through this approach. However, its application is often associated with disadvantages such as high costs, low throughputs and challenges in optimizing emerging hit compounds.[18] To address these limitations, machine learning (ML) and deep learning models are increasingly used to exploit phenotypic-based data, often in conjunction with molecular structures or multiomics knowledge. This integration enhances the accuracy of modern translational precision medicine.[12] For instance, Wu et al[19] developed a user-friendly web server called DeepCancerMap to accelerate the discovery of anticancer drugs.

In another study, Qin et al[12] created an MLP-based regression model to predict the bioactivity of compounds targeting the Erα protein, demonstrating efficacy with a loss value of 0.0146, highlighting its potential for screening active compounds for breast cancer management. Similarly, Bonanni et al[2] constructed several machine learning models using compounds with highly consistent cell-based antiproliferative assay data to predict ligand activity for the PC-3 and DU-145 prostate cancer cell lines. Additionally, He et al[7] developed predictive models using fingerprints and molecular descriptors for 13 different breast cancer cell lines, showcasing the potential of these techniques in enhancing cancer therapy.

Although the reported computational models have provided valuable insights into the discovery of anticancer agents, machine learning models employing drug-like descriptors remain elusive. Note that drug-like properties are crucial in the development of machine learning models for drug discovery and response prediction.[3] Specifically, studies have shown that incorporating ADME (Absorption, Distribution, Metabolism and Excretion) data into predictive models can significantly improve the accuracy of drug response predictions.[9] Addressing this gap, the present study focuses on extensively training machine learning models using data from the Genomics of Drug Sensitivity in Cancer (GDSC) database. This approach aims to predict the

antiproliferative cellular activity of compounds specifically against aggressive lung cancer cell lines including CaLu-06, HCC-78, NCI-H322, NCI-H358 and NCI-H522.

By leveraging the extensive genomic and pharmacological data available in GDSC, the study seeks to enhance the precision and efficacy of drug response predictions, ultimately contributing to the development of more effective therapeutic strategies for lung cancer.

Material and Methods

Dataset preparation and descriptor generation: Initially, the IC_{50} values for anti-cancer drugs against five lung cancer cell lines: CaLu-06, HCC-78, NCI-H322, NCI-H358 and NCI-H522, were obtained from the GDSC database.[5] Consequently, we processed the acquired data using the following steps: (1) Only compounds with reported IC_{50} values were kept and those lacking bioactivity data were discarded; (2) The bioactivity data were converted to the standard unit, μM; (3) Compounds with IC_{50} values ≤10μM were classified as active and those with higher values as inactive.[7] Finally, the five cell lines with more than 50 active and 50 inactive molecules were retained. All datasets employed for the models in this research are publicly accessible at https://github.com/Gokulakrish13/Machine-Learning-Classifiers-for-Predicting-Active Molecules-Against-Lung-Cancer-Cells.git. In the current investigation, to characterize the drug-like information of the molecules, 53 distinct type ADME descriptors were generated using Qikprop module of the Schrödinger suite.

Data Pre-processing: The pre-processing steps collectively aimed to enhance the quality of the dataset, improve model performance and facilitate more reliable predictions in subsequent machine learning tasks.[16] During the data pre-processing phase, the dataset was loaded and subjected to basic exploratory data analysis to comprehend its structure and characteristics. Numerical features were then normalized and standardized using the StandardScaler and MinMaxScaler respectively, to ensure uniformity and mitigate the impact of varying scales on model performance.[20]

Following this, the categorical features were encoded using LabelEncoder to convert them into numerical representations suitable for modelling. The outliers were detected using the Z-score method with a threshold of 3 and those identified were subsequently removed from the dataset. A high number of outliers in certain columns suggested these features might be irrelevant for the analysis.

A correlation analysis was performed by generating a correlation matrix to investigate the relationships among features. This analysis revealed several features with low or negligible correlation with the target variable, suggesting they were likely irrelevant for modelling purposes. For instance, some features displayed minimal correlation coefficients, further supporting the decision to exclude them from further analysis. Features lacking clear trends or patterns were identified. These insights, combined with the domain knowledge, guided the selection of only the most pertinent features for the final dataset. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalances in the dataset.[6] Finally, feature selection techniques, including recursive feature elimination with cross-validation (RFECV), were utilized to meticulously identify and retain the most pertinent features for the modeling process.[4]

Machine learning model construction: A systematic approach was followed in constructing machine learning algorithms to ensure the development of robust models that accurately predict the target variable. Five conventional machine learning algorithms including Logistic Regression, Naïve Bayes, Random Forest, K-Nearest Neighbor and Support Vector Machine, were employed for the development of classification models. Each algorithm was implemented using Scikit-Learn's library, with appropriate hyperparameters chosen either through manual tuning or automated techniques like grid search and randomized search.[13] Finally, the dataset was partitioned into training and testing sets using the train_test_split function to streamline the process of model development and evaluation.

Performance Evaluation of Models: Various performance metrics, including accuracy, precision, recall, F1-score and area under the ROC curve (AUC), were utilized to assess the effectiveness of the models. To ensure robust evaluation, k-fold cross-validation was employed, which helped to gauge the models' generalization performance and mitigate the risk of overfitting.[21] The best-performing model was selected based on a comprehensive analysis of these performance metrics, with additional consideration given to its interpretability and suitability for the specific task of predicting antiproliferative activity against lung cancer cell lines. These meticulous steps collectively contributed to the development of robust and reliable machine learning algorithms capable of making accurate and insightful predictions on the dataset.

Results and Discussion

Dataset analysis and Pre-processing: In the data pre-processing phase, the dataset was subjected to several key transformations to ensure its suitability for subsequent analysis and modelling. Initially, the dataset underwent thorough exploration, including an examination of column names and statistical summaries to gain insights into structure and characteristics. The `missingno` library was used to identify and visualize the null values present in the dataset. This analysis showcased that no missing values were found which ensured the integrity of the dataset. Then by utilizing the `StandardScalar` from the scikit-learn library, we performed standardization, which transformed the data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms sensitive to scale.

Further, `LabelEncoder` from the scikit-learn was used for converting categorical labels into numerical format that facilitate their integration into machine learning models. Accordingly, the labels 'Active' and 'Inactive' were represented numerically as [0, 1]. Similarly, the cell line data were encoded as [1, 2, 3, 4, 0] for 'HCC-78', 'NCI-H322', 'NCI-H358', 'NCI-H522' and 'CaLu-06' respectively.

To avoid undue impact on the model's performance, outliers were detected by applying the z-score method, setting the threshold at 3. This widely recognized statistical technique corresponds to a confidence interval of approximately 99.7%. Data points that occur beyond three standard deviations from the mean, are considered statistically rare and therefore, potential outliers.[1] In the present study, the Z-score method identified 41 descriptors with outliers, which were then imputed with median values. Finally, the pre-processed data was saved for further analysis and model development, ensuring the integrity and reproducibility of subsequent efforts.

Selection of important features for model construction: Feature selection is the process of identifying and selecting the most informative features from a dataset, aiming to improve model performance by reducing dimensionality and eliminating irrelevant or redundant features. The present study employed correlation analysis and RFECV to identify the subset of features that contribute most to predictive accuracy. In the present study, the 51 features were subjected to correlation analysis to identify any redundant features by examining the correlations between the features using heatmap. From figure 1, it can be observed that 12 positively correlated features ('Cell_Line', 'mol_MW', 'dipole', 'SASA', 'volume', 'dip^2/V', 'QPpolrz', 'QPlogPC16', '#ringatoms', '#in56', '#nonHatm') are the most relevant for model training and evaluation.

Additionally, RFECV was also employed using a Random Forest classifier as the base model, which iteratively selects the most informative features for machine learning models while evaluating performance through cross-validation, thereby enhancing model interpretability and generalization. The results of RFECV show the same set of optimal features, ensuring that the selected features were highly correlated with each other. Ultimately, the models were developed using these 12 features.

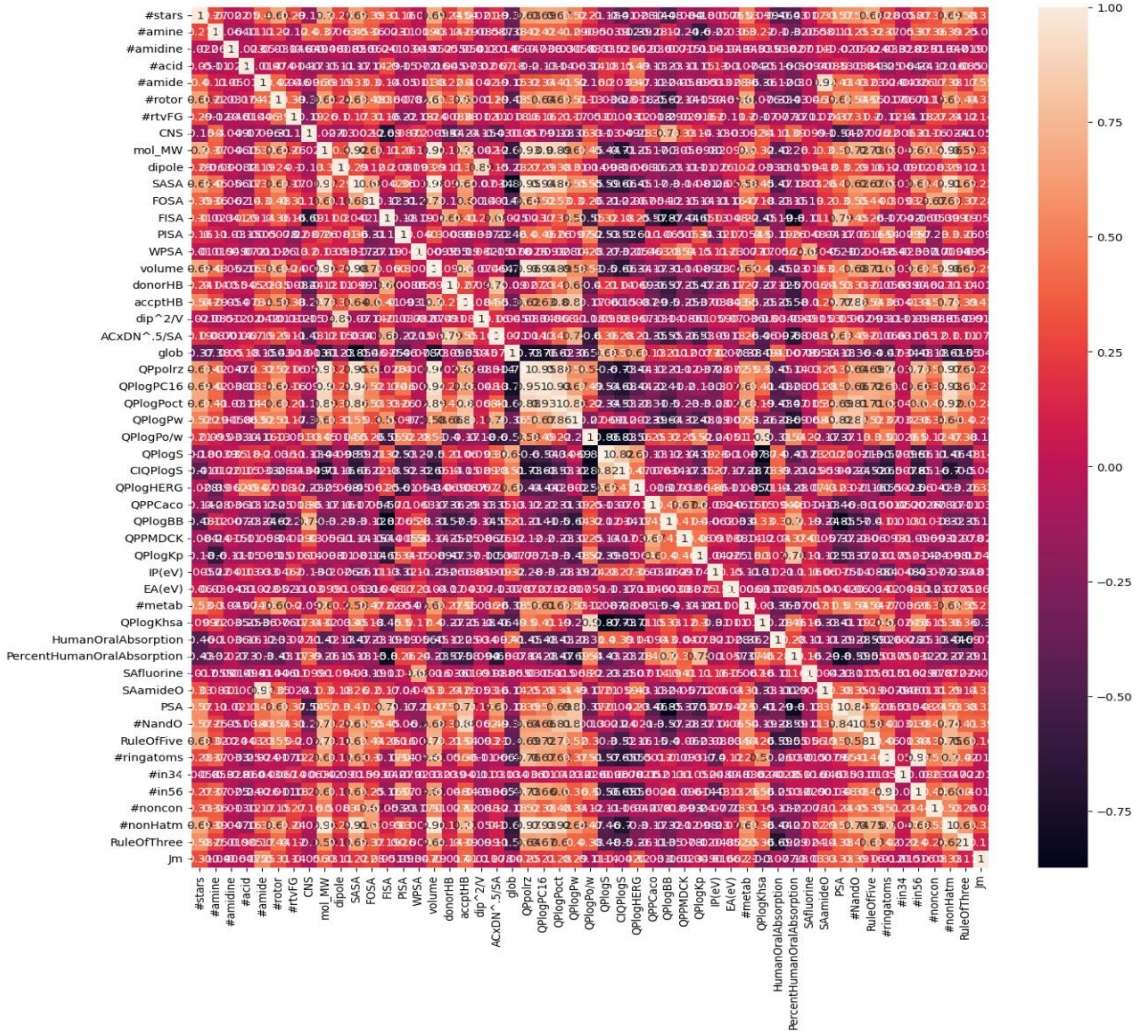
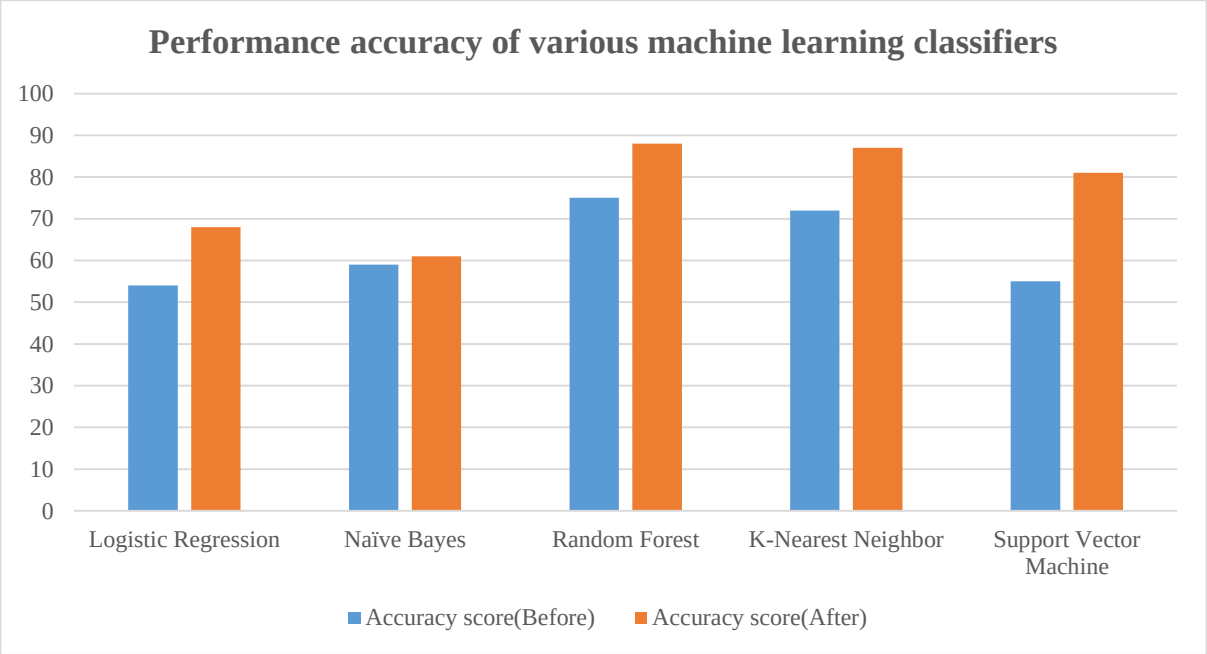
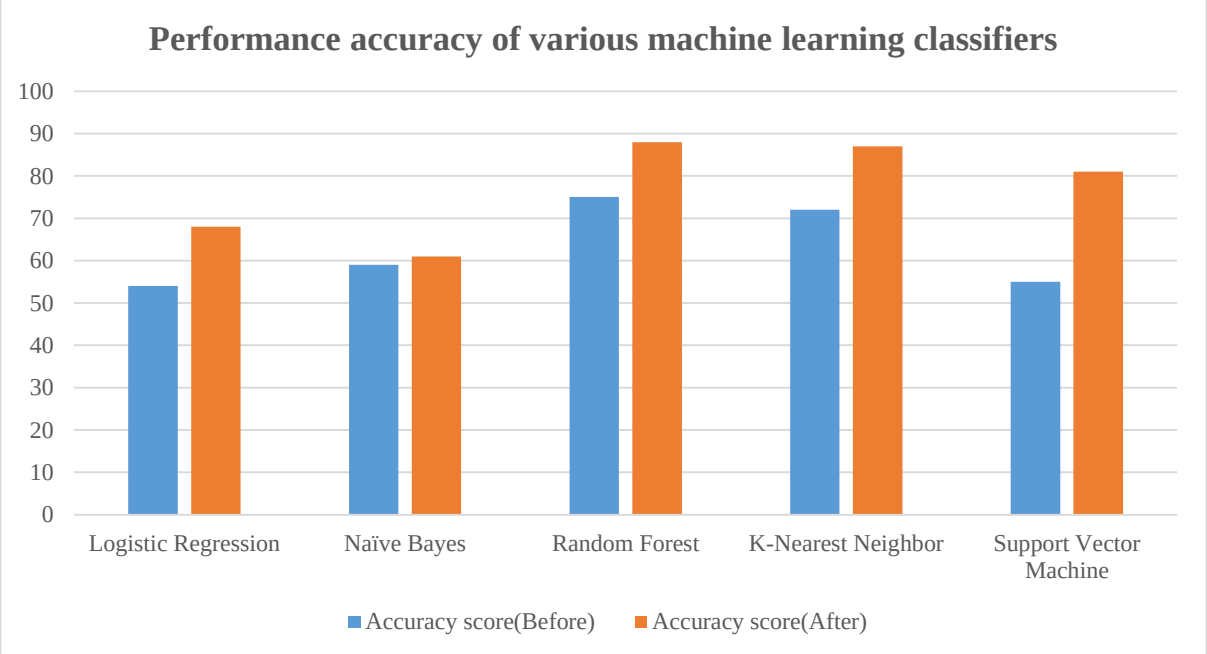


Figure 1: Heatmap visualization depicting the interrelationships between the various features and the target variable within the dataset



(a)



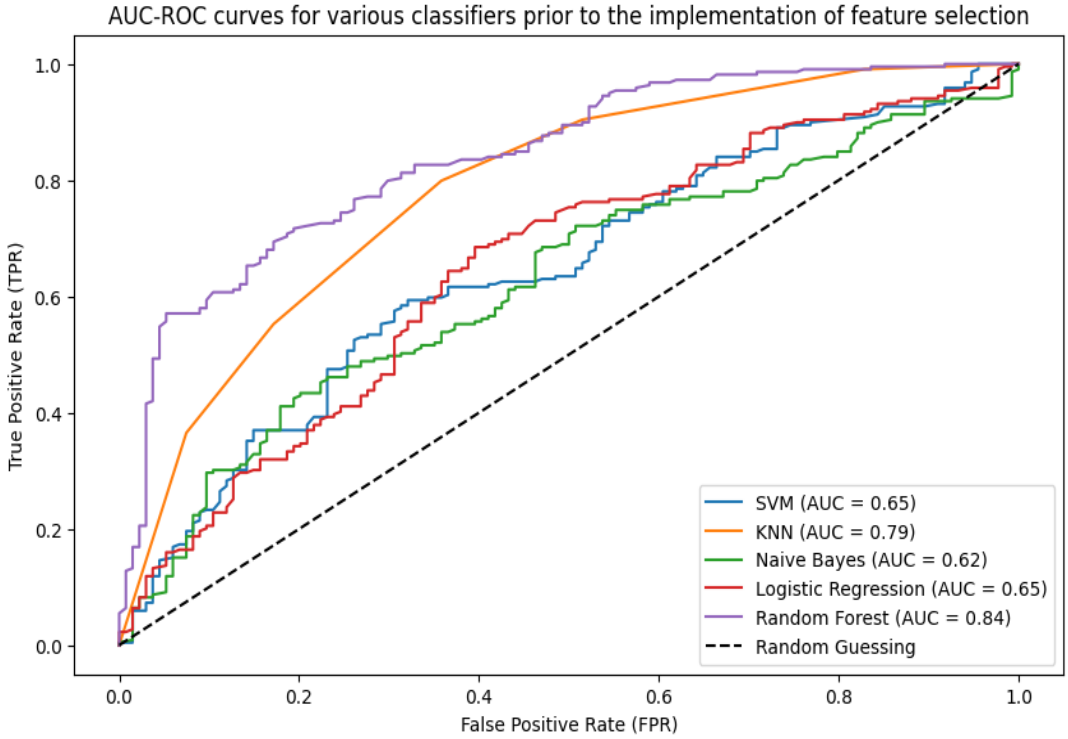
(b)

Figure 2: Comparison of the performance accuracy of various machine learning classifiers (a) Training set (b) Test set

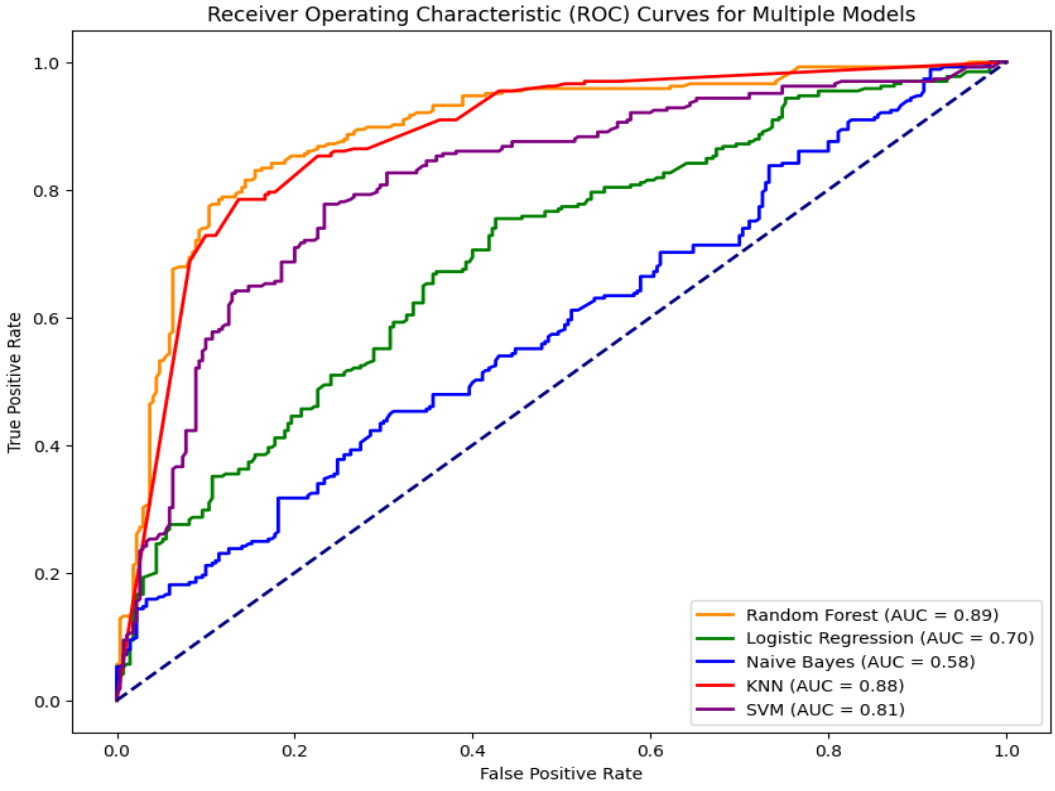
Performance evaluation of the developed models: Table 1 and table 2 showed the performance of the five classifiers before and after the feature selection process. The simplest technique to evaluate a model's performance is to measure its accuracy. Similarly, figure 2 and figure 3 depicted the performance of the different classifiers in distinguishing active and inactive drugs before and after the feature selection approach. In the present study, the random forest classifier with hyperparameters {'max_depth': 13, 'n_estimators': 78} emerged as the top performer, achieving an accuracy of 0.80, a recall of 0.77, a precision of 0.82 and an F1-score of 0.79, along with an AUC value of 0.85. This indicates its ability to accurately classify both positive and negative instances while maintaining a balance between recall and precision.

Similarly, logistic regression, using {'C': 6.11, 'penalty': 'l2', 'solver': 'liblinear'}, displayed lower performance, with recall, precision and F1-score all at 0.59 and an AUC of 0.64. The KNN algorithm demonstrated notable performance, achieving an accuracy of 0.76, precision of 0.78, recall of 0.73 and an F1-score of 0.76, along with an AUC value of

0.84, using the hyperparameters {'metric': 'manhattan', 'n_neighbors': 21, 'weights': 'distance'}, showcasing its ability to effectively classify instances based on their nearest neighbors. In contrast, the SVM classifier, with hyperparameters {'C': 7, 'kernel': 'rbf', 'gamma': 'scale'}, exhibited the lowest performance, with an accuracy of 0.54, precision of 0.62, recall of 0.24, F1-score of 0.34 and an AUC value of 0.55. Collectively, the evaluation metrics highlight that the random forest algorithm surpasses other classifiers in predicting drug responses in lung cancer cell lines.



(a)



(b)

Figure 3: AUC-ROC curves for various classifiers (a) before feature selection (b) after feature selection

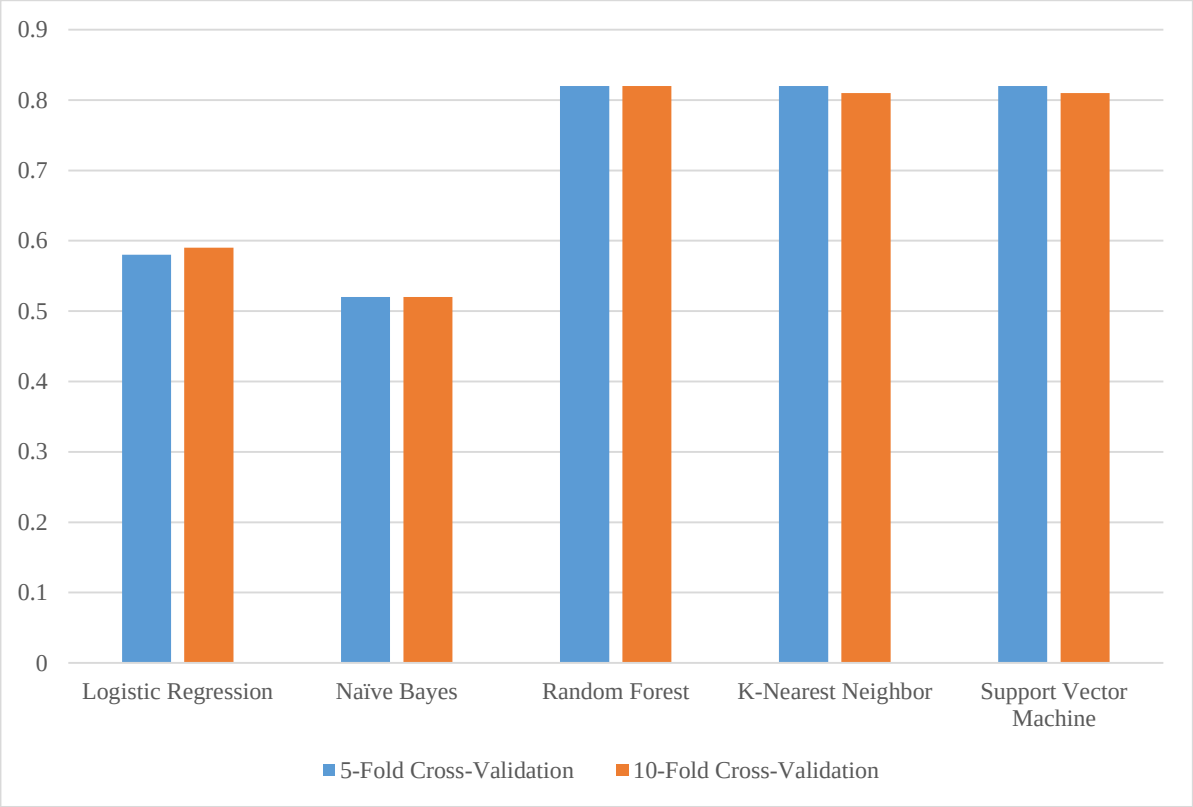


Figure 4: Comparative analysis of various machine learning classifiers through the implementation of K-Fold cross-validation

Table 1
Performance metrices before feature selection

Classification Algorithms	Accuracy score	Precision	Recall	F1-Measures	AUC value
Logistic Regression	0.651	0.66	0.86	0.75	0.65
Naïve Bayes	0.60	0.69	0.64	0.67	0.62
Random Forest	0.767	0.815	0.80	0.81	0.84
K-Nearest Neighbor	0.74	0.78	0.80	0.79	0.79
Support Vector Machine	0.62	0.62	0.33	0.77	0.65

Table 2
Performance metrices after feature selection

Classification Algorithms	Accuracy score	Precision	Recall	F1-Measures	AUC value
Logistic Regression	0.65	0.64	0.67	0.65	0.70
Naïve Bayes	0.55	0.55	0.51	0.53	0.58
Random Forest	0.83	0.83	0.83	0.83	0.89
K-Nearest Neighbor	0.80	0.81	0.79	0.82	0.87
Support Vector Machine	0.75	0.76	0.73	0.75	0.81

Table 3
K – Fold cross-validation score for different classifiers after feature selection

Classification Algorithms	5-Fold Cross-Validation	10-Fold Cross-Validation
Logistic Regression	0.58	0.59
Naïve Bayes	0.52	0.52
Random Forest	0.82	0.82
K-Nearest Neighbor	0.82	0.81
Support Vector Machine	0.82	0.81

Performance evaluation of cross validation: In the realm of model evaluation, employing cross-validation techniques is crucial for robustly assessing machine learning algorithms. In a 5-fold cross-validation, where the data is partitioned into five subsets for iterative model training and testing, the SVM classifier demonstrated a mean accuracy of 0.54 ± 0.0226, indicating a moderate performance level. In contrast, the KNN classifier exhibited a notably higher mean accuracy of 0.78 ± 0.0200, suggesting a strong predictive advantage with relatively low variance. Similarly, the Naive Bayes classifier achieved a mean accuracy of 0.56 ± 0.0181, while the Logistic Regression model and the Random Forest classifier attained mean accuracies of 0.57 ± 0.0101 and 0.79 ± 0.0268 respectively (Table 3).

To gain further insights into model performance, a 10-fold cross-validation framework was carried out. The results are depicted in figure 4. Despite a slight decrease in mean accuracy, the SVM classifier maintained competitive performance, achieving an accuracy of 0.53 ± 0.0377. Notably, the KNN classifier exhibited robustness with a consistent mean accuracy of 0.80 ± 0.0168, signifying its reliability across diverse data partitions. Meanwhile, the Naive Bayes classifier demonstrated a stable mean accuracy of 0.56 ± 0.0288, corroborating its reliability in varied cross-validation scenarios.

The Logistic Regression model displayed a mean accuracy of 0.58 ± 0.0310, while the Random Forest classifier sustained its effectiveness with a mean accuracy of 0.80 ± 0.0337. Both methods provide robust performance estimates, with 10-fold cross-validation typically offering higher accuracy at the expense of increased computational cost. Collectively, these findings highlight that the random forest consistently exhibited the highest mean accuracies across both 5-fold and 10-fold splits, underscoring its reliability and stability in classification tasks.

Conclusion

This study developed five predictive models using machine learning techniques to identify novel drug candidates for lung cancer treatment, assembling datasets comprising of 692 active and 1,071 inactive compounds across five commonly utilized lung cancer cell lines for *in vitro* antiproliferative assessments. Among the models evaluated: Logistic Regression, Naïve Bayes, Random Forest, K-Nearest Neighbor and Support Vector Machine, the Random Forest algorithm performed the best, achieving an accuracy of 0.80 and an AUC of 0.85, effectively distinguishing between active and inactive compounds. These findings indicate that the Random Forest model can guide future drug discovery efforts and personalized treatment strategies in lung cancer management.

Looking ahead, an online platform and local version of software based on these well-established models could be developed to significantly contribute to research aimed at designing and discovering new anti-lung cancer agents. As the database of compound toxicity data for lung cancer and normal cell lines expands, we will incorporate additional predictive models in future studies.

Acknowledgement

The authors thank the Management of Vellore Institute of Technology for providing the facilities to carry out this research work.

References

1. Berendrecht W., Van Vliet M. and Griffioen J., Combining statistical methods for detecting potential outliers in groundwater quality time series, *Environmental Monitoring and Assessment*, **195**, 85 **(2023)**

2. Bonanni D., Pinzi L. and Rastelli G., Development of machine learning classifiers to predict compound activity on prostate cancer cell lines, *Journal of Cheminformatics*, **14(1)**, 77 **(2022)**

3. Calado C.R., Bridging the gap between target-based and phenotypic-based drug discovery, *Expert Opinion on Drug Discovery*, **19(7)**, 789-798 **(2024)**

4. Chahar R.K. and Singh A.S., Using Machine Learning Techniques for Outlier Detection Application, 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), *IEEE*, 238-241 **(2022)**

5. Garnett M.J., Edelman E.J., Heidorn S.J., Greenman C.D., Dastur A., Lau K.W. and Benes C.H., Systematic identification of genomic markers of drug sensitivity in cancer cells, *Nature*, **483(7391)**, 570-575 **(2012)**

6. Gulati V. and Raheja N., Efficiency Enhancement of Machine Learning Approaches through the Impact of Preprocessing Techniques, 6th International Conference on Signal Processing, Computing and Control (ISPCC), *IEEE*, 191-196 **(2021)**

7. He S., Zhao D., Ling Y., Cai H., Cai Y., Zhang J. and Wang L., Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells, *Frontiers in Pharmacology*, **12**, 796534 **(2021)**

8. Herrera-Juárez M., Serrano-Gómez C., Bote-de-Cabo H. and Paz-Ares L., Targeted therapy for lung cancer: Beyond EGFR and ALK, *Cancers*, **129(12)**, 1803-1820 **(2023)**

9. Joo M., Park A., Kim K., Son W.J., Lee H.S., Lim G. and Nam S., A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients, *International Journal of Molecular Sciences*, **20(24)**, 6276 **(2019)**

10. Krentzel D., Shorte S.L. and Zimmer C., Deep learning in image-based phenotypic drug discovery, *Trends in Cell Biology*, **33(7)**, 538-554 **(2023)**

11. Kunjir A., Joshi D., Chadha R., Wadiwala T. and Trikha V., A comparative study of predictive machine learning algorithms for COVID-19 trends and analysis, International Conference on Systems, Man and Cybernetics (SMC), *IEEE*, 3407-3412 **(2020)**

12. Qin Y., Li C., Shi X. and Wang W., MLP-based regression prediction model for compound bioactivity, *Frontiers in Bioengineering and Biotechnology*, **10**, 946329 **(2022)**

13. Ramesh P. and Veerappapillai S., Prediction of micronucleus assay outcome using in vivo activity data and molecular structure features, *Applied Biochemistry and Biotechnology*, **193**, 4018-4034 **(2021)**

14. Ramesh P., Karuppasamy R. and Veerappapillai S., Machine learning driven drug repurposing strategy for identification of potential RET inhibitors against non-small cell lung cancer, *Medical Oncology*, **40(1)**, 56 **(2022)**

15. Sadri A., Is target-based drug discovery efficient? Discovery and "off-target" mechanisms of all drugs, *Journal of Medicinal Chemistry*, **66(18)**, 12651-12677 **(2023)**

16. Salmani-Javan E., Jadid M.F.S. and Zarghami N., Recent advances in molecular targeted therapy of lung cancer: Possible application in translation medicine, *Iranian Journal of Basic Medical Sciences*, **27(2)**, 122 **(2024)**

17. Shobana G. and Priya D.N., Cancer drug classification using artificial neural network with feature selection, Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), *IEEE*, 1250-1255 **(2021)**

18. Tousi A. and Luján M., Comparative analysis of machine learning models for performance prediction of the spec benchmarks, *IEEE Access*, **10**, 11994-12011 **(2022)**

19. Wu J., Xiao Y., Lin M., Cai H., Zhao D., Li Y. and Wang L., DeepCancerMap: a versatile deep learning platform for target-and cell-based anticancer drug discovery, *European Journal of Medicinal Chemistry*, **255**, 115401 **(2023)**.

(Received 25th September 2024, accepted 04th December 2024)