Deep Graph Ensemble Convolutional Neural Networks for Drug Response Prediction from Multi-Omics Cancer Cell Lines

Mr. Vikram Kishor Abhang
Research Scholar,
Department of Computer Engineering,
MET'S Institute of Engineering, Nashik

Dr. Baisa L. Gunjal
Research Guide
Department of Computer Engineering,
MET'S Institute of Engineering, Nashik

***Abstract*— Targeted anticancer therapies could be improved with the use of computational approaches for drug sensitivity prediction. Personalized treatment is substantially improved by deep learning designs, which are often disregarded in favor of reduced error and more accurate forecasts. Cancer Genomics Study on Drug Sensitivity (GDSC) is used in this investigation. The recently released GDSC2 dataset, which employs a better test and medication screening approach than GDSC1, is the center of our study. A total of 809 cell lines and 198 medications make up the GDSC2 dataset. Out of all the medications that have over 750 cell lines, only 86 have been studied. We only take into account gene expression data for all drugs and cell lines that have developed full resistance to the medication. Implement some of the preprocessing steps as follows: Imputation, Harmonization (Normalization), Data Augmentation, and Feature Selection. The molecular and genetic characteristics of multi-omics data increase medication response prediction in our research model. The proposed research model improves structural and genomic feature representations using multi-scale graph feature representation. To learn critical gene interaction, the proposed approach uses transfer learning. Transfer learning uses gene oncology biological domain knowledge to gather gene interaction information. GCNNs are useful in drug discovery, including drug-target interaction prediction. These deep learning models use circular fingerprint concepts to extract useful feature representations from input during training. For the purpose of categorizing multi-omics data as either "sensitive" or "resistant," the suggested study model use a hybrid graph-based deep learning algorithm called an Ensemble Convolutional Neural Network to determine the best threshold values. Data such as F1-Score, Area under the Precision-Recall Curve, Sensitivity, Specificity, Precision, and False Positive Rate will be evaluated in order to validate the proposed research. To surpass top-tier deep learning models, we apply graph-based Ensemble Convolutional Networks (G-ECN) that include structural and genomic drug features. We also show high generalizability on a separate Cancer Cell Line Encyclopedia dataset and explainable findings in case studies that match existing knowledge.**

Keywords*— *Deep Graph Ensemble Convolutional Neural Networks, Drug Response Prediction, Multi-Omics Cancer Cell Lines, Cancer Genomics Study on Drug Sensitivity, and structural and genomic drug features

I. INTRODUCTION

Because there is a vast array of treatment options available in modern oncology, it is crucial to choose the one that will work best for each individual patient. A promising approach to improving the efficiency of these treatment choices is the integration of multi-omics profiling with prediction models powered by artificial intelligence. Nevertheless, the lack of a significant number of annotated samples and the very high complexity of the datasets continue to impede these promising advances [1]. Predicting how drugs will react in cancer cell lines may help doctors create individualized plans for each patient. Nevertheless, it is still difficult to forecast how a medicine would work [2]. Driving forces Cancer is a complicated illness that kills a lot of people throughout the world. Even when dealing with the same form of cancer, treatment approaches might differ from patient to patient. Treatment of various cancers, reduction of healthcare costs, and improvement of recovery rates may all be achieved via the use of precision medicine in cancer. Personalized cancer therapy is now a reality thanks to machine learning algorithms that can anticipate how drugs will interact with tumor and medication data [3]. Driving forces More than one million people will lose their lives to cancer in the European Union in 2022, making it one of the deadliest illnesses in the world. Cancer cells may develop resistance to various chemicals due to the fact that one tumor might have a wide variety of cell types with varying genotypes. Anticancer medications also have the potential to cause serious adverse effects, which might endanger the health of patients. [4]. In addition to a complete knowledge of cancer mechanisms and medication action mechanisms, it is crucial to accurately forecast how various cancer cell lines will react to medicines in order to design novel anti-cancer treatments [5]. Cancer develops when there are changes to the DNA sequences, which are the fundamental units of proteins. Mutations cause aberrant and sometimes deadly tissue development because proteins regulate cellular structure and function. Germline mutations account for 10% of cases of breast cancer, the most frequent malignancy in women. Many variables, including variations in genetic interactions, environmental exposure, and cancer stage, contribute to the fact that different people have diverse reactions to the same therapy [6]. Gene expression, the number of copies variation, and methylation indicators are the basis of many approaches to screening medication response. In contrast, neural networks as well as random forests are only two of the numerous machine learning methods that have recently been used to forecast how medications would react in cancer cell lines [7]. The therapeutic result may vary among individuals with an identical cancer type even when they have the same therapy, since cancer is a very heterogeneous disease. Patients may be able to save money and have a better chance of survival with the use of anticancer medication response prediction to

create individualized treatment plans. There has been a lot of buzz around approaches based on graph neural networks as of late, thanks to their remarkable performance on the drug response forecasting challenge. Most of these methods, however, use graph convolution to analyze bipartite graphs involving cell lines and drugs, without taking into account the fact that the two sets of data are fundamentally different [8]. Personalized medicine and the discovery of new drugs give rise to the issue of drug response prediction. Using the multi-omics information that is now available for over a thousand cancer cell lines and tissues, deep neural networks have been employed to enhance drug response prediction [9]. Prior models found that graph depictions of pharmacological properties outperformed strings or numbers when it came to learning. When data from many cell lines are combined, it improves the accuracy of response to drugs predictions. However, these models did reveal issues with extracting pharmacological features from graph representation and incorporating redundancy information from multi-omics data [10]. An important problem in pharmaceuticals and medicine has been the achievement of speedy discovery of anticancer medicines and their exact use. While in vitro drug evaluation and in vivo investigations are crucial to the development of novel anticancer medications, they are both resource-intensive and time-consuming. Predicting drug-cancer cell lineage response using classic machine learning approaches is challenging because such algorithms often rely on a single data source, which restricts their ability to understand cancer cells in their whole [11]. Anticipating how a cancer cell line will respond to a therapeutic drug is a significant area of research in modern oncology that can help with tailored tumor treatment. Although several machine learning methods have been developed for CDR predictions, integrating diverse data on cancer cell lines, medications, and their recognized responses remains a significant challenge [12]. Research often makes use of cancer cell lines as in-vitro tumor models. Genomic data and large-scale medication screening have allowed for faster right-drug selection for cancer patients. The key to success is accurate medication response prediction. Precision medicine relies on cancer multi-omics data to forecast medication response, but few approaches can integrate and effectively locate the fundamental low-dimensional manifold of this high-dimensional data set [13]. The foundation of individualized cancer therapy is reliable forecasting of CDR, which has been a longtime difficulty in contemporary oncology. To make CDR predictions, current computational methods model interactions between whole drugs and cell lines, but they don't account for the fact that interactions might be due to a small number of finer-level ‘subcomponents,' like the drug's privileged substructures or the cancer cell's gene signatures, leading to inexplicable predictions [14]. Driving forces predicting how a patient will react to a medicine is difficult, even while it is known that different people react differently to the same medication. There have been proposed solutions using single omics observations and networks that allow for the incorporation of molecular interactions into reasoning. Nevertheless, integrating the abundance of data found in many omics levels remains a complex task [15]. Driving forces Precise estimates of medication response are necessary for personalized cancer therapy. While current deep learning algorithms have shown some encouraging results, precision medicine requires even more accurate results. Drug geometrical and topological information may both enhance prediction accuracy [16]. In this age of precision medicine, there is an immediate need for anti-cancer treatment response data from cell lines in order to make personalized medical decisions. It is almost hard to do measurements using wet-experiments due to the high costs, lengthy procedures, and limited applicability. The development of reliable computer models for drug-cell line response prediction could serve as a springboard for other studies [17].

A. Research Motivation

In order to direct the development of anticancer medications, precise prediction of cancer therapy response is essential. Uncertainty about treatment effectiveness and patient heterogeneity make cancer drug response prediction a difficult topic in contemporary customized cancer therapy. In order to differentiate between the interactions of two chemical atoms, previous efforts have neglected to account for the information contained in a drug molecule's chemical bonds. The results of interactions between drugs and cancer cell lines may also be directly affected by this data. There is evidence that the medicine's properties and the patient's genetic traits significantly impact the outcomes of cancer therapy response. For this reason, it is critical to increase the prediction accuracy by developing quick, thorough, and accurate methodologies for chemical feature extraction and genomes integration.

B. Research Contributions

Our solution to these problems is the G-ECN, a new kind of multi-source heterogeneous graph convolutional neural network. Several subnetworks are devoted to sophisticated feature extraction from drug and gene multiomics data, respectively, and this architecture includes a generic data improvement module that utilizes sequence recombination and an updated complete graph convolutional neural network using edge features. Predicting IC50 sensitivity values of CCLs to medicines is a regression problem that is accomplished by feeding the aggregated feature set into a one-dimensional convolutional network. Following is a synopsis of the work's primary contributions:

- By concurrently updating the chemical atom (node) and bond (edge) embeddings, we construct a hybrid graph convolutional network. To guarantee that in-depth knowledge about the drug is entirely remembered, the co-updating approach offers a fresh viewpoint for learning a thorough drug representation.
- The structural and genetic aspects of the multi-omics data are helpful in enhancing drug response prediction in this suggested research paradigm. The suggested study model builds multi-scale graph feature representations to improve architectural and genomic feature representations.
- In order to learn crucial gene interactions, the suggested model incorporates the transfer learning idea. To better understand gene interactions, transfer learning allows us to use what is already known in the biological field of gene oncology. Predicting drug-target interactions is only one of

several drug development tasks that GCNNs have proven useful. In order to train, these DL models use the same circular fingerprinting concepts to extract useful feature representations from the input.

- A Deep Learning Model Based on Hybrid Graphs: To effectively forecast pharmacological response, the suggested study model constructs a hybrid graph-based deep learning simulation (Ensemble Convolutional Neural Network) that integrates domain expertise with crucial graph properties (genome and structural) from multi-omics data, and then divides this data into two categories: Sensitive and Resistant, determined by the optimal threshold values.
- In order to evaluate the performance of the suggested study activity, evaluation measures including sensitivity, specificity, and precision, false positive rates, F1-Score, and area under the curve of precision and recall will be assessed. Using G-ECNs trained with a combination of structural and genomic drug characteristics, we get better results than top-tier deep learning models. In addition, our findings are easily explicable and show strong generalizability on a distinct datasets from the Cancer Cell Line Encyclopedia. This is supported by case studies that are in line with what is already known.

II. Related Work

To forecast the anticancer medication response of individual patients using three forms of multiomics data, a new deep learning-based approach was suggested in [18]. In order to apply convolutional neural networks—which are able to represent very complicated correlations between variables while being resilient to the elevated dimensionality of the inputs—in the proposed DeepInsight-3D method, structured data is first converted to images. Specifically, they show that that approach may make good use of extra picture channels to accommodate information from many 'omics layers while clearly recording their link. When compared to two other suggested state-of-the-art approaches, DeepInsight-3D achieved superior performance. Better tailored treatment techniques for various tumors may be possible in the future because to these advancements. By combining copy number variations, gene expression, morphological photos of cell lines, with chemical structures of medicines, MMCL-CDR creates a multimodal approach to cancer treatment response prediction. Multimodal Contrastive Learning for Cancer Drug Responses is proposed in [19]. The goal of MMCL-CDR is to align cancer cell lines across different data modalities by learning cell line representation from omic and image information. The CDR prediction will then be enhanced as a result. Their method outperforms other cutting-edge methods in CDR prediction, according to extensive research. The experimental results demonstrate that the system may develop a more precise image of cell lines by combining morphology and multiomics information obtained from cell lines, which enhances the efficiency of CDR prediction. Medicine Effectiveness Leveraging Forked and Specialized systems is a new technique for predicting how a medicine will act, which was created in [20]. Using structural data on over 200 compounds and multi-omics data from over 65 cancer cell lines, their model excels in predicting drug sensitivity. Additionally, they examined the viability of using single-cell expression data for drug response prediction. During validation using datasets that included unknown cell lines or medications, DELFOS beat other modern algorithms on many error and correlation measures. All things considered, DELFOS effectively utilizes multi-omics data to predict the potential responses of thousands of drug-cell line combination to a certain therapy. In order to forecast the efficacy of kinase inhibitors in treating various cancers, CancerOmicsNet[21] employed a graph neural network using sophisticated attention propagation techniques. Unifying diverse information like as genomes, biological networks, inhibitory profiling, or gene-disease correlations into a single graph structure, CancerOmicsNet highlights the complicated nature of cancer overall. Following precise tissue-level cross-validation, CancerOmicsNet outperforms prior techniques with an area under the receiver operating characteristic of 0.83. One area where CancerOmicsNet excels is in making predictions about inhibitor therapeutic effects and cancer cell line responses to new data. According to the study, a drug response prediction method based on a graph neural network was used to forecast the sensitivity of HER2-positive breast cancer cell lines to several drugs [...]. The GDSC dataset was implemented to train the model. You may find the drug sensitivity of the many drug-cell line combinations in that dataset. The GDSC library's medications were transformed into a graph architecture using the RDKit program. The chemical bonds among molecules are represented by the edges in that graph, while the molecules' contents are represented by the nodes. The initial node embedding specifies the chemical properties of all elements. A multi-omics depiction of cancer cell line was obtained by consulting the CCLE library. The DRP model was trained using 197,818 drug-cell line combinations. Seven distinct HER2-positive breast cancer cell lines were used to evaluate the model's accuracy after training with the matching drugs from the training dataset. Using patient data and genetically categorized breast cancer, a system has been developed in [23] that can tailor therapy according on the individual's needs. For each given cell line and drug, that neural network system calculates the IC50 by analyzing omics characteristics obtained using an auto-encoder. Following the K-means clustering of IC50 readings, a threshold value is used to classify that IC50 score as responder or non-responder. With a performance level of 0.80, their model surpasses its predecessors. In [24], the authors present a method for predicting how a drug will work in a given situation by combining 1D CNNs with an attention mechanism and pathway networks. That method takes into account the topological the natural world of the pathways in order to identify the subpathways that have a strong relationship with drug response; then, it uses that feature to train CNNs to predict drug response. Consequently, the output results will reflect the occurrence probabilities of these two groups. In that study, the average identification accuracy was 84.6% utilizing five-fold cross-validation, which is 4.5% more than the direct random forest method

for AUC-based medication prediction. The findings show that a one-dimensional convolutional neural network using an attention mechanism is the most effective method for predicting how low-grade glioma patients would react to medications. The NIHGCN method, which relies on local interactions and heterogeneous graph convolution networks, is proposed by the author as a means of anticipating the overall reaction to anticancer medications [25]. A heterogeneous network is first constructed using drugs, cell lines, and data on known drug responses. An interaction model requires the linear conversion of drug molecular fingerprint and cell line gene expression before they can be included as node attributes. The interaction module consists of a layer for interacting with the neighborhood and a layer for doing graph convolution networks in parallel. At the node level, the PGCN layer aggregates characteristics from neighbors using graph convolution, while at the element level, the NI layer considers interactions. Predictions of drug response are made by computing linear correlation coefficients between cell line with drug feature representations. The most recent publications on state-of-the-art deep learning techniques are reviewed and summarized in [26]. Even while deep learning has come a long way in predicting how drugs will work, it still has a long way to go before it can handle drugs that aren't part of the training dataset. Specifically, their drug blind test revealed that the similarity-regularized matrix factoring technique outperformed all five of the deep learning methods that were studied. They describe the difficulties of using a deep learning technique to anticipate how a medicine will react and provide novel ways that deep learning might be combined with well-established bioinformatics studies to address these difficulties. Improved medication representation and less data duplication were goals of the GraTransDRP deep learning model suggested in [27]. Improving drug representation extraction with the Graph transformer was the first step. The next stage included teaching convolutional neural networks to identify techniques, transcriptomics, and mutations. Nevertheless, transcriptomics features may have dimensions of up to 17,737. Therefore, transcriptomics properties were KernelPCA-ed to flatten and improve their presentation before being fed into the CNN model. Finally, a response value was forecasted by integrating drug and omics data using a completely linked system. According to experimental data, their model outperforms state-of-the-art methods like GraphDRP and GraOmicDRP. To predict the efficacy of cancer medications, the GADRP model was presented in [28]. It depends on GCNs and AEs. By using a layered deep AE, they are able to extract low-dimensional representations from the properties of the cell lines. After that, they account for data on drug, cell line, and DCP similarity and construct a sparse drug cell line pair network. Following this, the over-smoothing problem may be reduced by learning DCP features using an attention-based GCN based on the first residual plus layer. Lastly, a fully connected network is used for prediction. The benchmarking findings on five datasets demonstrate that GADRP can achieve better prediction performance than baselines on all metrics. Trials demonstrating the capacity of GADRP to anticipate outcomes, such as unexpected DCP reactions, drug-cancer tissue links, and drug-pathway correlations, stand out. Provided a graph neural network method for CDR prediction with contrastive learning in [29]. With the use of drug chemical structures, known cancer cell line-drug reactions, and multi-omics profiles of cancer cell lines, GraphCDR constructs a graph neural network to enhance the CDR prediction's generalizability. To ensure consistency, a multi-task learning model is used, which includes a contrastive learning task. The high-accuracy CDR prediction relies on a mix of biological features, known cancer cell line-drug reactions, and contrastive learning; GraphCDR surpasses state-of-the-art methods in computational studies conducted in several experimental configurations. The ablation investigation reveals the basic components of GraphCDR. Experimental results suggest that GraphCDR has predictive capabilities and could be useful in guiding the selection of anti-cancer medications. A subcomponent-guided deep learning method for interpretable CDR prediction, SubCDR aimed to identify the most important subcomponents influencing response outcomes when it was first described in [30]. To put it technically, SubCDR uses a cascade of deep neural networks to deconstruct CDR prediction into its component parts, identify subcomponent pairwise correlations, and harvest a variety of functional subcomponents from drug and cell line profiles. Having an interaction form among subcomponents can give a traceable path, making it apparent which subcomponents offer extra information to the response result. Their extensive computational testing on the GDSC dataset proves that SubCDR is superior than cutting-edge methods for CDR prediction. By finding several predicted cases, they demonstrate how effective SubCDR is in determining response-driving subcomponents and using these subcomponents to find new therapeutic drugs. In [31], the DrDimont method was developed; it predicts pharmacological reactions by differentially evaluating multi-omics systems. It may be used to compare two scenarios and then use that information to forecast how drugs would react differently. The molecular interactions are the main focus of Dr. Dimont. It starts with omics-layer correlation to build condition-specific networks, which are then aggregated into heterogeneous multi-omics molecular networks. Integrative results are guaranteed by a new semi-local, path-based integration stage. By comparing the integrated networks that are distinct to each circumstance, differential predictions may be produced. It is possible to obtain the molecular differences that cause high differential drug scores, which means that DrDimont's predictions may be explained. Dual Branch Deep Neural Matrix Factorizing (DBDNMF) was proposed by the authors of [32] as a solution to these issues. In order to reconstruct the partially visible matrix, DBDNMF employs a multi-layer hidden neural network that learns a latent model of cell lines and drugs from flexible inputs. Results from experiments conducted on datasets such as the Cancer Cell Line Encyclopedia and the Genomics of Drug Sensitivity in Cancer demonstrate that the drug prediction algorithm is stable and reliable, surpassing state-of-the-art drug response forecasting approaches. In accordance with earlier research, hierarchical clustering reveals that cells from the same tissue subtype have a same pattern of response, and that medicines with comparable response levels target similar signaling pathways.

III. Problem Statement

Every year, cancer claims the lives of millions throughout the world. Although there have been several medications developed and made available for cancer treatment, the disease is still mostly unsolved. There is great promise in using computational predictive models to study and treat cancer. This might lead to better drug development and more personalized treatment plans, which in turn could reduce tumors, alleviate pain, and increase patients' lifespans. Recent research using deep learning algorithms to predict cancer patients' responses to pharmaceutical treatments has shown promising outcomes. To represent drug compounds, earlier efforts either used string-based approaches like SMILES or graph-based methods. Nonetheless, data gleaned from these two approaches may supplement one another in the quest for new pharmaceuticals. Learning a better possible drug representation is possible with the full use of both pieces of information. The wealth of information available in multi-omics data was largely disregarded in earlier studies, which relied on a single genetic profile to characterize cancer cell lines. Genomics multiomics features still have a lot of room to grow. There is a lack of integration and use of some genetic traits that have shown to be extremely informative with cancer. Four significant obstacles remain for current deep learning methods that rely on multi-omics data. To begin, one of the most important steps in using models to predict drug sensitivity is learning new information characteristics from omics data. The problem is that biomolecular datasets are often very feature-heavy yet sparse in sample size, making them high-dimensional datasets. Overfitting is a major concern when dealing with deep learning models. The second issue is that researchers have to put in a lot of time and effort to figure out how deep learning models work since they are opaque. Clinical success with black-box techniques is elusive since accurate diagnosis relies on doctors' familiarity with the disease's fundamental characteristics. Third, a major obstacle in multi-omics analysis is figuring out how to combine various data types; initial integration and delayed integration are the two basic approaches. A significant percentage of misaligned gene points are purposefully eliminated to assist feature fusion in the models that fuse the depictions of features learnt from every omics before classification, which causes data loss issues. This brings us to our fourth point: current multi-omics medication response prediction algorithms have disappointing outcomes and might be improved.

To begin, we clarify what the job of CDR prediction entails. The log-normalized half-maximal inhibitory concentrations (IC_50) values for 22490 cell line-drug pairings were obtained from GDSC and used to depict the drug response of cancer cell lines. Next, following earlier research, we classify the IC 50 values according to a cutoff derived from the stated maximum screening concentration. By using this classification system, we may divide the correlation between cell lines and pharmacological reactions into two extremes: "sensitive" and "resistant."

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{response}_{ij} \leq \text{threshold}_j \\ 0 & \text{other} \end{cases}$$

where response $_{ij}$ is the half-maximum inhibitory concentration (IC_50) among the i-th cell line and the j-th drug, where threshold _j stands for the sensitivity threshold of the j-th drug. After constructing 7809 sensitive pairs and 14681 resistant pairings using 254 cell lines and 311 medicines, we get the final result.

The analytical model f is used to forecast the reply r of cancer c to the therapy by medication d in a DRP model, which may be expressed as r=f(d,c). A CNN architecture with weights acquired by backpropagation implements the function f. In order to forecast the response, this formulation is required for pancancer plus multi-drug ^1 prediction models, which need representations of both the medication and the cancer. As an exception, there are drug-specific models that are developed to provide predictions for a particular medication or group of pharmaceuticals, such as those that have a common mechanism of action (39). The formula for these models, which learn only from cancer traits, is r=f_D (c). Another kind of model is the multi-task learning model, which, given a representation of cancer as an input, may provide several outputs, each of which can yield a prediction for a different medication. With the multi-task formulation, the model may learn from more drug response data and take advantage of drug-specific similarities, leading to better overall generalizability than with drug-specific models.

IV. Proposed Work

A. System Model

Lack of full, accurate, skew, and amount of data is one of the biggest obstacles to efficient DRP. As a solution, we developed a flexible, multimodal neural network that can learn from different types of cell line profile data separately before intelligently combining them to forecast how each cell line would react to each medication that has been tried. The main problem of data sparsity is solved by this method. This enables the system to learn from datasets that include varying subsets of molecular profile data, even when the cell line and medication combinations are different. The drug response target prediction pipeline consists of three stages: preprocessing, graph creation, feature extraction and fusion, and prediction. Instead of selecting hyperparameters at random, we include these procedures into a hyperparameter optimization strategy that seeks for better omic processing module design features.

B. Data Preparation

The first stage in developing a prediction model is often data preparation, which calls for knowledge of statistical methodologies and bioinformatics. In this phase, DL framework APIs are used to organize training and test sets made up of aggregated heterogeneous data types that have been preprocessed. The drug response dataset that was constructed using N samples, represented by $S = \{d, c, r\}_{i=1}^{N}$,

covers drug (d), cancer (c), and response (r) representations. Typically, while generating DL models, it is preferred to use bigger datasets because, as shown using various cell line datasets, prediction generalization is predicted to increase with a higher number of training examples. The significance of data preparation has been further emphasized by recent data-centric research, which argues that, in addition to dataset size, effective information representations and the selection of an appropriate training set are equally crucial for improving predictions.

C. Genomics of Drug Sensitivity in Cancer (GDSC) database

We use the Genomics of Drug Sensitivity in Cancer (GDSC) database, which contains cancer therapeutic genomic data, in this investigation. For the purpose of drug sensitivity prediction, this dataset is extensively examined using statistical and machine learning methods. Models based on cell line and drug similarity, models to forecast drug sensitivity and target identification using lasso and elastic networks, and quantitative structure-activity relationship (QSAR) research utilizing kernelized Bayesian matrix decomposition are just a few examples.

We take a selection of GDSC mutation information, cell line annotation, and drug IC50 information, which includes targets, signaling pathways, information on point mutations and copy number variation, various phenotypes for 518 oncology medications in 988 cell lines, and IC50 values for a few genes. Out of the 35 medications available in the GDSC database, we chose 14 to be FDA-approved targeted therapies, 16 to be treatments with unambiguous targets but not yet FDA-approved, and 5 to be non-specific cancer treatments without goals. These experimental pharmaceuticals will be used in a drug sensitivity research. The pancreatic cancer dataset—which includes RNA expression data for 43 organoids tested against 26 drugs—was retrieved from the Pancreatic Cancer PDO Library (PCPL). Values of AUC were used for the purpose of measuring and reporting the drug's effectiveness.

D. Preprocessing

a. **Imputation:** Multi-omics data contains missing values, which can affect the accuracy of the prediction model. Conducting imputation for missing values is necessary during the analysis of multi-omics data. Imputation techniques can be utilized to fill in missing values in the data.

b. **Harmonization (Normalization):** In data preparation of multi-omics data, harmonization is applied as an initial process for making the data in a consistent format and controlling data quality measures. This process helps to minimize the batch effect in multi-omics data.

c. **Data Augmentation:** The issue of unbalanced class distributions and its effect include, including over-fitting, will be addressed through advanced augmentation techniques. This pre-processing step makes further data analysis more productive by balancing data

d. **Feature Selection:** The feature selection technique will be applied to recognize the most informative features for drug response prediction. In this model, the feature selection technique helps deal with multi-omics data's high dimensionality.

e. **SMILES**

According to the similarity principle, pharmacophores or chemical structures that induce gene expression similarity in cell lines are either structurally identical or have partial overlap. So, to get the substructure information out of the drug SMILES sequences that are entered, we use a transformer model. SMILES is a molecular encoding technique that integrates several pieces of information about molecules, including their connectivity structure, atomic and bond configurations, ring size, stereochemical properties, and more. The chemical structure and connectivity of several SMILES representations for the same molecule may be consistent. This discovery paves the way for data augmentation in molecular property forecasting, which in turn improves the model's ability to mine SMILES's deep knowledge for task-relevant chemical traits. We use SMILES permutation to increase the number of SMILES in the job. A new drug-genome pair is formed when every augmentation data molecule is re-entered into the training process together with the genomic data that corresponds to the original molecule. For the sake of training, we consider the new data set to be an independent instance. To prevent data leakage issues with data augmentation, we limited our operations to the drug compounds in the training set.

F. Graph Construction

It is possible to naturally portray a medication as a graph due to its unique chemical structure. The vertices stand for chemical atoms, while the edges indicate bonds. As a result, the medication set may be simply represented as $\{\mathcal{D}_n = (\mathbf{V}_n, \mathcal{A}_n)|_{n=1}^{M}\}$ where $\mathbf{V}_n \in \mathbb{R}^{N_n \times C_{node}}$ and $\mathbf{A}_n \in \mathbb{R}^{N_n \times N_n \times C_{edge}}$ comprise the nth drug's edge information in the adjacency matrix and feature matrix. N_n is the atomic number of the nth medicine, C_{node} and C_{edge} each node and edge have a certain amount of feature channels. Atomic properties are represented by the rows of the characteristic matrix. An edge exists between the nodes demonstrated by the horizontal and vertical coordinates, and the element is the attribute of that edge if one of the neighboring matrix elements is a one-hot vector. An edge does not exist if it is 0. We describe a new CGCN design which handles node and edge information concurrently to fully retain the deep details of the drug graph representation, making up for the UGCN's missing edge information. With a layer-wise procedure, the CGCN implemented to the i-th drug can be described as f(D_i):

$$V^{(l+1)} = \mathrm{Update}\left(V^l, \sigma\left(E_k^{(l)} V^l W_k^{(l+1)} + b_k^{(l+1)} \mid k \in (1, \ldots, C_{edge})\right)\right)$$

where V^l is the node feature matrix in the l th layer, $\sigma(.) is the activation function.$ W^{l+1} and b^{l+1} are variables that can be taught. E_k^l represents the edge feature matrix for the kth edge. With this perspective, the adjacency matrix A_v is really just a collection of subadjacency matrix

structures, the exact number of which is defined by the edge feature dimensions. An expression that captures a matrix of adjacency A_n is:

$$A_n = \left\{ E_{n,k} \middle|_{k=1}^{e_{edge}} \right\}, E_{n,k} \in \mathbb{R}^{N_n \times N_n}.$$

In addition, there are two processes to upgrade the edge features. A vector e_(i,j) representing the connection among nodes i and j is defined in the first step :

$$e_{ij}^{(l+1)} = \sigma\left(\left(V_i^{(l+1)} \parallel V_j^{(l+1)} \right) W_{ij}^{(l+1)} + b_{ij}^{(l+1)} \right).$$

$e_{i,j}^{(l+1)}$ is the l+1th layer relation vector between nodes i and j, $\sigma(.)istheactivation$ function, $V_i(l)$ and $V_i(l+1)$ layer l's and layer l+1's node feature vectors, respectively, W_{ij}^{l+1} and b_{ij}^{l+1} are variables that can be taught. After that, for each edge that connects nodes i and j, we update its feature vector E_ij by:

$$E_{ij}^{(l+1)} = \text{Update}\left(E_{ij}^{l}, e_{ij}^{(l+1)} \right)$$

Update signifies the update function, E_{ij}^{l} serves as the matrix of edge features connecting nodes i and j in layer l. By defaulting to l=2, we build a two-layer CGCN network, expanding on the work of UGCN. In order to make sure the model works, we're going to configure the first layer to update the edges and leave the second layer out. The bottom layer will not receive edge updates, and all of the initialization procedures discussed in the discussion section will proceed to the middle layer with them.

G. Feature Extraction and Fusion

We get omics data for 254 cell lines, including gene expression and copy number variation. Among the thirteen distinct cancers represented by these 254 cell lines are those of the skin, digestive tract, blood, and lungs. To create all-encompassing omics models of the cell lines, these datasets are used. We log-normalize the gene levels of expression from the GDSC data to Transcripts Per Million (TPM) values before processing them. up addition, zeroes are used to fill up any missing values in the copy number variation and gene expression databases. After that, we normalize the gene expression data using Gaussian regularization:

$$\exp r_j = \frac{(\exp r_j - z_j)}{\sigma_j}$$

where $\exp r_j$ represents the j-th gene's expression features., μ_j and σ_j are the average and variability of the j-th gene.

We define $\mathbf{M}^g \in \mathbb{R}^{m \times d^g}$ as the matrix of features including gene expression, where m is the count of lines of cancer cells, d^g is a feature vector for a single cell line, and each line represents a gene expression dimension. Using a late-integration strategy, we train individual neural layers to aggregate omics feature characteristics. A representation matrix of f dimensions is constructed by encoding the gene expression characteristic of cell lines. $\mathbf{z}_g$.

The fusion and output procedure incorporates several aspects to produce the final output once the drug and cell representations have been extracted.

Computation of correlation coefficients

Let $(x_{i,j}, x_{i,k})$ signify the drug replies of drug $i \in [1, n]$ in cell lines $j, k \in [1, m]$ given as the I_55, AUC, or drug relevance score, respectively. Following that, the P-value for the correlation $r_{j,k}$ among cell lines j, k as well as for all medications, n is calculated as:

$$r_{j,k} = \frac{\sum_{i=1}^{n} (x_{i,j} - \bar{x_j})(x_{i,k} - \bar{x_k})}{\sqrt{\sum_{i=1}^{n} (x_{i,j} - \bar{x_j})^2 \cdot \sum_{i=1}^{n} (x_{i,k} - \bar{x_k})^2}}$$

where $\bar{x_i}$ and $\bar{x_j}$ represent, for all medicines, the average response of cell lines j and k, respectively. For the above calculation, we only included the medications that were common to both cell lines, as it is normal for not all pharmaceuticals to be evaluated in both cell lines. Similarly, taking into account the rank values of x, we calculate the Spearman correlation coefficient for every pair of cell lines across all medicines.

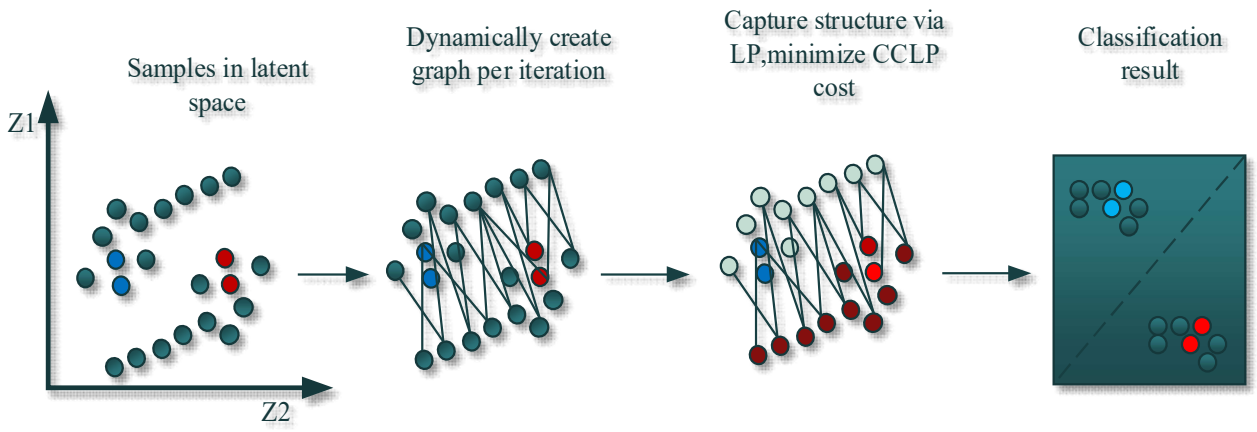
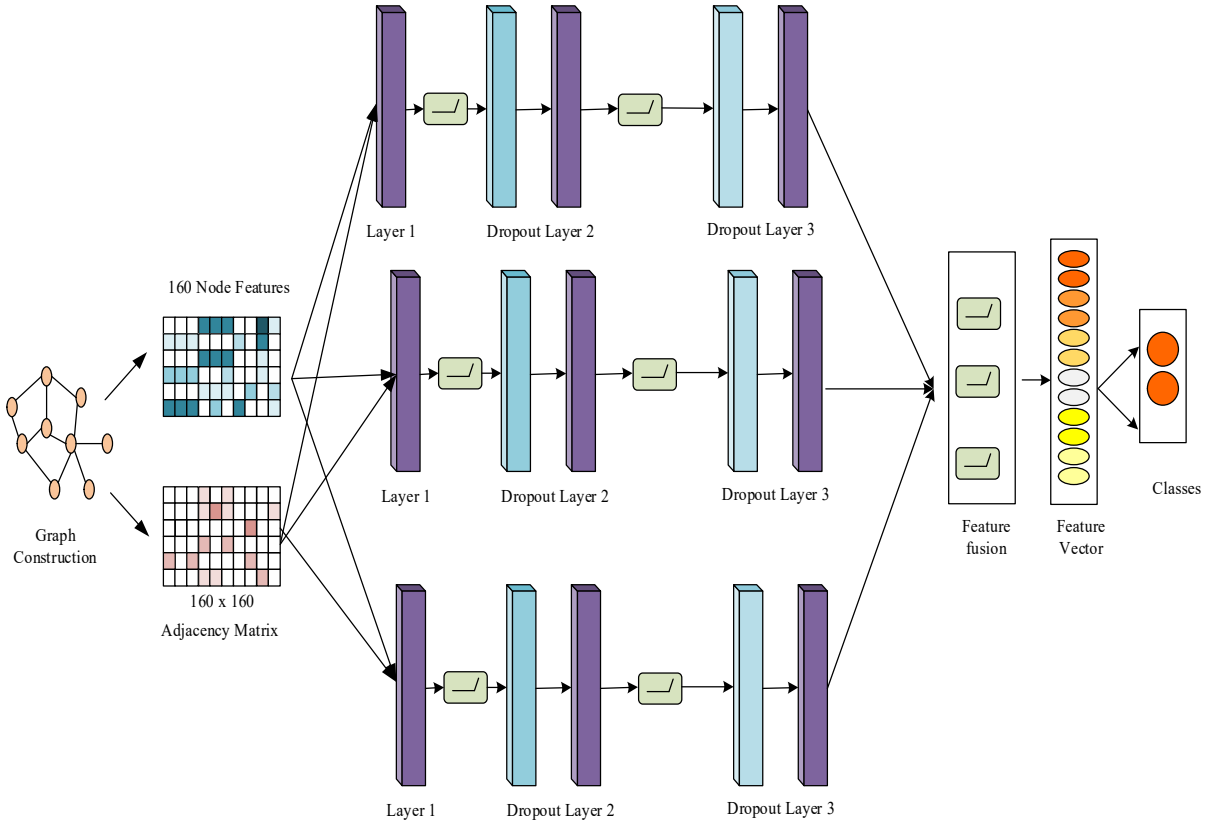


Fig.1. Graphical Model



We began by labeling the whole pharmaceutical corpus and creating collection D using SMILES string characters. The tokenized set is denoted as T. When a new label appears in the labeled set T, it is added at the greatest continuous occurrence frequency until either the size of D meets the maximum length δ or no numerous label surpasses the threshold μ. A series of drug substructures S={S_1,S_2,…,S_i } containing i atoms is produced by this procedure. Using the encoder module in the transformer model, we specify substructure sequences as matrices in order to capture contextual semantic information $M^S \in \mathcal{R}^{l\times\zeta}$, ζ is the maximum length of the drug's substructure sequence, and l is the length of the substructure. In matrix M_i^S, the i-th column represents the structure index of the i-th drug sequence substructure as a one-hot vector. Meanwhile, a one-hot vector is built to collect the drug substructure's location information $I_i \in R^{\zeta}$, which, using elements 1 and 0, indicate the substructure's position information. Consequently, we create an entirely novel representation D_i through the addition of together the drug's position information and the substructure information's representation:

$$D_i = W_e M_i^S + W_p I_i$$

where W_e and W_p are variables that can be taught. The many attention levels of the transformer determine the possible connection Z_i between the substructures:

$$Z_i = \text{Softmax}\left(\frac{(D_i W_q)(D_i W_k)^T}{\sqrt{d}}\right)(D_i W_v)$$

Each drug's final expression is obtained by feeding the output into a fully linked Feed-forward Network (FFN):

$$FFN(x) = \max(0, Z_i W_1 + b_1)W_2 + b_2$$

where W_1, W_2, b_1 and b_2 are learnable parameters.

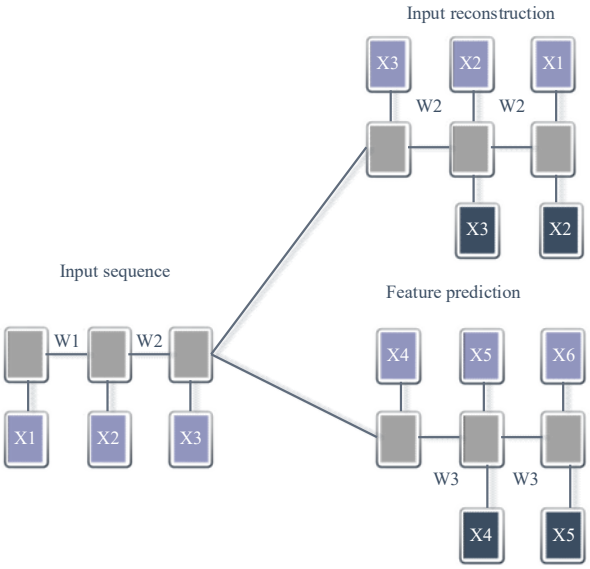


Fig.2. Graph based Ensemble Model

A ReLU-activated linear layer is the first step in the same series of processes that are applied to the transcriptome and interactome features; an additional multi-head attention layer is then used to include the atom characteristics. The attention mechanism improves the model's representational capability by encouraging a greater level of interaction among these characteristics. The two linear layers or the two multi-head attention layers in an interactome feature pipeline or the two transcriptome feature pipelines often use parameter sharing. It is well-known that this strategy improves the model's generalizability by letting it recognize and capitalize on similarities across various input feature types. The need for generality becomes less pressing, however, when our method's objective is to forecast pharmacological reactions for learnt cell lines and medications. This kind of thinking allows the model to concentrate on obtaining more nuanced learning by keeping the parameters among these layers separate. Here is the calculation for the attention layer's output:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k The atomic features are input into two separate linear layers to create matrices K and V, where is the size of the features. To get the matrix Q, the encoded transcriptome or interactome feature is put into a linear layer that does not have the activation function. Combining the results from the two multi-head attention layers yields the molecular feature. A ReLU-activated linear layer subsequently encodes the interactome feature. This layer applies a transformation to the input while maintaining the original input space's dimensionality. A comparable procedure is used to process the transcriptome feature. After that, the two linear layers' outputs are added along with the molecular feature. One set of completely linked layers with dimensions [768+512, 512, 256, 128,1] takes this aggregated feature representation as its input. All of these layers are activated using the ReLU function.

In this study, we applied Shap (SHapley Additive exPlanations) analysis in order to get a better understanding of how each feature contributes to predictions, particularly in more complicated models such as ensemble learning or deep learning. Through the process of assigning the contribution of each characteristic to the overall forecast, it offers a transparent assessment of the significance of the features. That being said, which is absolutely necessary in order to comprehend the biological repercussions of our results... In addition to this, SHAP assists in comprehending the significance of various graph aspects and gene connections in influencing forecasts.

If the IC50 value is less than 1, the block will be red to indicate high sensitivity; however, users have the freedom to select sensitivity through comparison of the IC50 numbers of that drug in different cell lines. This is how drug sensitivity for GDSC while CCLE is determined. Among cancer cell lines, GDSC's drug sensitivity dataset is among the biggest. From GDSC, which has information on 224 medicines evaluated against various bladder cancer cell lines, we retrieved data relevant to pharmaceuticals used to treat bladder cancer. The mean IC50 is the average drug sensitivity across all accessible cell lines in the database. The red cells represent the sensitive cell lines. This graph compares the IC50 values of several medications to the average IC50 and to other cell lines that were chosen for the study. Additionally, CCLE includes data from 13 clinically relevant medications that were evaluated in-house against 10 cell lines, as well as a dataset of 24 pharmaceuticals that are clinically relevant. The CTRP team evaluated over 475 chemicals on bladder cancer cell lines, and you may download the area under the curve (AUC) that was created for each of these lines. Applying the R extreme value software package, we transform the AUC values and classify outliers as either resistant or sensitive. Included in GDBC are all these datasets.

The final result is an estimate of the half-maximal inhibiting concentration's natural logarithm. Due to its shown efficacy in regression issues, the MSE is used as the loss function to train the DL framework.

V. Results & Discussion

A. Experiments

Our 35-drug list includes 30 targeted medications and 5 non-targeted chemotherapeutic treatments; of the 14 targeted drugs, 16 are FDA-unapproved and 14 are in clinical usage. Indices such as the model's sensitivity, specificity, precision, and accuracy, as well as the F1-score, are used to assess the outcomes. Finally, we assess the

relationship between the physiological and biological importance of targets preserved by our proposed approach NDSP throughout feature selection and conduct enrichment analysis on them using the metascape platform. We next look at the medicine and illness in question.

The data was preprocessed by collecting and sorting the cell lines, which are multi-omics samples, into responsive and non-sensitive categories according to the binarized IC50 values for every medication.

B. Z-score transformations

For each medication, we calculated the z-score translations of the drug reactions x_ij over all tested cell lines, or more precisely:

$$Z_{i,j} = \frac{x_{i,j} - \bar{x}_i}{\sigma_i}$$

where $\bar{x}_i, \sigma_i$ reflect the average and dispersion of medication i's reaction across every line of cells it was evaluated on, correspondingly. We independently checked that the precomputed zscored data were completed across all cell lines before using them in the GDSC analysis.

C. Model development

Building NN architectures and optimizing model hyperparameters (HPs) are part of model development. Typical heuristics used by NN developers include relying on intuition, experimenting, and incorporating structures from adjacent domains. Selecting the architecture, learning algorithms, and fundamental NN modules are all part of this procedure. Researchers have explored several DL approaches because to the diversity of information representation for malignancies and medications, as well as the possible use of DRP models in various pre-clinical and clinical contexts.

D. Model training and validation

The GDSC dataset was used for 5-fold cross-validation to verify the model on learnt cell lines and medications. To improve the accuracy of the forecasts, the 5-fold predictions were averaged. The hyperparameters that were suggested were utilized for the baseline with precision. When separating the GDSC dataset into a training set and a test set, no duplicate cell lines or medications were present, allowing for verification on unlearned cell lines with pharmaceuticals. The implementation of a 25-fold cross-validation ensured statistical significance and allowed for the use of an incredibly huge training set.

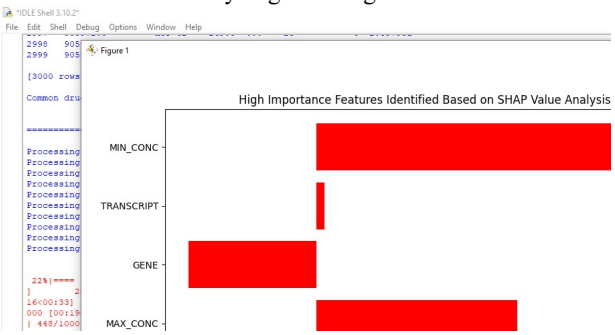


Fig.4. SHAP for Feature Computation

Scatter Plot of Genomic Feature

The scatter plot of the GENOMIC_FEATURE indicated the distribution and variability of gene and transcript data. It helps to identify the diverse set of genomic characteristics among the samples, which could have implications for drug response prediction.

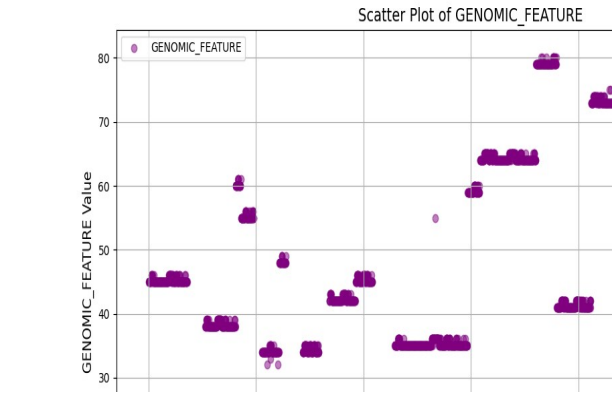


Fig.5. Scatter Plot of Genomic Feature

Scatter Plot of Structural Feature

The scatter plot for STRUCTURAL_FEATURE demonstrated the range and distribution of minimum and maximum concentration values. This variability is crucial as it may influence the effectiveness of drug treatment.

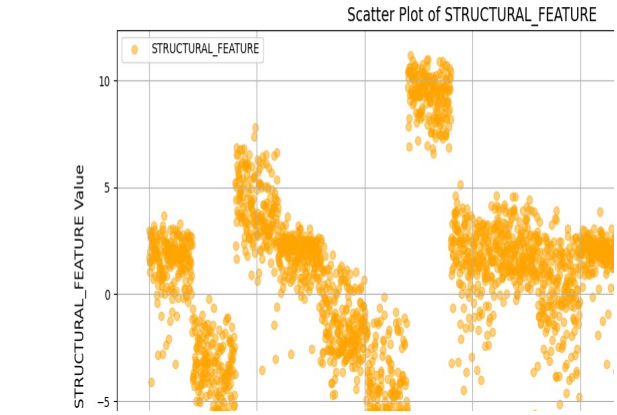


Fig.6. Scatter Plot of Structural Feature

Scatter Plot of Combined Genomic and Structural Feature

The combined features scatter plot illustrated the relationship between GENOMIC_FEATURE and STRUCTURAL_FEATURE. The color gradient representing combined values indicates how these features interact, revealing potential correlations that could be leveraged in predictive modeling.

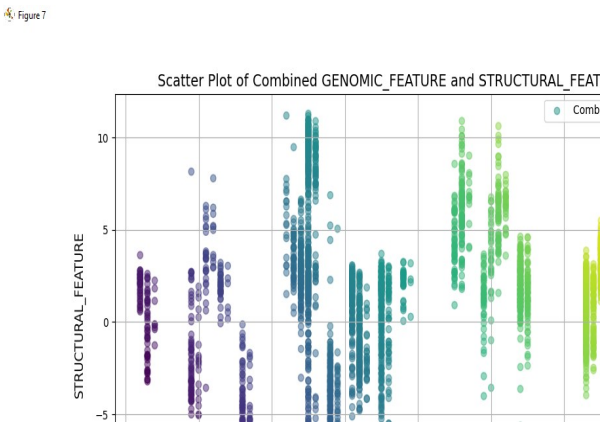


Fig.7. Scatter Plot of Combined Genomic and Structural Feature

Selected Features

Table 1: Selected Features

DTI
Protein Structure & Sequence
GB, and GA
Drug and Target Similarity
DNA Methylation
Image Data
Genomic Sequence
PPI
DDI, and DCI

Through the computation of the correlation among the anticipated and known IC50 values, we were able to assess the accurate prediction performance of the model. In addition to this, the loss that was sustained throughout the process of training the model is documented. A comparative analysis of these measurements is carried out using methods that are analogous to them, such as CNN, GNN, and Random Forest models. For the purpose of doing more research on the model, we used a total of 25 random cross-validation sets to choose the one that produced the best results.

In accordance with the findings of earlier research, we make use of the following four assessment metrics: area under the curve (AUC) and area under the precision-recall curve (AUPR), mean squared error (MSE), and R2.

- One way to measure how accurate a model is by looking at its receiver operating characteristic (ROC) curve, which compares the true positive rate (y-axis) with the false positive rate (x-axis). When assessing binary classification models, researchers often use the area under the ROC curve, abbreviated as AUC. The area under the ROC curve is represented by AUC, and its value may be anywhere from 0 to 1. One strong statistic for evaluating models is area under the curve (AUC), which is unaffected by the ratio of positive to negative data.
- One way to measure AUPR is using the AUPR metric. Here, recall is on the x-axis and accuracy is on the y-axis. An easy way to gauge how well a model is doing is via AUPR. Improved model performance is shown by AUC and AUPR values that are closer to 1, which also suggest increased recall and accuracy.
- Our loss function, which measures the discrepancies between the actual and projected values, is the mean squared error (MSE). Assume that i and n_c are the index and count of cell lines, respectively, and that j and n_d are the index and count of medications. Calculating MSE entails

$$\mathrm{MSE}(y,\hat{y}) = \frac{1}{n_c \times n_d}\sum_{j=1}^{n_d}\sum_{i=1}^{n_c}(y_{ij}-\hat{y}_{ij})^2$$

Where $\hat{y}_{ij}$ stands for the value of the expected pharmacological reaction in a cell line i and drug j, and y_{ij} denotes the actual worth of their answer in that context. In general, a reduced MSE indicates stronger predictive ability from a model, and a zero MSE indicates that the model is approaching flawless prediction. Still, there is no maximum value for MSE.

The magnitude of the goal value, which in our research is the drug response level assessed by log∗IC50, may considerably impact the evaluation of MSE, a simple metric for regression model. Also, we used coefficients of correlation to make our model evaluations more fair R^2 how much of the model's output variables (i.e., drug reactions) can be understood by looking at the model's independent variables R^2 is computed by

$$R^{v}(y,\hat{y}) = \frac{1}{n_d}\sum_{j=1}^{n_d}\left(1-\frac{\sum_{i-1}^{n_c}(y_{ij}-\hat{y}_{ij})^2}{\sum_{i=1}^{n_c}(y_{ij}-\bar{y}_j)^2}\right)$$

where $\bar{y}_j$ is the mean of the drug j response levels across all cell lines. R^2 has a range of $[-\infty,1]$ on the trial version.

Contrary to MSE, greater R^2 greater model performance is shown, and when R^2 equals to one, flawless prediction is attained. R2 equals 0 when the model arbitrarily predicts that all outputs will be the average of true labels. In addition, it might be negative on occasion if the model consistently produces lower results than averages.

At first, we tried out feedforward DGEM on individual omics datasets. The outcomes of the medication response prediction were shown in Table 1 together with the mean squared error (MSE) and coefficient of determination (R2).

We contrasted the outcomes using dense and graph embeddings for mRNA expression, mutations, and CNV data. The functional interactions of biomolecules from interactome data may be used as prior knowledge using graph embeddings.

Table 2: Selected Features

Feature Selected	Classifier	ROC	Accuracy	Precision	Recall	F1-score
2	SVM	0.8457	0.7763	0.7457	0.8068	0.7694
	RF	0.7832	0.7037	0.5939	0.8137	0.6673
	XGBoost	0.8573	0.7925	0.8242	0.7607	0.7990
	Proposed	0.9071	0.8386	0.8362	0.8410	0.8383
4	SVM	0.8588	0.7788	0.7474	0.8103	0.7718
	RF	0.8004	0.7208	0.6416	0.8000	0.6969
	XGBoost	0.8970	0.8412	0.8567	0.8256	0.8437
	Proposed	0.9327	0.8634	0.8652	0.8615	0.8637
6	SVM	0.8645	0.7771	0.7457	0.8085	0.7700
	RF	0.8050	0.7242	0.6382	0.8103	0.6984
	XGBoost	0.8939	0.8335	0.8635	0.8034	0.8384
	Proposed	0.9253	0.8565	0.8447	0.8684	0.8549
8	SVM	0.8687	0.7797	0.7457	0.8137	0.7721
	RF	0.8020	0.7293	0.6126	0.8462	0.6937
	XGBoost	0.9007	0.8215	0.8379	0.8051	0.8245
	Proposed	0.9233	0.8488	0.8447	0.8530	0.8483
10	SVM	0.8777	0.8155	0.7628	0.8684	0.8054
	RF	0.7972	0.7233	0.5751	0.8718	0.6754
	XGBoost	0.9156	0.8352	0.9044	0.7658	0.8460
	Proposed	0.9368	0.8745	0.8737	0.8752	0.8745
12	SVM	0.8931	0.8079	0.7833	0.8325	0.8031
	RF	0.8132	0.7455	0.6485	0.8427	0.7183
	XGBoost	0.9203	0.8676	0.8805	0.8547	0.8694
	Proposed	0.9235	0.8497	0.8464	0.8530	0.8493
14	SVM	0.8946	0.8155	0.7986	0.8325	0.8125
	RF	0.8096	0.7319	0.6672	0.7966	0.7135
	XGBoost	0.9254	0.8736	0.8754	0.8718	0.8739
	Proposed	0.9317	0.8599	0.8652	0.8547	0.8608
16	SVM	0.91645	0.7797	0.7457	0.8137	0.7721
	RF	0.8020	0.7293	0.6126	0.8462	0.6937
	XGBoost	0.9007	0.8215	0.8379	0.8051	0.8245
	Proposed	0.9233	0.8488	0.8447	0.8530	0.8483

A small number of deep neural networks were included in a recent review that compared approaches for medication prediction in cancer lines.

Table 3: Comparison

	MSE	**AUC**	**AUPR**	**R2**
Our work	0.28 ± 0.01	0.9563	0.9743	0.90 ± 0.01
CNN	3.02 ± 0.17	**0.9890**	**0.9501**	0.10 ± 0.02

The receiver operating characteristic (ROC) curves for our model and other cutting-edge CDR prediction approaches are shown in Figure 2. When compared to other approaches, our model's constantly larger ROC curve shows how well it learns cell line and drug representations automatically, which leads to accurate prediction of cancer medication response.

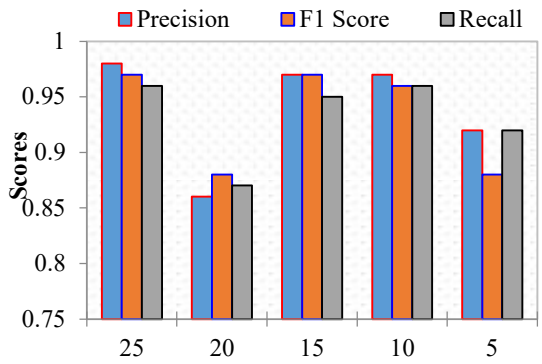


Fig 8:- K-Cross Fold Validation

Molecular drug structures and genomic, transcriptome, and epigenome data from cell lines are combined in a proposed Deep Ensemble model for cancer treatment response prediction. The input features were successfully captured using convolutional neural networks (CNNs) and fully connected networks (FCNs), which extract representations of those characteristics. In the end, a Deep Graph Ensemble model was used to forecast IC50 by merging these characteristics.

When it comes to knowing how people may react to certain medications, there are a number of benefits to combining and analyzing multidimensional information using computational methods. The intricate relationships among genes, proteins, metabolites, and drug responses may be better understood with the use of multi-omics data in drug response prediction. Better patient outcomes and a tailored medical revolution might be within reach with this integrated technique. Applying convolutional neural networks (CNNs) to the problem of drug response prediction is fraught with difficulties. Among them, you may find a high demand for network parameters, a small number of samples, and inputs that are highly dimensional and heterogeneous due to data coming from several omics platforms. The end result is that deep neural networks need a huge amount of data to be trained properly. While prior CNN efforts focused on only two kinds of omics data, our model provides a generic approach to integrating a wide variety of omics data. We suggested two embedding techniques, dense embedding and graph embedding, to deal with the high dimensionality of omics data types. In addition, we showed how graph embeddings made it possible to include interactome data. An effective method for merging data from several omics sources was provided by the attention layer. At the very last layer, we used an attention method. Investigating the use of attention mechanisms at hidden levels is also an option. Also, it's worth looking at ways to further decrease the amount of trainable factors. While the cost of obtaining omics data continues to decline, there is a growing need for innovative computational methods that can efficiently integrate multi-omics data for applications like individualized diagnosis and therapy.

VI. Conclusion

Deep learning designs, which are often ignored in favor of lower error and more accurate predictions, are associated with a significant improvement in personalized therapy. For the purpose of this inquiry, the Cancer Genomics Study on Drug Sensitivity (GDSC) is used. At the core of our investigation is the newly made available GDSC2 dataset, which, in comparison to GDSC1, utilizes a more effective method for screening for tests and medications. The GDSC2 dataset is comprised of 280 different drugs and 809 different cell lines in total. Only 86 of the drugs that have more than 750 cell lines have been investigated from a scientific standpoint. This means that we only take into consideration the gene expression data for all of the medications and cell lines that have established complete resistance to the medicine. The following are some of the preparation stages that should be implemented: imputation, harmonisation (normalization), data augmentation, and feature selection. Within the framework of our study model, the molecular and genetic properties of multi-omics data contribute to an improvement in the prediction of medicine response. Through the use of multi-scale graph feature representation, the study model that has been suggested enhances the representations of structural and genetic features. Transfer learning is the method that is suggested for the purpose of learning about key gene interactions. Transfer learning is a method that gathers information on gene interactions by using knowledge about the biological area of gene oncology. In the process of drug development, GCNNs are helpful, particularly in the prediction of drug-target interactions. The circular fingerprint ideas are used by these deep learning models in order to retrieve valuable feature representations from the input while they are being trained. To obtain the optimal threshold values for the purpose of classifying multi-omics data as either "sensitive" or "resistant," the proposed research model employs a hybrid graph-based deep learning technique known as an Ensemble Convolutional Neural Network. This approach is used to categorize the data. For the purpose of validating the suggested study, several types of data, including F1-score, Area under the Precision-Recall Curve, Sensitivity, Specificity, Precision, and False Positive Rate, will be analyzed. We use graph-based Ensemble Convolutional Networks (G-ECN) that include structural and genomic drug characteristics in order to outperform the most advanced deep learning models. In addition, we demonstrate a high degree of generalizability on a distinct dataset from the Cancer Cell Line Encyclopedia, as well as discoverable conclusions in case studies that correspond to previously acquired information.

References

[1] Nguyen, G.T., Vu, H.D., & Le, D. (2021). Integrating Molecular Graph Data of Drugs and Multiple -Omic Data of Cell Lines for Drug Response Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19, 710-717.

[2] Almutiri, T., Alomar, K.S., & Alganmi, N.A. (2023). Predicting Drug Response on Multi-Omics Data Using a Hybrid of Bayesian Ridge Regression with Deep

Forest. International Journal of Advanced Computer Science and Applications.

[3] Saini, H., Ahn, M., Ward, G., Kucia-Tran, J.A., Gewinner, C., Ferrari, N., Brothwood, J., Bevan, L., Davis, M.P., Fazal, L., Sims, M., O'Reilly, M., Chessari, G., Ferraldeschi, R., Lyons, J.F., Wallis, N., & Thompson, N. (2024). Abstract 667: Identification of biomarkers of response to MDM2 inhibition in solid tumours using computational, multi-omics approaches. Cancer Research.

[4] Zhang, Y., Lian, L., & Yang, X. (2023). NrGe-DTL: a computational framework for cancer drug response prediction based on deep transfer learning from combined denoised genomic profiles and chemical structure embedding of drugs. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 440-445.

[5] Prasse, P., Iversen, P., Lienhard, M., Thedinga, K., Herwig, R., & Scheffer, T. (2022). Pre-Training on In Vitro and Fine-Tuning on Patient-Derived Data Improves Deep Neural Networks for Anti-Cancer Drug-Sensitivity Prediction. Cancers, 14.

[6] Branson, N., Cutillas, P.R., & Besseant, C. (2024). Understanding the Sources of Performance in Deep Learning Drug Response Prediction Models. bioRxiv.

[7] Wang, Y., Yang, Y., Chen, S., & Wang, J. (2021). DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. Briefings in bioinformatics.

[8] Wang, C., Lye, X.H., Kaalia, R., Kumar, P., & Rajapakse, J. (2021). Deep learning and multi-omics approach to predict drug responses in cancer. BMC Bioinformatics, 22.

[9] Zhang, H., Goedegebuure, S.P., Ding, L., Hawkins, W.G., DeNardo, D.G., Fields, R.C., Chen, Y., Payne, P.R., & Li, F. (2023). M3NetFlow: a novel multi-scale multi-hop multi-omics graph AI model for omics data integration and interpretation. bioRxiv.

[10] Zuo, Z., Wang, P., Chen, X., Tian, L., Ge, H., & Qian, D. (2021). SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. BMC Bioinformatics, 22.

[11] Tak, S., Han, G., Leem, S., Lee, S., Paek, K., & Kim, J. (2024). Prediction of anticancer drug resistance using a 3D microfluidic bladder cancer model combined with convolutional neural network-based image analysis. Frontiers in Bioengineering and Biotechnology, 11.

[12] Peng, W., Chen, T., & Dai, W. (2021). Predicting Drug Response Based on Multi-Omics Fusion and Graph Convolution. IEEE Journal of Biomedical and Health Informatics, 26, 1384-1393.

[13] Abinas, V., Abhinav, U., Haneem, E.M., Vishnusankar, A., & Nazeer, K.A. (2024). Integration of autoencoder and graph convolutional network for predicting breast cancer drug response. Journal of bioinformatics and computational biology, 22 3, 2450013 .

[14] Li, M., Wang, Y., Zheng, R., Shi, X., Li, Y., Wu, F., & Wang, J. (2021). DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18, 575-582.

[15] Huang, S., Hu, P., & Lakowski, T.M. (2021). Predicting breast cancer drug response using a multiple-layer cell line drug response network model. BMC Cancer, 21.

[16] Park, H., Yamaguchi, R., Imoto, S., & Miyano, S. (2022). Xprediction: Explainable EGFR-TKIs response prediction based on drug sensitivity specific gene networks. PLoS ONE, 17.

[17] Wang, X., Zhu, H., Jiang, Y., Li, Y., Tang, C., Chen, X., Li, Y., Liu, Q., & Liu, Q. (2022). PRODeepSyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network. Briefings in Bioinformatics, 23.

[18] Sharma, A., Lysenko, A., Boroevich, K.A., & Tsunoda, T. (2022). DeepInsight-3D for precision oncology: an improved anti-cancer drug response prediction from high-dimensional multi-omics data with convolutional neural networks. bioRxiv.

[19] Li, Y., Guo, Z., Gao, X., & Wang, G. (2023). MMCL-CDR: enhancing cancer drug response prediction with multi-omics and morphology images contrastive representation learning. Bioinformatics, 39.

[20] Piochi, L.F., Preto, A.J., & Moreira, I.S. (2023). DELFOS—drug efficacy leveraging forked and specialized networks—benchmarking scRNA-seq data in multi-omics-based prediction of cancer sensitivity. Bioinformatics, 39.

[21] Pu, L., Singha, M., Ramanujam, J., & Brylinski, M. (2022). CancerOmicsNet: a multi-omics network-based approach to anti-cancer drug profiling. Oncotarget, 13, 695 - 706.

[22] Kim, K., Ju, H., Kim, K., Kang, C., Kang, C., & Woo, S. (2023). Prediction of Drug Sensitivity of HER2-Positive Breast Cancer Cell Line via Graph Neural Network. 2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD), 1-2.

[23] Vishnusankar, A., Unniyattil, A., Haneem, E.M., Abinas, V., & Nazeer, A. (2022). Deep neural network aided multi-omics drug response prediction for breast cancer. 2022 IEEE 19th India Council International Conference (INDICON), 1-6.

[24] Mingxun, Z., Zhigang, M., & Jingyi, W. (2022). Drug Response Prediction Based on 1D Convolutional Neural Network and Attention Mechanism. Computational and Mathematical Methods in Medicine, 2022.

[25] Peng, W., Liu, H., Dai, W., Yu, N., & Wang, J. (2022). Predicting cancer drug response using parallel heterogeneous graph convolutional networks with neighborhood interactions. Bioinformatics.

[26] Chen, Y., & Zhang, L. (2021). How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Briefings in bioinformatics.

[27] Chu, T., & Nguyen, T.T. (2021). Graph Transformer for drug response prediction. bioRxiv.

[28] Wang, H., Dai, C., Wen, Y., Wang, X., Liu, W., He, S., Bo, X., & Peng, S. (2022). GADRP: graph convolutional networks and autoencoders for cancer drug response prediction. Briefings in bioinformatics.

[29] Liu, X., Song, C., Huang, F., Fu, H., Xiao, W., & Zhang, W. (2021). GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. Briefings in bioinformatics.

[30] Liu, X., & Zhang, W. (2023). A subcomponent-guided deep learning method for interpretable cancer drug response prediction. PLOS Computational Biology, 19.

[31] Hiort, P., Hugo, J., Zeinert, J., Müller, N., Kashyap, S., Rajapakse, J., Azuaje, F.J., Renard, B.Y., & Baum, K. (2022). DrDimont: explainable drug response prediction from differential analysis of multi-omics networks. Bioinformatics, 38, ii113 - ii119.

[32] Liu, H., Wang, F., Yu, J., Pan, Y., Gong, C., Zhang, L., & Zhang, L. (2024). DBDNMF: A Dual Branch Deep Neural Matrix Factorization method for drug response prediction. PLOS Computational Biology, 20.

Dr. Baisa L. Gunjal working as Head & Professor in Department of Information Technology Engineering at Amrutvahini COE, Sangamner, India. Her research area is Deep Learning and Image Processing. She is having 27+ years of teaching experience.

Mr. Vikram Kishor Abhang, Ph.D research scholar in Department of Computer Engineering at Ph. D. Research Center, MET's Institute of Engineering, Nashik, India. His research area is Drug Response Prediction and Deep Learning. Currently he is working as Assistant professor in Department of Computer Engineering at Amrutvahini COE, Sangamner, India. He is having 13+ years teaching experience.