2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI) | 979-8-3315-1852-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/IC3ECSBHI63591.2025.10990710

Enhancing Drug Sensitivity Prediction in Cancer Cell Lines Using Multi-Omics Data and Machine Learning

Sara Amjad
Department of Computer Engineering
Aligarh Muslim University
Aligarh, India
saraamjad0302@gmail.com

M. Azhar Aziz
Interdisciplinary Nanotechnology Center
Aligarh Muslim University
Aligarh, India
azhar.inc@amu.ac.in

M. M. Sufyan Beg
Department of Computer Engineering
Aligarh Muslim University
Aligarh, India
mmsbeg@hotmail.com

***Abstract*—This study integrates diverse omics datasets, including proteomic, transcriptomic, and miRNA profiles from cancer cell lines, with drug sensitivity (AUC) data for 285 drugs. We identified that combining proteomics, transcriptomics, and miRNA data provides the best predictive model. Using the k-means clustering algorithm, we grouped cell lines based on these multi-omics data. Subsequently, drug sensitivity (AUC) for 285 drugs was predicted using a 5-fold cross-validation Random Forest Classifier for each cluster separately. The resulting five clusters were thoroughly evaluated, leading to drug suggestions with over 75% accuracy for further investigation. Our findings indicate that high-quality clusters derived from the optimal set of multi-omics features improve the performance of the Random Forest algorithm in predicting drug responses, compared to low-quality clusters obtained from a random set of features. We also compared drug similarities within clusters 0, 1, and 2, analyzing the results with a heatmap. Our findings show improved similarity between drugs in cluster 0 compared to clusters 1 and 2, based on hydrogen bond donor, hydrogen bond acceptor, and molecular weight properties. This aligns with the silhouette scores, which is higher for cluster 0 compared to clusters 1 and 2. These results suggest that, in high-quality clusters, drugs with high prediction accuracy tend to exhibit greater similarity in their properties compared to those in low-quality clusters. This supports a cluster-centric approach in precision oncology, as high-quality clusters, indicated by high silhouette scores, lead to more accurate and consistent drug response predictions.**

Keywords—drug similarity, multi-omics, cancer cell lines, machine learning

I. Introduction

In recent years, the integration of multi-omics data has revolutionized the field of precision oncology, enhancing our ability to predict drug responses and personalize treatment strategies. In our research work, we have detected that when proteomic and transcriptomic data are combined, the predictive performance of machine learning model improves, leading to better drug response prediction. This observation can be validated by a study, in which the performance of drug sensitivity models trained on proteomic or transcriptomic data surpassed genomic-based models for most drugs [1]. Moreover, we found that the results are further improved when proteomic and transcriptomic data is combined with miRNA data of the cancer cell lines. This finding indicates that the combination of proteomic, transcriptomic, and miRNA data for cancer cell lines is highly informative and can be exploited for drug response prediction to find better predictive results compared to other omics combinations. Unlike single-omics approaches [2,3], our method goes further by incorporating multi-omics data, to provide a more comprehensive understanding of cancer heterogeneity. In comparison to kESVR [4], which uses combination of Principal Component Analysis (PCA), k-means clustering, and Support Vector Regression (SVR) to predict drug response based on single omics (gene expression), our approach offers wider analysis by incorporating multi-omics data, allowing for broader understanding of cancer cell lines. Hence, enabling more precise personalized treatment strategies. Use of Machine Learning techniques can boost the processing of data in oncology, these powerful techniques can be applied in diagnosis, prognosis, and treatment modulation [5]. The study affirming influence of miRNA expression on chemotherapy response in cancer treatment [6], highlights the importance of miRNA data in drug response prediction. Furthermore, supporting the inclusion of miRNA data in the optimized feature set for drug response prediction. The predictive models including artificial intelligence have shown high potential in preclinical environments. However, adaptation of these models in clinical setting remains a challenge [7]. In spite of the pandemic, cancer death rates continued to decline from 2019 to 2020 by 1.5%, contributing to an overall reduction of 33% since 1991. Advances in treatment have led to rapid declines in mortality for leukemia, melanoma, and kidney cancer [8]. This progress increasingly reflects the efficacy of innovative therapeutic approaches and the critical role of precision medicine in oncology. Recent studies [7] show that there is a need for better predictive models for personalizing cancer treatment. This underscores the importance of advancing machine learning based models to improve cancer drug response prediction.

II. Data And Implementation

A. Data and Software Details

We performed drug response prediction using various combinations of pan-cancer omics data of cancer cell lines, including proteomic (RPPA - Reverse Phase Protein Array)[9], transcriptomic (gene expression)[9,10], genomic (hotspot mutation, copy number absolute)[9,10], metabolomic[11], and miRNA data[9]. The combination of proteomic (RPPA), transcriptomic (gene expression), and miRNA data provided well-defined clusters and high-accuracy predictions for a comprehensive list of drugs. Specifically, we utilized proteomic (RPPA)[9], transcriptomic (gene expression)[9,10], and miRNA data[9] alongside drug sensitivity (AUC) values for 285 drugs from the Genomics of Drug Sensitivity in Cancer (GDSC1) database[12]. All required data for 1921 pan-cancer cell lines were downloaded from DepMap portal. Scaling was performed, followed by dimensionality reduction using t-SNE (t-distributed Stochastic Neighbor Embedding) for each omics dataset.

The multi-omics and drug sensitivity data were integrated using SQL, and Python was used for implementing clustering and machine learning algorithms followed by graph generation. During integration, cell lines with missing feature values were excluded, and drugs with many missing values were removed. Remaining missing values for drugs were handled by skipping those entries during machine learning algorithm execution. Drug details and similarities were obtained using PubChemPy, a Python library for accessing PubChem data [13], integrated within the Spyder development environment.

B. *Implementation*

This study builds upon our previous work [14], where we demonstrated that multi-omics data outperforms single-omics data and that a cluster-centric approach enhances the performance of machine learning models. In this current research, we have refined our previous findings and further improved the feature set to achieve a more efficient model. We experimented with different feature sets, to obtain the set of features which provides with good quality cluster as well as high prediction accuracy for maximum number of drugs. Table I. represents different feature set with their silhouette scores. Additionally, we have analysed the relationships between drugs within the same cluster using heat maps, represented in Fig 1.

We compared different sets of omics features from cancer cell lines to identify the combination that yields the best performing model for predicting drug response using machine learning. Our analysis indicated that the combination of proteomics, transcriptomics, and miRNA provided significantly better results than other feature combinations. We experimented with different sets of cancer cell line features to obtain the best set of clusters using k-means. We evaluated the clusters based on their silhouette scores and divided the dataset into five different clusters. The result of which is presented in Table I. When all omics features were considered, the best cluster obtained had a silhouette score of 0.3734, while the other clusters had silhouette scores close to 0.1. Using miRNA data alone for clustering, resulted in a silhouette score of 0.7579 for the best cluster. However, using only miRNA features for drug response prediction with machine learning algorithm led to poor result[14]. The same was the case with metabolomics and the combination of transcriptomics and miRNA. Additionally, using different sets of omics features for clustering and drug response prediction negatively impacted predictive performance of the machine learning algorithm for the best performing cluster(cluster 0).The combination of few feature set for clustering and drug response prediction is represented in Table II. The same colour row when compared showed a slight better performance for highest quality cluster where same feature set was selected for clustering and drug response prediction. The similar pattern was observed for different set of features. Therefore, we tested various combinations of omics features mentioned in Table 1 to identify the best fit for both clustering and drug response prediction. Our findings showed that using proteomics, transcriptomics, and miRNA for both clustering and drug response prediction yielded the best results, specifically for cluster 0.The clusters obtained through k-means were saved in a MySQL database as separate tables. Each cluster table (cluster0, cluster1, cluster2, cluster3, and cluster4) was passed to a Random Forest classifier with 5-fold cross-validation to predict drug responses. Drugs with response prediction accuracy above 75% were identified. This procedure was repeated for each of the five cluster tables, and the prediction accuracy for drugs with more than 75% accuracy was represented in a bar graph in Fig 1. We selected cluster 0 with a silhouette score of 0.58, and clusters 1 and 2 with silhouette scores of 0.31 for further comparison. Clusters 3 and 4 were not used for any further study as they had extremely low silhouette scores of less than 0.3. We compared the similarity between drugs in cluster 0 to the similarity between drugs in clusters 1 and 2, representing this comparison through a heatmap. We selected the top 3 drugs which were predicted with maximum accuracy in each cluster by analysing the bar graph in Fig 1. This led to a comparison between high prediction accuracy drugs as well as overcoming any bias that would be induced due to the presence of an unequal number of drugs identified by the machine learning algorithm with high prediction for each cluster. The heatmap analysis, which measured drug similarity based on hydrogen bond acceptors, hydrogen bond donors, and molecular weight, showed that drugs obtained from the well-defined cluster (cluster 0) were more similar than drugs obtained from the low-quality clusters (clusters 1 and 2). These results indicate that the machine learning algorithms that are trained on the high-quality clusters may show higher level of similarity in drugs, which have high prediction accuracy compared to low-quality clusters. Thus, supporting a cluster-centric approach in precision oncology. However, further analysis is required to reach more conclusive result. We evaluated the structural similarity of drugs using three molecular features: molecular weight, hydrogen bond donors, and hydrogen bond acceptors. We calculated the similarity score for each molecular feature for every pair of drugs. The similarity score was calculated as the ratio of the minimum value to the maximum value for the given feature between the drug pair. This method ensured the score ranged from 0 to 1. We calculated similarity matrices for each molecular feature and also an overall mean similarity matrix. These matrices hold pairwise similarity scores between drug pairs. For every drug pair, we calculated the similarity scores for each feature and stored the result in corresponding similarity matrix. The overall mean similarity matrix was obtained by averaging the similarity scores across all considered features. To visualize the similarity data, we generated heatmaps for the overall mean similarity matrix using the Seaborn library. These heatmaps offered an intuitive visual representation of the similarity scores between the drugs, with color gradients indicating the level of similarity. Represented in Fig 1.

III. Data Validation And Discussion

In our study, we performed 5-fold cross-validation for drug response prediction using a random forest classifier. We compared the drug similarity within different clusters. Specifically, we analysed the drug similarity within cluster 0 and compared it to clusters 1 and 2. Our findings indicate that high-quality clusters, such as cluster 0, exhibit higher drug similarity for drugs which have higher prediction accuracy. The bar chart representation in Fig. 1 shows cluster 0 preforms better by predicting maximum number of drugs with accuracy more than 75% compared to cluster 1 and 2, this finding aligns with our previous work [14]. Overall performance of the model in our study was evaluated by considering the number of drugs with prediction accuracy above 75% and their mean accuracy. The detailed analysis shows the robustness of our approach in predicting the drug responses for long list of drugs. These findings validate the potential of machine learning algorithms in transforming cancer treatment and patient outcomes. As well as, it highlights the integration of multi-omics data for improved drug response prediction [14]. In this study, we optimised the feature set and obtained best results with the combination of proteomics, transcriptomic and miRNA data. Our work aligns with prior research, which draws attention to the importance of multi-omics integration in precision oncology [15]. A notable finding in our study is the superior performance of best quality cluster (cluster 0), which showed higher drug similarity compared to low quality clusters (cluster 1 and 2), and identification of optimized feature set that best fits the clustering and drug response prediction model. Our comparative analysis of drug similarities, based on properties such as hydrogen bond donors, hydrogen bond acceptors, and molecular weight, further validates the clustering approach. The higher similarity of drugs within cluster 0, as depicted in the heatmap, indicates that well-defined clusters can provide more accurate predictions for structurally similar drugs. This finding is very crucial for the development of targeted therapies, as it suggests that clustering based on multi-omics data can identify groups of cell lines that respond similarly to specific drugs. However, our study also revealed that using different sets of omics features for clustering and drug response prediction negatively impacts prediction accuracy. This emphasizes the need for a consistent set of features throughout the analytical process to obtain high predictive performance. Though our findings are promising, they have several limitations that need to be addressed in future. Firstly, our study is focused on a specific combination of omics data, thus exploring additional omics features, such as epigenomic, could further improve the accuracy of predictive model. Secondly, inclusion of clinical data, such has treatment histories and patient outcomes, could help in developing a comprehensive understanding of drug responses. Future work should also consider refining the clustering algorithm to handle the complexity of high dimensional multi-omics data more efficiently. In conclusion, our study demonstrates the potential of integrating multi-omics data for drug response prediction. The finding supports a cluster-centric approach in precision oncology, where high-quality cluster derived from comprehensive omics data can lead to more accurate and consistent drug response predictions. It identifies the optimal feature set and also studies the similarity between drug pairs which have high prediction accuracy. The approach could help in improving drug response prediction , ultimately leading to better patient outcomes.

TABLE I. SILHOUETTE SCORES FOR DIFFERENT OMICS FEATURE COMBINATIONS ACROSS CLUSTERS

Omics Features	**Silhouette Score**				
	Cluster 0	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***
Proteomic + Transcriptomic + Genomic + Metabolomic + miRNA	0.3734	0.1475	0.0724	0.1416	0.1536
Proteomic + Transcriptomic + Genomic	0.2630	0.1113	0.0996	0.1114	0.1202
Transcriptomic + Genomic + Metabolomic	0.2162	0.1191	0.0922	0.0997	0.1182
Genomic + Metabolomic + miRNA	0.2833	0.1468	0.1335	0.0912	0.1156
Proteomic + Transcriptomic + miRNA	0.5897	0.3130	0.3132	0.2713	0.2389
Proteomic + Transcriptomic	0.5144	0.2694	0.2552	0.2131	0.2757
Transcriptomic + Genomic	0.1219	0.1522	0.1101	0.1611	0.0834
Genomic + Metabolomic	0.2083	0.1242	0.1495	0.1181	0.1181
Transcriptomic + miRNA	0.7014	0.3532	0.3377	0.2636	0.3051
Proteomic + miRNA	0.5883	0.3239	0.2265	0.2071	0.3055
Proteomic	0.3856	0.2714	0.3147	0.2528	0.2528
Transcriptomic	0.5166	0.3120	0.2777	0.2856	0.1688
Genomic	0.1289	0.1590	0.1552	0.1435	0.1643
Metabolomic	0.6807	0.3850	0.3276	0.3774	0.4956
miRNA	0.7579	0.2940	0.4369	0.4272	0.4676

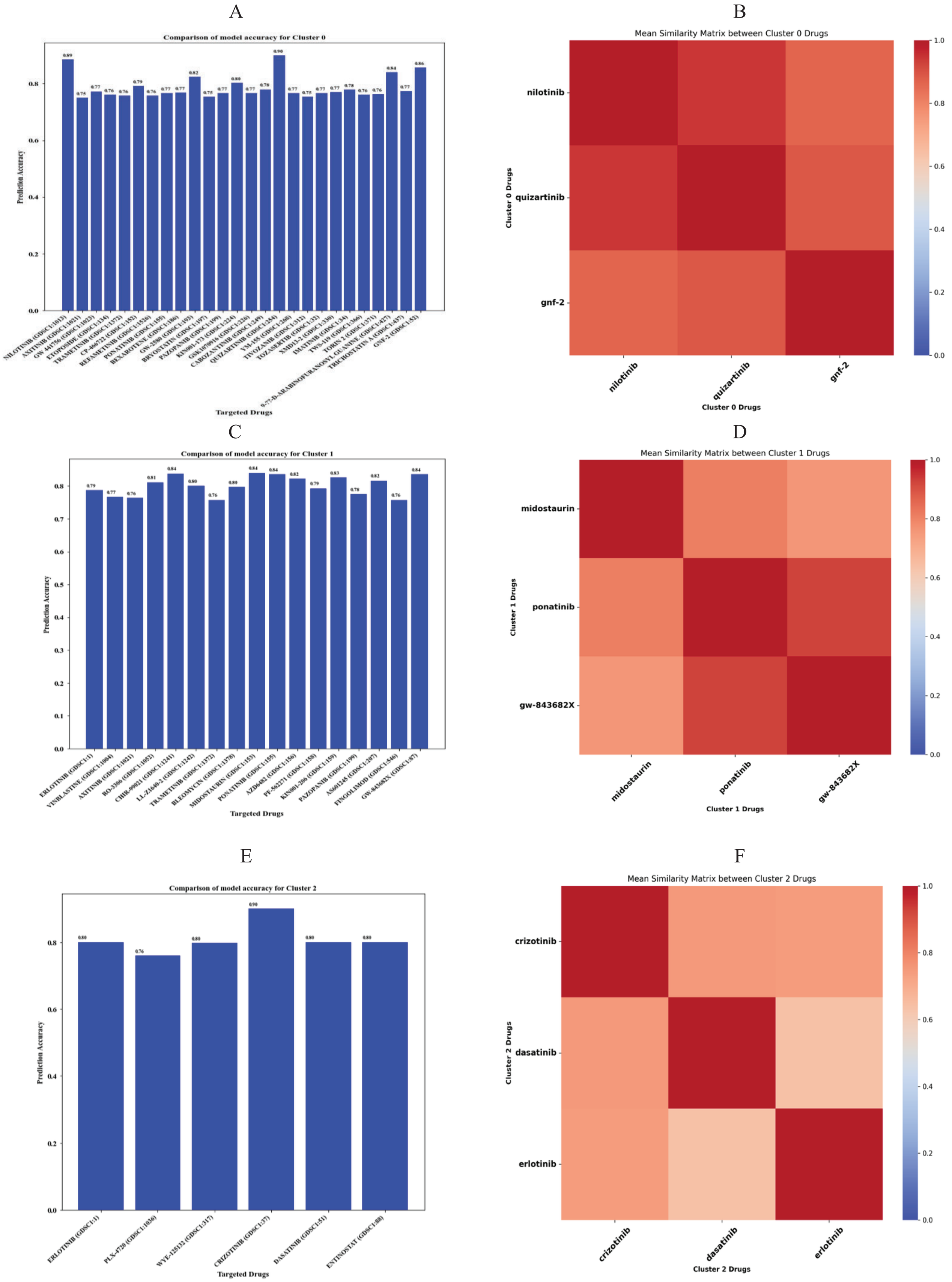

Fig. 1. A) Accuracy for drug response prediction for cluster 0, B) Mean Similarity Matrix between the top three high prediction accuracy drugs of Cluster 0, C) Accuracy for drug response prediction for cluster 1, D) Mean Similarity Matrix between the top three high prediction accuracy drugs of Cluster 1, E) Accuracy for drug response prediction for cluster 2, F) Mean Similarity Matrix between the top three high prediction accuracy drugs of Cluster 2.

IV. Novel Approach to Drug Response Prediction

Our study demonstrates the critical role of selecting appropriate omics features for both clustering and drug response prediction in cancer cell lines. The integration of proteomic, transcriptomic, and miRNA data not only enhances cluster definition but also improves the accuracy of drug response predictions. This study introduces a top-down approach to drug sensitivity prediction, integrating proteomic, transcriptomic, genomic, metabolomic, and miRNA profiles of cancer cell lines with Drug Sensitivity (AUC) data for 285 drugs. Unlike bottom-up approaches that typically focuses on predefined drug lists and predict responses based on single or multi-omics data, we first considered five omics features with the maximum number of drugs for which sufficient data was available. After clustering the cancer cell lines based on multi-omics data using k-means clustering algorithm, we refined the drug list by isolating those with prediction accuracy exceeding 75%. The use of pan-cancer data, multi-omics feature, k-means together with random forest classifier and long list of drugs makes the approach unique and robust. Moreover, focusing on the drugs with high prediction accuracy by eliminating those with accuracy less than 75% leads to reduction of complexity. This approach can be further refined and used for analyzing multi-omics data and developing better understanding of relation between cell line omics features and its utilization in prediction of drug response. Moreover, it can also be used to study relationship between different drug pairs to identify new patterns and relations as done in this study.

V. Conclusion

Our study demonstrates the critical role of selecting appropriate omics features for both clustering and drug response prediction in cancer cell lines. The integration of proteomic, transcriptomic, and miRNA data not only enhances cluster definition but also improves the accuracy of drug response predictions. The k-means clustering algorithm, coupled with Random Forest classifiers, proved effective in identifying high-quality clusters and predicting drug responses with over 75% accuracy. The comparative analysis of drug similarities within different clusters highlighted that well-defined clusters, indicated by higher silhouette scores, may lead to similar prediction accuracy for similar drugs, which have high prediction accuracy value. This finding supports the use of a cluster-centric approach in precision oncology, where high-quality clusters of similar cell lines can enhance the predictive performance of machine learning models. Future research should focus on refining clustering algorithms and exploring additional omics features to further improve the accuracy and reliability of drug response predictions. Additionally, expanding the scope of drug similarity measures and incorporating clinical data could provide deeper insights and drive the development of more effective cancer therapies.

TABLE II. DRUG RESPONSE PREDICTION ACCURACY USING VARIOUS OMICS FEATURE COMBINATIONS FOR CLUSTERING

Omics Datatype		Clusters				
		Cluster 0	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
Proteomic + Transcriptomic + Genomic + Metabolomic + miRNA (Same feature set for clustering and drug response prediction)	Mean Accuracy	0.80	0.78	0.78	0.79	0.82
	Number of drugs with accuracy more than 75%	22	6	8	9	8
Proteomic + Transcriptomic + miRNA (Same feature set for clustering and drug response prediction)	Mean Accuracy	0.79	0.79	0.80	0.77	0.81
	Number of drugs with accuracy more than 75%	25	17	7	5	3
Proteomic + Transcriptomic + Genomic + Metabolomic + miRNA(Clustering) Proteomic + Transcriptomic + miRNA(Drug response prediction)	Mean Accuracy	0.80	0.79	0.77	0.78	0.82
	Number of drugs with accuracy more than 75%	21	7	9	10	4
Proteomic + Transcriptomic + miRNA(Clustering) Proteomic + Transcriptomic + Genomic + Metabolomic + miRNA (Drug response prediction)	Mean Accuracy	0.79	0.79	0.81	0.78	0.80
	Number of drugs with accuracy more than 75%	21	16	10	4	4

Acknowledgment

ChatGPT, an AI language model developed by OpenAI, was used for language correction and structural editing.

References

[1] M. Rydenfelt , M. Wongchenko ,B. Klinger,Y. Yan ,N. Blüthgen, “The cancer cell proteome and transcriptome predicts sensitivity to targeted and cytotoxic drugs,” Life Sci Alliance, vol. 2, 2019.

[2] M. P. Menden, F. Iorio, M. J. Garnett, U. McDermott, C. H. Benes, et al., “Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties,” PLoS ONE, vol 8,2012.

[3] E. A. Clayton, T. A. Pujol, J. F. McDonald, P. Qiu, “Leveraging TCGA gene expression data to build predictive models for cancer drug response,” BMC bioinformatics, vol. 21,2020.

[4] A. Majumdar, Y. Liu ,Y. Lu ,S. Wu,L. Cheng, “kESVR: An Ensemble Model for Drug Response Prediction in Precision

Medicine Using Cancer Cell Lines Gene Expression," Genes (Basel)., vol. 12, 2021.

[5] C. Nardini, "Machine learning in oncology: a review," Ecancermedicalscience, vol. 14, 2020.

[6] M. Konoshenko, P. Laktionov, "The miRNAs involved in prostate cancer chemotherapy response as chemoresistance and chemosensitivity predictors," Andrology, pp 51-71,2022.

[7] R. Rafique,S. M. R. Islam ,J. U Kazi, "Machine learning in the prediction of cancer therapy," Comput Struct Biotechnol J., vol. 19,2021, pp 4003-4017.

[8] R. L Siegel, K. D Miller,N. S. Wagle, A. Jemal, "Cancer statistics," CA Cancer J Clin., vol. 73, 2023 pp 17-48.

[9] M. Ghandi, F. W Huang, J. Jané-Valbuena , G. V Kryukov, C. C Lo ,et al., "Next-generation characterization of the Cancer Cell Line Encyclopedia ," Nature, vol. 569, 2019, pp 503-508.

[10] DepMap, Broad (2023). DepMap 23Q4 Public. Figshare+. Dataset.

[11] H. Li , S. Ning , M. Ghandi , G. V Kryukov , S. Gopal, et al.,"The landscape of cancer cell line metabolism," Nat Med., vol. 25, 2019, pp 850-860.

[12] F. Iorio, T. A Knijnenburg , D. J Vis , G. R Bignell , M. P Menden, et al., "A Landscape of Pharmacogenomic Interactions in Cancer," Cell ,vol. 166, 2016 ,pp 740-754.

[13] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, et al., "PubChem 2023 update,"Nucleic Acids Research, vol. 51, 2023, pp D1373–D1380 .

[14] S. Amjad, M. A. Aziz, M. M. S. Beg, "Precision Oncology: Deciphering Drug Sensitivity Through Multi-Omics Clustering and Machine Learning in Cancer Cell Lines," ICECET, 2024, pp 1-6.

[15] Subramani, S. Verma, S. Kumar, A. Jere and K. Anamika, "Multi-omics Data Integration, Interpretation, and Its Application," Bioinform Biol Insights, vol. 14, January 2020.