

Tailoring therapies: a frontier approach to pancreatic cancer with AI-driven multiomics profiling

Janiel Jawahar, Paramasivan Selvi Rajendran

Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India

Article Info

Article history:

Received Feb 15, 2024

Revised Apr 4, 2024

Accepted Apr 13, 2024

Keywords:

Drug response prediction
Ensemble feature selection
Molecular descriptors
Multiomics
Pancreatic cancer
Personalized therapeutics
TabNet

ABSTRACT

Pancreatic cancer is often diagnosed at an advanced stage when treatment options are limited. Being one of the deadliest cancers that mandates longer medication and treatment phases, there is an inevitable need to have the knowledge of drug response of anti-pancreatic cancer drugs before it is recommended for a patient. AI-driven drug response prediction has proven potential to personalize treatment strategies, improve therapeutic outcomes, and reduce adverse effects and treatment costs for cancer patients. In this research work, we have accounted for the use of different drug descriptors and their core structures known as scaffolds along with three cell line features, chromatin profiling, reverse phase protein array, and metabolomics data to build a feature engineered dataset for drug response prediction tested on various computational learning models. The 53 unique drugs against 18 unique pancreatic cancer cell lines were taken as the raw dataset. The initial dataset having a large dimension was feature selected using an ensemble method derived from five different techniques. The dataset was evaluated on various computational methods and an accuracy of 89% was achieved using the TabNet architecture. Furthermore, the common scaffolds that were persistently found among the drugs that possess high IC50-valued drug clusters were also recorded.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Janiel Jawahar

Department of Computer Science and Engineering, Hindustan Institute of Technology and Science

Padur, Chennai, Tamil Nadu, India

Email: jani.hits23@gmail.com

1. INTRODUCTION

Cancer is an illness with various forms and complexities. It's crucial to understand that among patients with the same cancer type the effectiveness of anti cancer treatments can vary [1]. In particular pancreatic cancer is an aggressive malignancy known for being detected at stages when limited treatment options and having an unfavorable prognosis. Pancreatic cancer contributes significantly to the cancer mortality rate requiring prolonged periods of medication and treatment due to its increasing prevalence. Detecting this disease early poses a challenge since its often diagnosed when it has already reached advanced stages. Developed countries in North America and Europe have rates of pancreatic cancer with the United States having one of the highest rates. The majority, 80% of individuals diagnosed with pancreatic cancer have either locally advanced or distant metastatic disease. Unfortunately these cases have low chances of long term survival with an overall 5 year survival rate below 10%. However individuals who are diagnosed with cancer in its early stages have the opportunity to achieve a cure by undergoing a combination of surgical procedures chemotherapy treatments and radiotherapy sessions [2]. Hence it is of importance to develop a comprehensive comprehension of pancreatic cancer and the medications that are recommended for its

treatment. This knowledge has the potential to enhance the chances of survival and minimize the overall mortality rate.

Pancreatic cancer has a lower occurrence comparatively while age plays a role as a risk factor and it is not practical to screen the entire population solely based on age due to the expensive clinical tests involved and the possibility of false positive results in many patients. Therefore prioritizing the utilization of automated systems, for detecting cancer is crucial to improve accuracy and treatment outcomes [3]. Furthermore it's important to understand and predict how anticancer drugs recommended for therapy will work. Artificial intelligence (AI) which refers to computer systems using advanced learning algorithms to mimic intelligence and perform tasks has been successfully applied in oncology. It has contributed to advancements in diagnosing and treating gastrointestinal cancers, including pancreatic cancer. These advancements include the use of techniques aimed at enhancing patient prognosis [1]-[3]. Personalized therapy, also referred to as precision oncology, involves tailoring medications to suit individual patients. The use of population based drug doses for patient care often leads to variations in treatment outcomes and side effects among individuals [4], [5]. The imperfections in cancer medicine delivery systems can be attributed to their complexity [4]. The primary objective of precision cancer therapy is to enhance the effectiveness of treatment [6]. However it is equally important to understand the factors associated with drug response in order to adopt a comprehensive approach to precision medicine in the field of oncology [4], [7], [8]. Establishing correlations between genes and medications can be challenging due to the intricate nature of genomics and drug sensitivity data. To address these challenges various computational techniques have been developed, with machine learning algorithms proving successful in achieving this goal [8]-[10]. Incorporating a range of cell line features such as chromatin profiling, reverse-phase protein arrays (RPPA) and metabolomics data provides a holistic understanding of the molecular characteristics of cell lines. Consequently, this enhances the precision and significance of models for drug response while offering valuable insights. Medical experts believe that by studying tumors at the molecular level they can develop customized treatments that target specific subgroups of tumors and even individual patient characteristics. This approach has the potential to greatly improve the effectiveness of therapies and yield better treatment outcomes [11].

When working with datasets that have dimensions in different fields like document and image analysis biomedical data and others it is crucial to carefully select a relevant subset of features [12]. The emergence of datasets with a number of columns or features has caught the attention of many researchers in the field of feature selection. Several techniques have been proposed in studies to forecast drug sensitivity [13]. However these approaches lack the effectiveness to accurately predict drug sensitivity. This is because no single selection algorithm can guarantee the results in terms of stability and predictive performance [14]. It is widely acknowledged that not all columns contain information and having irrelevant or noisy columns can confuse machine learning algorithms thereby hindering their effectiveness. As a result researchers have been exploring the efficacy of approaches that combine multiple selectors to achieve optimal outcomes [13]-[16]. This study focuses on examining the utilization of drug descriptors and drug molecular core structures along with three cell line features such as chromatin profiling, RPPA and metabolomics data for predicting the drug response to anti pancreatic cancer drugs. To manage the vast variety of features in cell lines we utilized an ensemble approach for selecting features. The ensemble outcomes were achieved by implementing five techniques, for feature selection.

2. RELATED STUDIES

In their study researchers led by Qiu *et al.* [7] discovered genes that can predict the response, to drugs with high reliability and accuracy. They trained machine learning models using biomarker features, including gene expression profiles, mutation profiles, pathways, methylation and copy number variations. In order to improve the accuracy of predictions additional information such, as the chemical composition of drugs was also incorporated into the models. Lanka *et al.* [8] developed an ensemble machine learning algorithm called (ELAFT), which aimed to predict the effectiveness of anti cancer drugs. By harnessing the strengths of machine learning techniques, like regression classifiers random forest (RF) classifiers, k-nearest neighbor (KNN) classifiers and support vector machines (SVM) this method achieved impressive accuracy levels surpassing 90%. Pearson correlation was employed to enhance the accuracy further.

Suphavilai *et al.* [9] proposed a method, for predicting how cancer drugs would respond using a recommender system approach. The method involves mapping drugs and cell lines into a latent space called (CaDRReS) which helps identify similarities between medications and cell lines. The authors utilized the genomics of drug sensitivity in cancer (GDSC) and the cancer cell line encyclopedia (CCLE) dataset to evaluate their approach. According to their findings, CaDRReS has shown performance compared to progressive methods, in accurately predicting drug response. Tan *et al.* [3] have come up with a method to

predict how cancer cells will respond to a chemical used in chemotherapy. They used a combination of machine learning algorithms and two unique signatures derived from gene expression profiles of cancer cell lines that were exposed to the drug. Furthermore, Partin *et al.* [17] have proposed the utilization of multi omics data such as gene expression, protein expression and DNA methylation to anticipate the response of drugs. Summary graphs were created to depict the 61 distinct deep learning based models that were selected for examination. Alwi *et al.* [18] utilizing catalogue of somatic mutations in cancer (COSMIC) datasets, conducted a study that explores the correlation between resistant mutations and resistance to cancer medications by considering protein structures. They also developed a predictive model capable of identifying which mutation is responsible for drug resistance, encompassing both inherent resistance (which is observed before treatment) and acquired resistance (which occurs after treatment). Similarly the DualGCN method, introduced by Ma *et al.* [19], presents an innovative approach for predicting cancer drug response. This method specifically targets the shortcomings of current techniques by effectively transferring knowledge from in vitro cancer cell lines to single-cell and clinical data, without the need for extensive single nucleotide variant data. DualGCN demonstrated potential for use in clinical and single-cell data, independent of extensive single nucleotide variant (SNV) data. The authors also highlighted the importance of considering the intricate tumor microenvironment and the limitations of SNV-based models in understanding drug response.

Zhu *et al.* [20] have proposed a combination approach that involves three different types of models, namely light gradient boosting machine (LightGBM), single deep neural network (sDNN) and time delay neural network (tDNN), while tDNN is a two subnetwork DNN and sDNN is a single network DNN, LightGBM is a decision tree (DT) model. The authors trained and evaluated these models using cancer cell line encyclopedia (CCLE), cancer therapeutics response portal (CTRP), and NCI60. Their model achieved an accuracy rate of 82% on CCLE datasets, 85% on CTRP datasets and 80% on NCI60 datasets. An inevitable constraint in handling the gene expression data is the huge dimensionality of features they possess. The advantage of adding more cell line features in model training to help better prediction would also potentially increase the dataset dimension thereby making it difficult to handle and becomes computationally expensive. This narrows the way of finding the best features to select of the dataset before training. Seijo-Pardo *et al.* [21] have provided an ensemble feature selection method for handling high-dimensional data such as gene expression data. They have done this using feature-class and feature-feature mutual information, by combining optimal subsets selected by various filters such as Chi-square, Info Gain, Gain Ratio, Relief F, and Symmetric Uncertainty. They claimed to have achieved excellent results through this approach by validating in different machine learning algorithms such as the KNN, RF, and SVM applied on two networks and five gene expression datasets. Similarly, Mera-Gaona *et al.* [16] have reported the use of ensemble advantage in feature selection through a theoretical and practical framework that aids in comprehending the fundamental ideas and connections involved in combining different feature selection algorithms. Jin *et al.* [22] have created a sophisticated deep learning algorithm called differential genes screening TabNet (DGS-TabNet) that can classify Alzheimer's disease using genetic data for both binary and multiclass scenarios. The DGS-TabNet model has surpassed various state-of-the-art deep learning and classical machine learning models, achieving an impressive accuracy of 93.80% for binary classification and 88.27% for multi-class classification on a gene expression dataset. Similarly, Arik *et al.* [23] and Kita *et al.* [24] have utilized TabNet in their research, achieving superior results in managing large-scale medical tabular data.

3. MATERIAL AND METHODS

To build a resourceful feature-engineered dataset, cell line data such as chromatin profiling data, metabolomics data, reverse phase protein array data and drug line data such as drug inhibition data (IC50) for cell lines were downloaded from the GDSC and CCLE database. RDKit and Datamol [25] libraries were used to process the drug simplified molecular-input line-entry system (SMILES) feature. Google's Colab in browser and Microsoft Visual Studio Code were used as IDE for running the python scripts. All the code implementation and results discussed were done on a 12th Gen Intel Core i5-1240P-1.70 GHz processor with a 16GM RAM machine.

3.1. Experiments on drug molecular descriptors

The raw drugline data downloaded from the GDSC site had information such as drug name, drug experimented cancer cell name, their inhibition results, pathway name and putative target as features. A Python script was written to retrieve the SMILES data for the corresponding drug names. This was achieved through an API offered by PubChem, an open chemistry database maintained by the National Institutes of Health (NIH). Using the open source python libraries, RDKit library, and Datamol [25] library, multiple features from the SMILES data were derived and built into our dataset.

Using the k-means clustering algorithm, the dataset was clustered into three classes, high, moderate and low inhibition based on the inhibition values. This new feature was used as the target class for training and model validation. The correlation between features was calculated using pearson correlation and the pairwise similarity between features was calculated using the cosine similarity method. The feature importance chart Figure 1, Table 1 derived using the cosine similarity and the correlation heatmaps for each cluster Figures 2-4. The integration of RDKit's Murcko scaffold approach transforms molecular structures into simplified scaffolds. The conversion of RDKit molecular objects into SMILES representations further enhances the comparative analysis of our datasets. The resulting count of occurrences for each unique scaffold in the dataset provides valuable information regarding the structural diversity and prevalence of specific frameworks. This approach proves to be instrumental in identifying common structural motifs within each inhibition cluster, paving the way for a more targeted and efficient exploration of chemical space. The core structures Figures 5-7 that are found only in high-inhibition drugs were also found and recorded.

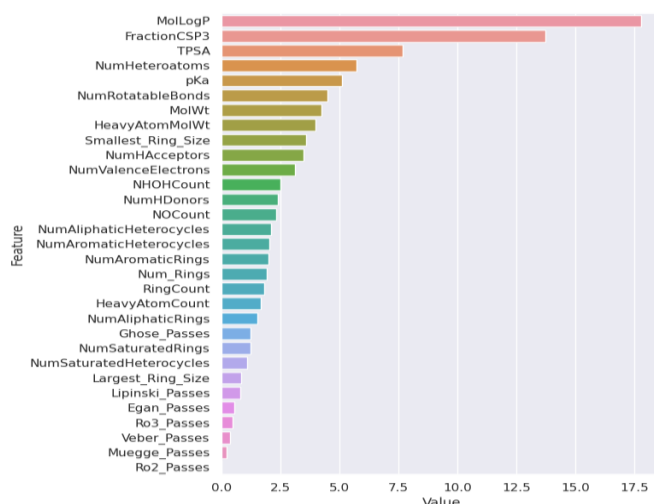


Figure 1. Feature importance score of drug descriptors

Table 1. Feature importance score of drug descriptors in each cluster

Drug descriptors	Cluster-0 feature importance's	Cluster-1 feature importance's	Cluster-2 feature importance's
Largest_Ring_Size	151.2131	208.188	62.8674
Smallest_Ring_Size	151.2422	208.1102	62.8406
pKa	151.113	208.1469	62.7967
MolLogP	151.3571	208.0691	62.6538
MolWt	151.0908	205.4052	62.4535
HeavyAtomCount	151.3571	208.1639	62.8298
HeavyAtomMolWt	150.7802	208.2461	62.8288
NHOHCount	150.6921	207.949	62.8036
NOCCount	150.3723	207.576	61.8106
NumHAcceptors	150.4577	208.2221	62.8583
NumHDonors	151.446	207.4655	62.8466
NumHeteroatoms	151.2559	208.0154	62.6891
NumRotatableBonds	151.2948	208.1327	62.6538
NumValenceElectrons	150.4779	206.3011	62.4535
NumAromaticRings	150.2994	203.4908	62.764
NumSaturatedRings	150.9827	207.386	62.6463
NumAliphaticRings	151.376	207.4655	62.8288
NumAromaticHeterocycles	151.1853	205.4052	62.8036
NumSaturatedHeterocycles	150.9936	208.1634	62.8583
NumAliphaticHeterocycles	151.1327	208.1639	62.6891
RingCount	151.1836	208.2461	62.6538
FractionCSP3	151.3733	207.949	62.4535
TPSA	151.0664	208.2419	62.764
Veber_Passes	151.3904	207.576	62.8288
Ghose_Passes	151.31	208.1666	62.8036
Muegge_Passes	151.4269	208.2221	62.8583
Ro3_Passes	151.3571	208.0219	62.8466
Egan_Passes	150.6921	207.4655	62.6891
Ro2_Passes	150.4577	208.141	62.764
Lipinski_Passes	151.2559	208.0154	62.8288

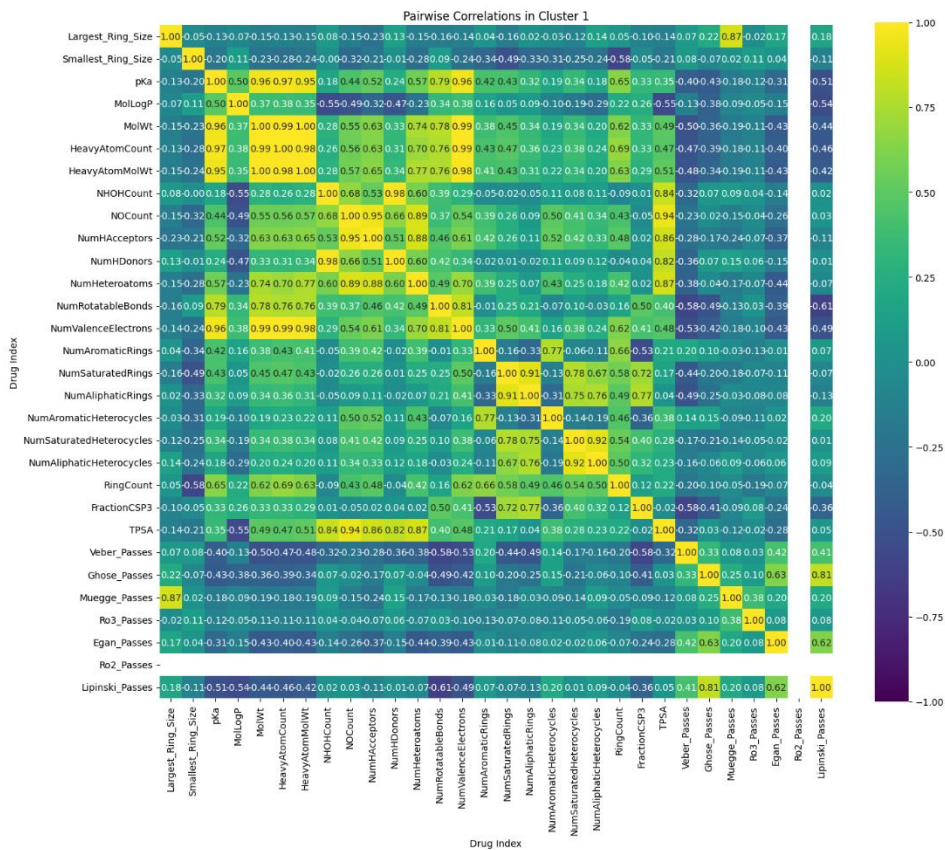


Figure 2. Correlation heatmap of derived drug descriptors in moderate inhibiting cluster

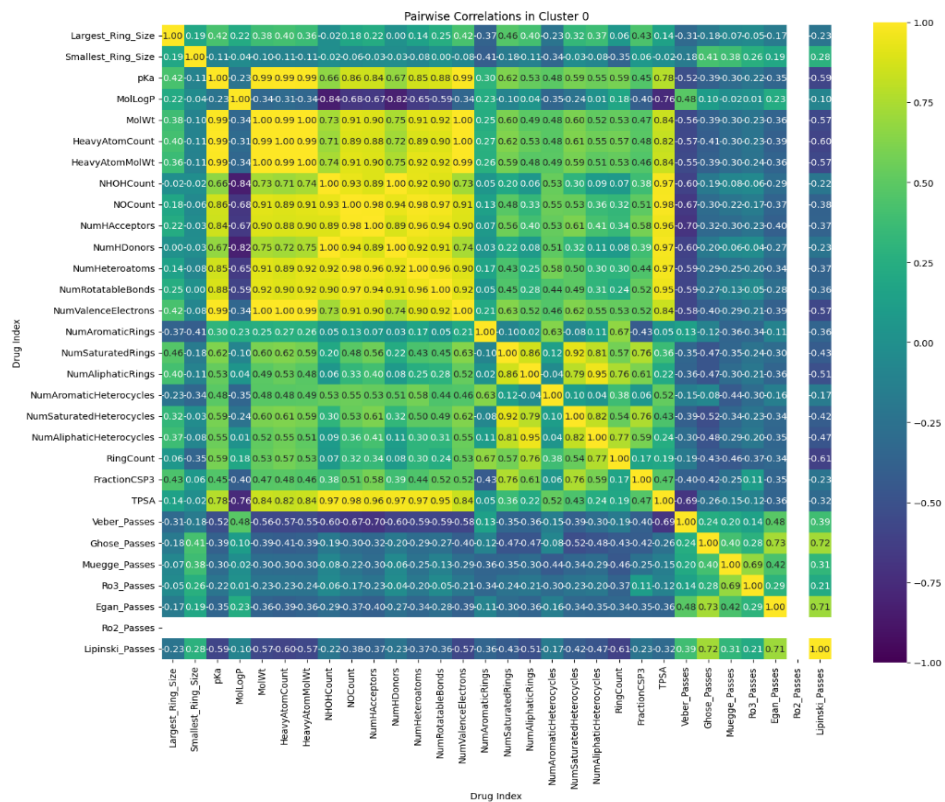


Figure 3. Correlation heatmap of derived drug descriptors in low inhibiting cluster

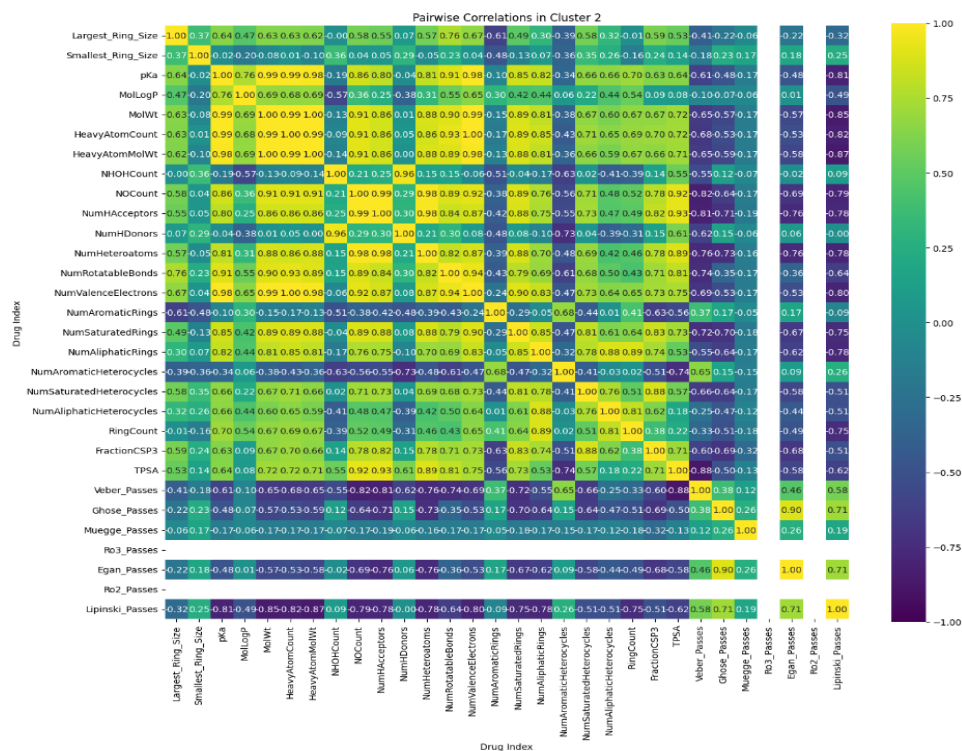


Figure 4. Correlation heatmap of derived drug descriptors in high inhibiting cluster

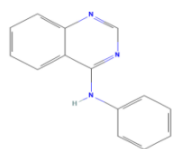


Figure 5. 4-Anilinoquinazoline

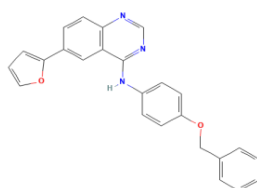


Figure 6. (4-Benzyloxy-phenyl)-(6-furan-2-yl-quinazolin-4-yl)-amine

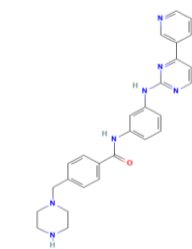


Figure 7. 4-(piperazin-1-ylmethyl)-N-[[3-[(4-pyridin-3-yl)pyrimidin-2-yl] amino] phenyl] benzamide

4. EXPERIMENTS ON CELL MOLECULAR DESCRIPTORS

4.1. An ensemble approach to feature selection

IC50 data from GDSC and chromatin profiling, reverse phase protein array, and metabolomics data from CCLE were downloaded. CCLE data's high dimensionality necessitated feature selection to avoid performance degradation in machine learning models. Ensemble feature selection methods, combining strengths and mitigating weaknesses of individual algorithms, were employed to address bias [14], [16], [21]. A diverse range of techniques was selected to capture various aspects of feature importance, providing a holistic perspective. Python scripting enhanced predictive performance and deepened understanding of genomic landscapes. Five different methods ensured a well-rounded feature selection process, mitigating bias in selection outcomes [14], [16], [20], [21].

5. METHOD

To address the high dimensionality of cell line data and improve accuracy and prediction efficiency, an effective feature selection process was implemented. Employing multiple feature selection methods, each targeting different types of relationships within the dataset, 35 features were selected from each approach. Through max voting, the top 35 features among the total 175 were chosen. Figures 8-10 in the results section detail these selected features for each cell line dataset.

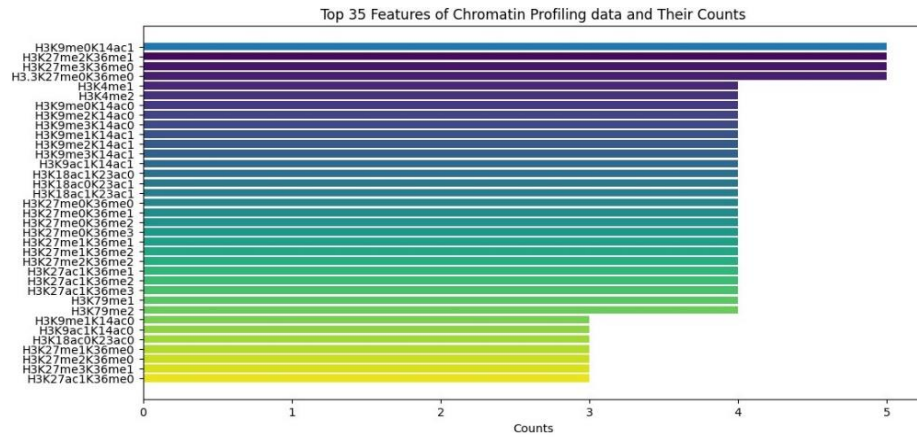


Figure 8. Results of top 35 max voted chromatin profiling data features

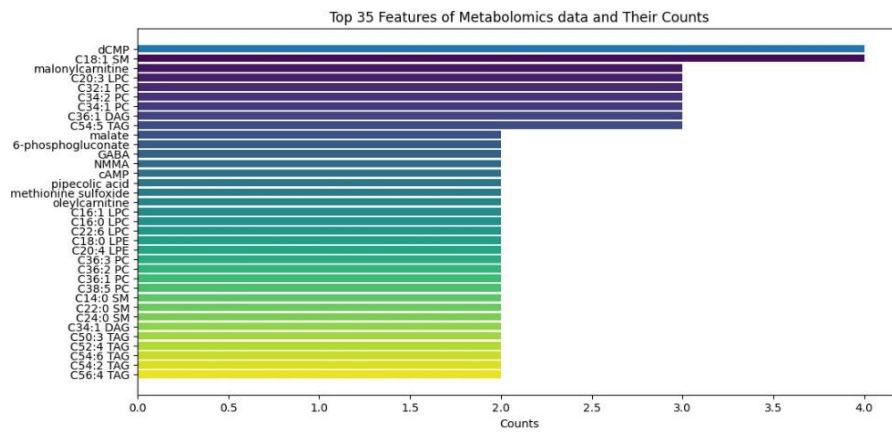


Figure 9. Results of top 35 max voted metabolomics data features

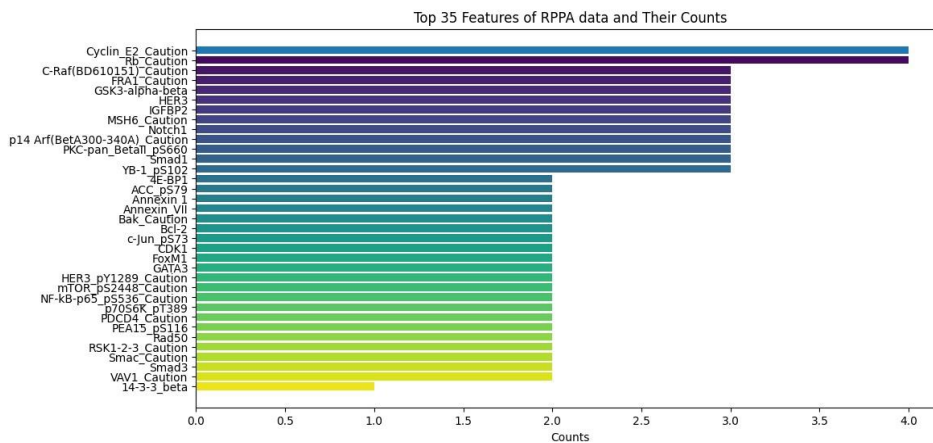


Figure 10. Results of top 35 max voted RPPA data features

The architectural diagram Figure 11 is illustrated below. The dataset underwent testing, training, and validation using various computational methods, including Naive Bayes, generalized linear model, logistic regression, fast large margin, deep learning, DT, RF, gradient boosted trees, and SVM, with different feature combinations. Gradient boost and TabNet showed superior performance, with TabNet outperforming gradient boost by 10%. TabNet [22]-[24] a recently developed deep learning model, specializes in handling tabular data complexities. Its attention mechanism enables selective focus on significant features during both

training and decision-making, capturing intricate relationships within structured datasets. With an encoder-decoder architecture and sequential decision-making, TabNet refines predictions by considering multiple features. Its sparse feature selection capability, driven by the attention mechanism, identifies critical attributes, enhancing both accuracy and interpretability. TabNet's sequential attention approach improves interpretability and facilitates more efficient learning by prioritizing the most significant features.

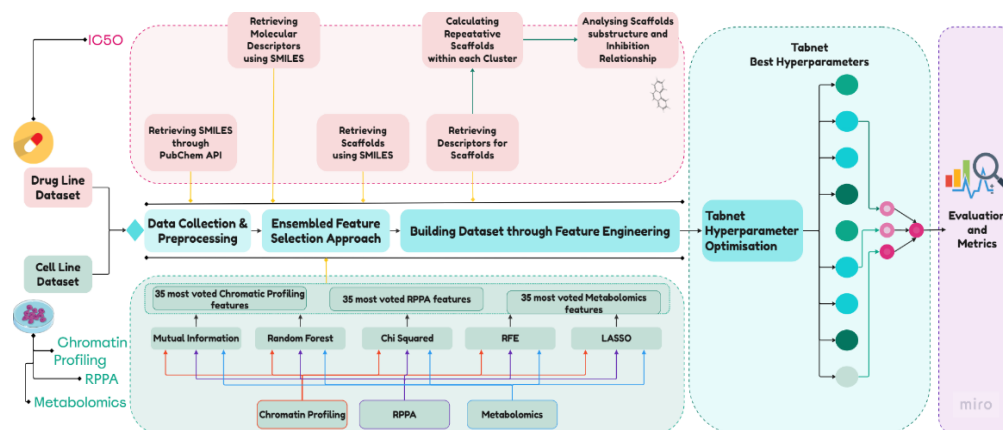


Figure 11. Architectural diagram

6. RESULTS AND DISCUSSION

A Python programming script was used to create an ensemble of feature selection techniques. The top 35 most voted features were used for the model training. The results show that chromatin profiling data had the highest correlation between feature selection methods with at least three of the feature selection methods picking same features and all of the methods able to identify four of the features.

Similarly the correlation between different drug descriptors for each drug cluster was analyzed and plotted. Figure 4 describes the feature importance score chart of few of the highly ranked features that are highly correlated. And also, the results from analysing the scaffolds structure suggest that the structures 4-nilinoquinazoline Figure 5, (4-Benzyloxy-phenyl)-(6-furan-2-yl-quinazolin-4-yl)-amine Figure 6, and 4-(piperazin-1-ylmethyl)-N-[3[(4-pyridin-3-ylpyrimidin-2-yl)amino]phenyl]benzamide Figure 7 are present only in drugs that possess high IC50 values and all contain a shared quinazoline core. This core consists of a benzene ring fused to a pyrimidine ring and serves as a fundamental structural element in these compounds. Despite this common core, each compound possesses distinct aromatic rings, substituents, and functional groups, which contribute to their individual chemical identities. The dataset was first trained on nine different in-silico prediction methods and the gradient boost algorithm consistently produced better prediction results. But Tabnet, which is a deep learning architecture, specifically designed to handle large sized tabular data outperformed gradient boost with much better accuracy. Through optimized hyper parameter tuning techniques, the accuracy was improved further. Both the results in Figure 12 and Table 2 were recorded. The data and code used in this study are available at <https://github.com/JJaniel/Drug-Response-Prediction-for-Pancreatic-Cancer-using-Multiomics-profiling>.

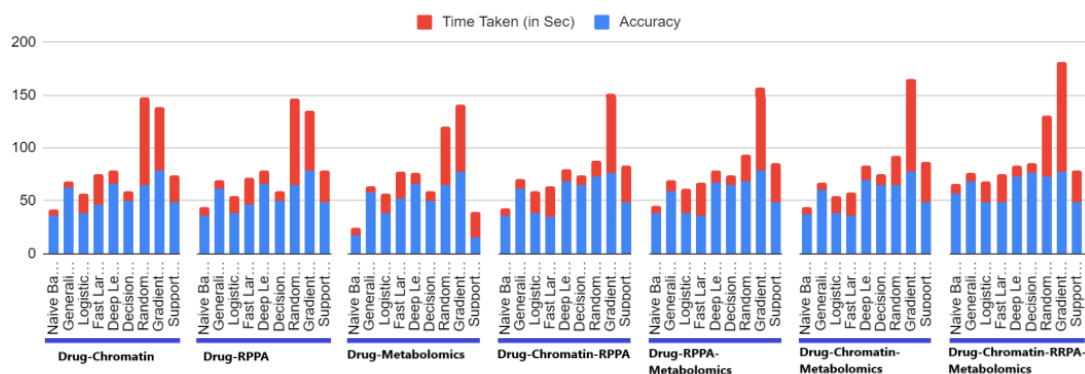


Figure 12. Accuracy metrics on different feature combinations

Table 2. Accuracy metrics on TabNet classifier

Hyperparameters	Iterations									
n_d	11	61	61	61	31	31	31	75	75	75
n_a	51	38	38	13	32	32	32	32	32	32
n_steps	4	4	4	4	2	2	2	5	5	5
gamma	1.36	1.54	1.54	1.45	1.00	1.00	1.00	2.75	2.75	2.75
n_independent	1	2	2	1	1	1	1	4	4	4
n_shared	2	1	1	2	1	1	1	1	1	1
lambda_sparse	0.00085	2.56692	2.56692	0.000395	0.00055	0.00055	0.00055	0.005	0.0054	0.00546
optimizer_params:	0.0021	0.00842	0.0084	0.00897	0.00998	0.00998	0.00998	0.009	0.0098	0.00980
{ lr }								80		
{ step_size }	8	12	12	15	15	15	15	13	13	13
{ gamma }	0.969	0.83	0.83	0.98	0.86	0.869	0.869	0.85	0.85	0.85
mask_type	entmax									
Dropout								0.40	0.40	0.40
Accuracy	81.18	83.5	88.2	82	82.3	85.88	83.52	79.7	85.7	87.0

7. CONCLUSION

Our research, utilizing AI techniques, has successfully utilized drug molecular descriptors, core structures and key cell line features to predict the effectiveness of anti-pancreatic cancer drugs with an impressive accuracy rate. The TabNet architecture emerged as the top performer closely followed by the gradient boost model. Additionally our analysis of the scaffold structures of these drugs has revealed a significant finding: the presence of a quinazoline core among highly inhibiting drug clusters. This shared scaffold feature consistently appears across potent drug clusters and holds profound implications. By integrating these core structures with other relevant drug and cell line features, we can facilitate precise predictions and gain insightful analyses. Our study not only provides accurate predictions of drug response but also offers valuable insights into shared core structures among distinct inhibiting drug clusters. This approach paves the way for deriving and examining insights based on scaffold structures, promoting a progressive approach in precision medicine and future drug design efforts. Through this comprehensive approach, our aim is to drive advancements in cancer treatment strategies ultimately leading to improved patient outcomes and therapeutic efficacy.

FUNDING

This research is funded by the Indian Council of Medical Research (ICMR). (Sanction no: ISRM/12(125)/2020 ID NO.2020-5128 dated 10/01/21).




REFERENCES

- [1] D. Placido *et al.*, “A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories,” *Nature Medicine*, vol. 29, no. 5, pp. 1113–1122, May 2023, doi: 10.1038/s41591-023-02332-5.
- [2] M. G. Dinesh, N. Bacanin, S. S. Askar, and M. Abouhawwash, “Diagnostic ability of deep learning in detection of pancreatic tumour,” *Scientific Reports*, vol. 13, no. 1, p. 9725, Jun. 2023, doi: 10.1038/s41598-023-36886-8.
- [3] M. Tan, O. F. Özgül, B. Bardak, I. Ekşioğlu, and S. Sabuncuoğlu, “Drug response prediction by ensemble learning and drug-induced gene expression signatures,” *Genomics*, vol. 111, no. 5, pp. 1078–1088, Sep. 2019, doi: 10.1016/j.ygeno.2018.07.002.
- [4] P. S. Rajendran and K. R. Kartheeswari, “Anti-cancer drug response prediction system using stacked ensemble approach,” in *Lecture Notes in Networks and Systems*, vol. 436, 2022, pp. 205–218.
- [5] J. C. Costello *et al.*, “A community effort to assess and improve drug sensitivity prediction algorithms,” *Nature Biotechnology*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, doi: 10.1038/nbt.2877.
- [6] C. Wang, X. Lye, R. Kaalia, P. Kumar, and J. C. Rajapakse, “Deep learning and multi-omics approach to predict drug responses in cancer,” *BMC Bioinformatics*, vol. 22, no. S10, p. 632, Nov. 2021, doi: 10.1186/s12859-022-04964-9.
- [7] K. Qiu, J. H. Lee, H. B. Kim, S. Yoon, and K. Kang, “Machine learning based anti-cancer drug response prediction and search for predictor genes using cancer cell line gene expression,” *Genomics and Informatics*, vol. 19, no. 1, p. e10, Mar. 2021, doi: 10.5808/gi.20076.
- [8] J. Lanka *et al.* “ELAFT: an ensemble-based machine-learning algorithm that predicts anti-cancer drug responses with high accuracy,” *Journal of Oncology Research*, vol. 4, no. 1, 2021.
- [9] C. Suphavitai, D. Bertrand, and N. Nagarajan, “Predicting cancer drug response using a recommender system,” *Bioinformatics*, vol. 34, no. 22, pp. 3907–3914, Nov. 2018, doi: 10.1093/bioinformatics/bty452.
- [10] Y. F. Lin *et al.*, “Predicting anticancer drug resistance mediated by mutations,” *Pharmaceuticals*, vol. 15, no. 2, p. 136, Jan. 2022, doi: 10.3390/ph15020136.
- [11] R. Nussinov, C. J. Tsai, and H. Jang, “Anticancer drug resistance: an update and perspective,” *Drug Resistance Updates*, vol. 59, p. 100796, Dec. 2021, doi: 10.1016/j.drug.2021.100796.
- [12] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, “Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data,” in *Pacific Symposium on Biocomputing*, Nov. 2014, pp. 63–74, doi: 10.1142/9789814583220_0007.




- [13] A. Sharma and R. Rani, "Ensembled machine learning framework for drug sensitivity prediction," *IET Systems Biology*, vol. 14, no. 1, pp. 39–46, Feb. 2020, doi: 10.1049/iet-syb.2018.5094.
- [14] N. Hoque, M. Singh, and D. K. Bhattacharyya, "EFS-MI: an ensemble feature selection method for classification," *Complex & Intelligent Systems*, vol. 4, no. 2, pp. 105–118, Jun. 2018, doi: 10.1007/s40747-017-0060-x.
- [15] P. S.i Rajendran and M. Sivannarayana, "Multi head graph attention for drug response prediction," in *Proceedings - 2023 3rd International Conference on Smart Data Intelligence, ICSMDI 2023*, Mar. 2023, pp. 407–414, doi: 10.1109/ICSMDI57622.2023.00078.
- [16] M. Mera-Gaona, D. M. López, R. Vargas-Canas, and U. Neumann, "Framework for the ensemble of feature selection methods," *Applied Sciences (Switzerland)*, vol. 11, no. 17, p. 8122, Sep. 2021, doi: 10.3390/app11178122.
- [17] A. Partin *et al.*, "Deep learning methods for drug response prediction in cancer: predominant and emerging trends," *Frontiers in Medicine*, vol. 10, Feb. 2023, doi: 10.3389/fmed.2023.1086097.
- [18] Z. Bin Alwi, "The use of SNPs in pharmacogenomics studies.," *The Malaysian journal of medical sciences : MJMS*, vol. 12, no. 2, pp. 4–12, 2005, [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/22605952><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3349395>.
- [19] T. Ma, Q. Liu, H. Li, M. Zhou, R. Jiang, and X. Zhang, "DualGCN: a dual graph convolutional network model to predict cancer drug response," *BMC Bioinformatics*, vol. 23, 2022, doi: 10.1186/s12859-022-04664-4.
- [20] Y. Zhu *et al.*, "Ensemble transfer learning for the prediction of anti-cancer drug response," *Scientific Reports*, vol. 10, no. 1, p. 18040, Oct. 2020, doi: 10.1038/s41598-020-74921-0.
- [21] B. Seijo-Pardo, V. Bolón-Canedo, I. Porto-Díaz, and A. Alonso-Betanzos, "Ensemble feature selection for rankings of features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9095, 2015, pp. 29–42.
- [22] Y. Jin *et al.*, "Classification of Alzheimer's disease using robust TabNet neural networks on genetic data," *Mathematical Biosciences and Engineering*, vol. 20, no. 5, pp. 8358–8374, 2023, doi: 10.3934/mbe.2023366.
- [23] S. Arik and T. Pfister, "TabNet: attentive interpretable tabular learning," *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 8A, no. 8, pp. 6679–6687, May 2021, doi: 10.1609/aaai.v35i8.16826.
- [24] K. Kita *et al.*, "Bimodal artificial intelligence using TabNet for differentiating spinal cord tumors-Integration of patient background information and images," *iScience*, vol. 26, no. 10, p. 107900, Oct. 2023, doi: 10.1016/j.isci.2023.107900.
- [25] H. Mary *et al.*, "datamol-io/datamol: 0.12.3," *Zenodo*, 2024. <https://doi.org/10.5281/zenodo.10535844>.

BIOGRAPHIES OF AUTHORS



Janiel Jawahar    has 7 years of Industrial experience in data analytics. He received his M.Tech. in Remote Sensing, B.E in Electrical and Electronics Engineering from Anna University and doing his Ph.D. under ICMR funded project titled "Development of Anti-Cancer Drug Response Prediction Model Using Ensemble Learning for Clinical Applications." at the Hindustan Institute of Technology and Science, Chennai, India. His research interests include machine learning, deep learning, remote sensing, and blockchain. He can be contacted at email: jani.hits23@gmail.com.



Dr. Paramasivan Selvi Rajendran    has 22 years of teaching experience and she received the B.E. degree and M.E in Computer science and engineering from Madurai Kamaraj University, and Ph.D. in Computer science and engineering from the National Institute of Technology (NIT), Trichy, India. Her current research interests include natural language processing, deep learning, and machine learning. She is a Life Member of the Indian Society for Technical Education (ISTE), and Computer Society of India. She authored two books and published research papers in 59 international Journals. She can be contacted at email: selvir@hindustanuniv.ac.in.