Clinical Case Reports and Studies
2025 Volume 9, Issue 1
DOI: 10.59657/2837-2565.brs.25.184

BioRes Scientia
Value your Research

Research Article

Open Access

Enhancing Lung Cancer Drug Prediction: From Traditional Methods to Deep Learning

Saeed Dhekra*, Huanlai Xing

School of Computing and Artificial Intelligence, Chengdu, Sichuan, China.

**Corresponding author: Saeed Dhekra.*

Abstract

With the rapid advancements in considering artificial intelligence in medicine, precision, robustness, and reliability predictions are crucial for investigating cancer drug cells' response to gene-drug interactions in lung cancer to precisely predict the drug response for cancer patients' treatments. In this work, we evaluated the predictive models of a multiple regression model for the response of lung cancer cells to a drug. We use a Gene Expression Omnibus data set of 507 and 1496 cells with multiple regression techniques for feature selection to improve model performance. machine learning models including linear regression, decision tree regression, random forest regression, gradient boosting regression, and support vector regression and evaluated performance using different metrics. We also use deep learning models such as long short-term memory (LSTM), neural network models, graph neural networks (GNN), and ResNet-50. The performance metrics evaluated include mean square error (MAE), root mean square error (RMSE), and R2 score. where ResNet-50 resulted as the best model. and showed superior performance in all key metrics, achieving the highest R2 value and the lowest root mean square error on the test set. Our results highlight the potential of deep learning models to capture complex cancer cells in lung gene-drug interactions with excellent opportunities for improving drug response predictions. ResNet-50 demonstrated superior performance with an MAE of 0.0163, an RMSE of 0.0976, and an R^2 score of -0.0014.

Keywords: lung cancer; cancer drug response prediction federated GATs; deep learning

Introduction

Cancer kills hundreds of thousands of lives worldwide each year, making it a serious public health concern. Because of its complexity and heterogeneity, cancer remains lethal despite enormous improvements in research and medical treatments [1]. The illness is caused by aberrant cells that proliferate uncontrollably, invade healthy tissues, and then affect the healthy cells as well. With localized tumors to metastatic disease or the spread of cancer cells to other parts of the body, it progresses through some stages. While late-stage tumors require vigorous and focused special treatments, early-stage cancers frequently require more sophisticated therapies. Individualized care for this condition, taking into account each patient's distinct genetic, biological, and medical requirements [2]. One of the main issues with cancer treatment is that patients with the same cancer mostly respond differently to chemotherapy due to the inherent het- erogeneity of cancer cells' responses to the drugs they are treated with. Accurate prognosis depends on an understanding of tumor heterogeneity within and between patients. The cancers of each patient show distinct genetic signatures and reactions to treatment, highlighting the necessity for prognostic instruments associated with these side effects [3]. Precisely anticipating a patient's reaction to therapy based on the molecular and clinical features of cells affected by cancer is essential [4]. The goal of customized cancer drugs is to modify care according to the needs of each patient to increase effectiveness and decrease negative effects [5]. Predicting a patient's reaction to a particular cancer medication is essential with the customized approach. The prediction of treatment response in cancer has been changed by recent developments in computational biology and bioinformatics. The development of complex prediction models has been made easier by the integration of multi-omics data, such as transcriptomics, proteomics, metabolomics, and genomes. These computational models, which are powered by deep learning (DL) and machine learning (ML) algorithms, reveal the complex biological patterns which help with better treatment and better medication selection [2]. Attractive approaches to achieving precise predictions have been made possible by the development of ML and DL algorithms [4].

Treatment outcomes are greatly impacted by the ability to forecast a patient's response to cancer medications with better in terms of accuracy. which enables cancer treatment doctors to choose the best course of medication, cut down on trial-and-error methods, and lessen the possibility of giving ineffective medications. Superior outcomes for patients and more economical utilization of healthcare facilities result from this advanced cutting-edge strategy [5]. Predicting drug responses in cancer treatment for several reasons is important. to create individualized therapy regimens that are based on each patient's unique hereditary and biological profile of genes, which boosts therapeutic efficacy aligned with the historical data of cancer treatment [6]. to precise drug response prediction improves patient quality of life by preventing unsuccessful therapies and minimizing negative side effects of drugs. to improve overall healthcare efficiency by allocating treatments to people who will benefit from them most effectively in all terms [7].

The term personalized healthcare refers to the customization of medical care to meet the unique needs of every patient. When choosing the best therapy for cancer patients, this method takes into account the genetic, cellular, and surgical features of the patient's tumour. The goal of personalized medicine is to minimize side effects and increase therapeutic benefits [8]. Even with potential advantages, there are a lot of difficulties in forecasting medication responses in cancer patients. Complexity is increased by the constantly changing characteristics of cancer and the impact of the tumour microenvironment [9]. Traditional ML techniques, including logistic regression (LR), support vector regression (SVR), and RFs, were used in the early attempts to predict how cancer responds to medications. to create prediction models, these methods concentrated on obtaining features from a variety of datasets, such as genomic, transcriptomic, and clinical data [10]. This was a seminal moment in the application of computational techniques to comprehend and forecast treatment-related responses from various cancer kinds according to their biological features. The development of DL, and in particular deep neural networks (DNNs), changed the landscape of cancer therapy prediction modelling dramatically. DNNs are excellent at extracting complicated features from raw data independently, eliminating the requirement for a laborious process of feature engineering. Furthermore, through the examination of tissue samples, twin convolutional neural networks (tCNNs) have demonstrated efficacy in the analysis of spatial data from histological pictures, allowing for predictions of treatment effectiveness [11]. Pengfei Liu et al., for instance, used the analysis of digital tumour slides to show how applicable tCNNs are for predicting immunotherapy responses in patients with colorectal cancer [12]. These findings highlight the development of more sophisticated computational methods that can handle complex datasets and uncover useful information for creating customized cancer treatment plans. In this study, we present a unique method that combines ML models for feature selection for big historical data and trains ML and DL-supervised learning models. By utilizing the relational data found in systems of cells, such as drug-target linkages, protein interactions, and genetic sequences for the lung cancer cell response prediction. to improve the accuracy and reduce the error of anticipating unique reactions to cancer treatments, our framework hopes to further precision oncology and improve outcomes of treatment for those diagnosed with cancer.

Literature Review

For easier multi-omics disease evaluation, investigators introduced Deep MOCCA. To improve prediction skills, this framework incorporates relationships between proteins and genes as well as transcripts and other sorts of omics data to improve prediction accuracy [13]. Deep MOCCA uses graph convolutional neural networks (CNN) to estimate patients' survival durations for 33 distinct cancer types in individual patient data. Thanks to the addition of a graph attention mechanism specifically designed to detect driver genes and prognostic markers with the CNN Model, the framework outperforms current survival prediction approaches. In an- other study, cancer images are conceptualized as unstructured networks, which presents a novel CNN technique with different hidden layers and activation functions [14]. Through attentiveness mechanisms, these network visualizations are

combined with embedded data obtained from gene expression patterns. In non-human small cell lung cancer, in particular, this integration enhances tumor classification by using patient survival results and resulting performance from other existing studies. This method facilitates spatial DNA profiling by efficiently identifying variations in tumors through the use of genetic expression data and pathological images. researcher in another study presents a graph neural network method for cancer drug reaction (CDR) forecasting that uses comparison learning [15], To forecast CDR, Graph CDR makes use of drug chemical compositions, established reactions, and multi-omics profiles of cancer cells of the liver. The approach uses contrastive learning in a multitask learning framework to improve the ability to generalize results in various experimental conditions. The significance of biological characteristics, established cell line-drug re- actions, and contrastive training in attaining precise CDR predictions is emphasized by elimination research. An additional study describes the Multi-View Contrastive Heterogeneous Graph Attention Network (MCHNLDA) in their work, which aims to anticipate the relationships between diseases and long non-coding RNAs (lncRNAs). MCHNLDA creates two different graph perspectives by utilizing a wealth of biological data about genes, cancer cell physical structure, and lncRNAs [16].

These perspectives enable collaboratively guiding graph-embedded data using cross-contrastive learning without the need for labelled data. To efficiently collect sequential structural information along meta-paths, the model combines an LSTM network with a heterogeneity context-dependent Graph Attention Network (GAT) as part of its attention mechanism. The author of the recent study studied over 600 individuals who were given immune checkpoint inhibitors (ICIs) utilizing ICI-net to forecast therapeutic responses throughout diverse types of cancer cells [17]. ICInet outperforms other clinical biomarkers for ICI responses and exhibits robust generalization to various kinds of cancer. Their algorithm selects immunotherapy-response-associated biomarkers and improves prediction accuracy in cancer; it obtains an area under the curve (AUC) of 0.85. MMCL-CDR (Multimodal Contrastive Learning for Cancer Drug Responses), a novel method that integrates multiple data modalities to predict cancer drug responses, is presented by [18]. By integrating DNA copy number variance, gene expression, cell type morphological photos, and drug chemical structures, MMCL-CDR boosts accuracy in prediction by synchronizing tumor cell lines across diverse data sources. Experimental results reveal MMCL- CDR's advantage over existing approaches for forecasting cancer treatment responses. The XMR model, an explainable multimodal neural network intended for responses to drug estimation, has been developed by [19]. The framework incorporates a graph neural network (GNN) to understand drug structural characteristics and an external neural net-work for learning genetic parameters. XMR enhances com- prehension and provides insights into predictive mechanisms by predicting medication reactions based on gene mutations and molecular structures through the incorporation of a multidimensional fusion layer. A thorough examination of DL models to forecast response to single-drug therapies was carried out by [20], revealing common practices and difficulties in the field. Their analysis sheds light on how DL applications are doing right now for predicting treatment responses. To anticipate the relationships between miRNA and drug resistance, it suggests the Attentive Multimodal Graph Convolution Network (AMMGC) [21].

AMMGC considerably outperforms current strategies in predicting connections among resistant drugs with miRNAs through the acquisition of residual representations of medications with miRNAs through the convolution of graph neural net- works and an attention neural network. Leveraging a heterogeneous network framework, the study presents drGAT, the graph-based DL approach to drug response prediction [22]. By forecasting drug response results and clarifying the effects of drugs via attention coefficients, the model reaches great accuracy as well as precision. Utilizing numerous gene interaction relationships, MRNGCN is a graph convolution network technique for discovering cancer driver genes [23]. To increase node and link prediction accuracy, MRNGCN combines gene characteristics discovered using heterogeneous graph convolution networks with a self- attention approach. Their study presents

NMGMDA, a computational model that uses nuclear norm minimization and graph attention networks to predict drug-microbe interactions. Using the similarities between medications and microorganisms that are already associated with one another, NMGMDA derives forecasting scoring for drug-microbe correlations [24]. research in which they use a network reconstruction technique to develop drug response forecasting as an associated forecasting issue among an extensive group of cells [25]. The approach they use demonstrates outstanding precision and biological significance by accurately classifying sensitive as well as resistant cancer cell- drug relationships by generating "network patterns" for both medicines and cell lines. A thorough analysis of numerous representation techniques is given in, which also highlights the drawbacks of each approach [26].

To provide insightful analysis for academics working on the subject, the article also looks at possible developments in these methods. To choose experiments that provide reaction data, which is essential for determining successful treatments and improving drug response forecasting models, look at a variety of active learning techniques. Their research shows that when it comes to finding successful medicines, active learning techniques typically perform better than a random process [27]. Leveraging the DL model presents shinyDeepDR, an online platform for computational screening of 265 possible anti-cancer medications. ShinyDeepDR makes medication identification easier by letting users enter mutation gene expression data from cancer samples without requiring high- performance computer technical expertise [28]. The most prevalent kind of liver cancer, hepatic cell carcinoma, has an "undruggable" gene that the scientists hope is curable. This serves as an example of the tool's efficacy.

Materials and Methods

To comprehensively investigate the gene expression changes in lung cancer cell lines treated with Gefitinib and Erlotinib, we utilized two publicly available datasets. These datasets were accessed from the Gene Expression Omnibus (GEO) databases GSE112274 and GSE149383, respectively, which archive and freely distribute high-throughput functional genomic data. The datasets were analyzed to understand the drugs' molecular mechanisms and identify potential biomarkers for drug response in non-small cell lung cancer (NSCLC). The study involved several steps, including data preprocessing, normalization, feature extraction, and the application of ML and DL models to predict drug responses based on gene expression profiles. The workflow diagram from the Figure 1 illustrates the steps followed in our study.

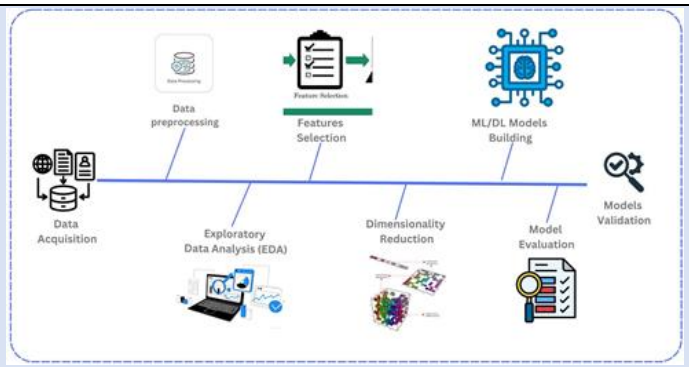

Figure 1: Framework Diagram.

Datasets

In this research, we employed two separate datasets to thoroughly examine the effects of the medications gefitinib and erlotinib on lung cancer cells. We made use of an extensive dataset that was produced from single-cell RNA- seq data Gene Expression Omnibus (GEO) datasets; particularly, GSE112274 single-cell RNA have 507 cells for the medicine Gefitinib [29], and GSE149383 includes 1496 cells for the drug Erlotinib for lung cancer cell lines [29].

Gefitinib Drug

The dataset is titled "Gefitinib Response in Lung Cancer Cell Lines" (Accession Number: GSE112274) [29]. The dataset GSE112274 is accessed from GSE112274 on GEO and is a public repository database for next-generation sequencing and other forms of high-throughput functional genomic data submitted by the scientific community. This focuses on gene expression changes in lung cancer cell lines treated with Gefitinib, an EGFR tyrosine kinase inhibitor used for treating non-small cell lung cancer (NSCLC). The organism studied is Homo sapiens (Human), and the plat- form used is an Affymetrix or Illumina gene expression array, with specific platform details available on the GEO page. The primary objective of the study is to investigate the gene expression changes induced by Gefitinib treatment. The dataset comprises 507 samples, with 200 samples rep- resenting lung cancer cell lines treated with Gefitinib being sensitive to the drug and the remaining samples serving as control cell lines treated with the drug being resistant, as shown in Figure 2. The samples in the dataset include both treated and control groups, providing a comprehensive view of the gene expression changes associated with Gefitinib treatment. The dataset allows researchers to compare treated and untreated samples to identify differentially expressed genes and potential biomarkers for drug response from Table 1.

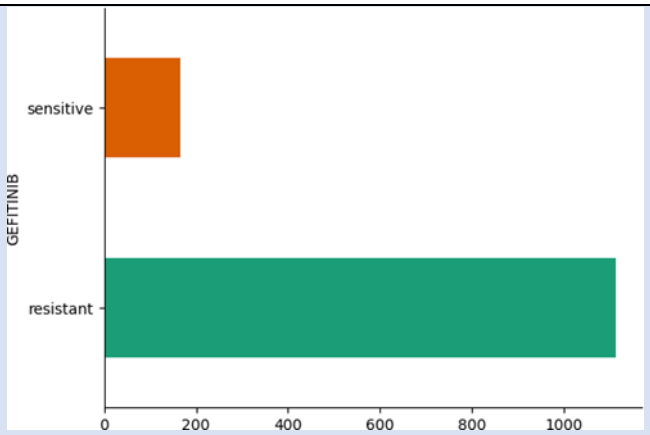

Figure 2: GEFITINIB Drug Distribution.

Table 1: Gefitinib Response in Lung Cancer Cell Lines dataset.

Group	Number of Cells
Treated (Gefitinib)	254
Control	253
Total	507

Erlotinib Drug

The dataset Lung Cancer Cells, with GEO Accession Number GSE149383 given the drug Erlotinib, focuses on investigating the gene expression changes in lung cancer cells treated with Erlotinib, an EGFR tyrosine kinase inhibitor used in the treatment of non-small cell lung cancer (NSCLC) [29]. Our primary objective is to investigate the gene expression changes treated by the Erlotinib drug. The dataset comprises a total of 1,496 cells given in Figure 3. The samples in the dataset include both treated and control groups, providing a comprehensive view of the gene expression changes associated with Erlotinib treatment. The dataset allows researchers to compare treated and untreated samples to identify deferentially expressed genes and potential biomarkers for drug response from Table 2.

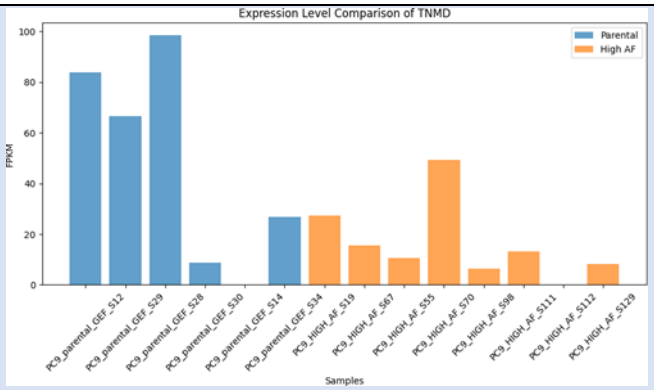

Figure 3: Expression Level Comparison of TSPAN6.

Table 2: Erlotinib Response in Lung Cancer Cell Lines dataset.

Group	Number of Samples
Treated (Erlotinib)	748
Control	748
Total	1496

Data Visualization

Genes Data Expression Levels

The expression levels of the TSPAN6 gene in comparison to a different gene in are shown Figure 3. The x-axis sample gene names are categorized into "Parental" and "High AF" groups [30]. while the y-axis represents the expression level in FPKM (Fragments Per Kilobase Million) units is the normalized counts [31]. The TSPAN6 gene is the protein encoded as a member of the transmembrane 4 superfamily. High AF gene samples exhibit significantly higher TSPAN6 gene levels compared to the Parental. This suggests that the TSPAN6 gene is up-regulated in the High AF group. The PC9_HIGH_AF_S67 in Figure 3 exceptionally high TSPAN6 expression level and is an outlier in the data set. The single nucleotide polymorphisms of the tenomodulin gene (TNMD) [32] expression levels between the parental and high-AF groups differ significantly, as seen in Figure 4. These instances show the existence of distributions with outliers and unsuitable distributions for modelling. To overcome these problems, we need to commence our investigation with thorough data pre-processing. The gene spread in the data set is shown in Figure 5-7. The word cloud's analysis of the genes and clusters offers important hints regarding the pathways in the data that reflect most frequently in the data.

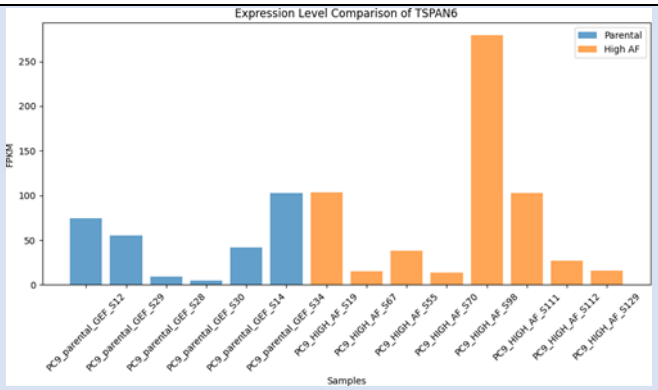

Figure 4: Expression Level Comparison of TSPAN6.

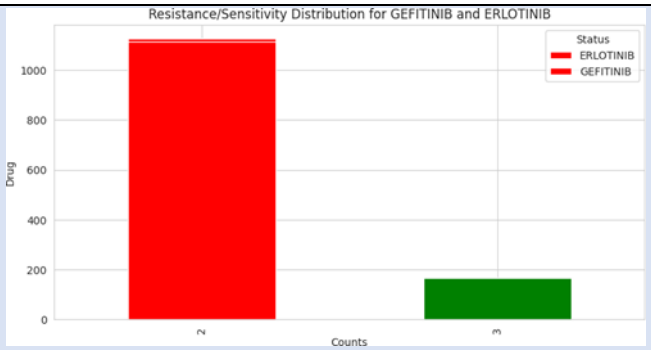

Figure 5: Resistance/Sensitivity Distribution for GEFITINIB and ERLOTINIB Drugs.

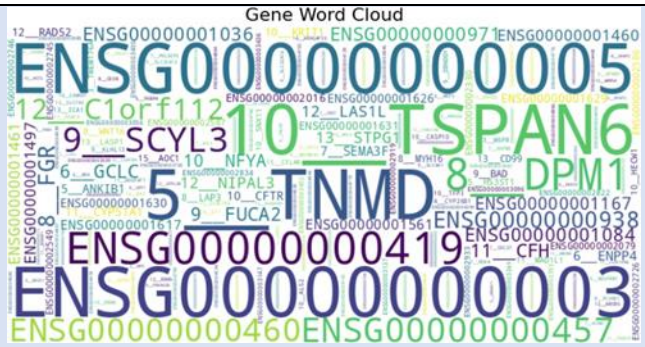

Figure 6: Analyzing the Gene Word Cloud of the dataset.

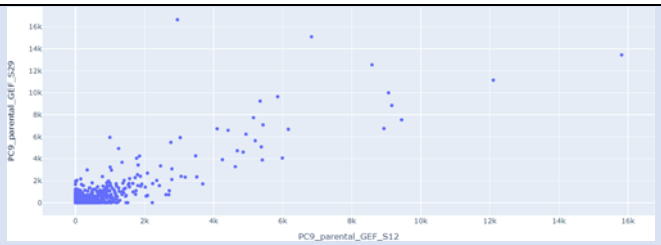

Figure 7: Expression Level Comparison.

Expression

In the comparison of two parental level genes, PC9_parental_GEF_S12 and PC9_parental_GEF_S29, we found a weak positive correlation between the expression levels of the PC9_parental_GEF_S12 and PC9_parental_GEF_S29 genes. Most of the data points cluster in the lower left corner and have relatively low expression levels for both genes. from Figure 8. Some points are located away from the main cluster, which are the outliers in the expression patterns.

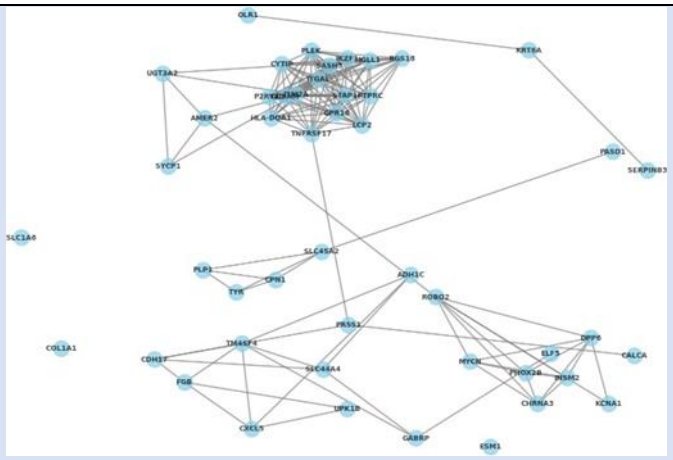

Figure 8: Gene Network Diagram.

Data Prepossessing

Data preprocessing is a critical step in preparing raw data for analysis, ensuring that it is in a suitable format for effective modelling and interpretation. Our study meticulously processed the gene expression data from the GSE112274 and GSE149383 datasets to enhance data quality and reliability. This process involved several key stages, including handling missing values, normalizing and scaling features, encoding categorical variables, and partitioning the data into training and testing sets. Each stage was carefully executed to address potential issues and ensure the data was robust and well-suited for subsequent ML and DL analyses. Through these preprocessing techniques, we aimed to optimize the performance of our predictive models and derive meaningful insights from the gene expression profiles. The Gene Interaction Network is the interaction between various genes in the data set. Each node in the network represents a gene, and the lines connecting the nodes indicate relationships between them in Figure 9. The key observation from the Gene Interaction Network is that the network is relatively dense, with many genes interconnected and complex. UGT3A2 and SYCPI genes have a large number of connections. These genes are considered hubs in the network and play central roles. The network has distinct clusters of genes that are more closely connected. This indicates functional modules within the biological system of lung cells. OLR1 and ESM1 genes have relatively few connections and are located on the periphery of the network. These genes might play more specialized roles in the biological processes of the lungs. The interactions and interdependence between the genes of interest are revealed in great detail by examining the gene interaction networks. The grouped trends, hub existence, and dense interconnected point to a complex and networked biological system. Important genes, operational modules, and putative regulatory mechanisms involved in the basic biological processes are identified by investigating the network in further pre-processing of the dataset. We use the Min Max Scaler as a scaling technique to normalize our features by transforming them to a common scale. We map the input data to a fixed range between 0 and 1, preserving the probability distribution of the data set. We scale our data set features by using Equation (1).

$$\text{Xscaled} = \frac{\text{Xoriginal} - \text{Xmin}}{\text{Xmax} - \text{Xmin}}$$

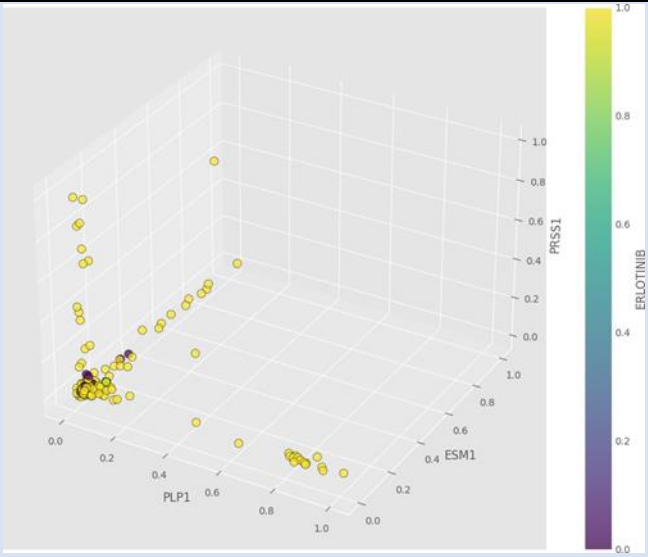

Figure 9: Erlotinib Feature Distribution.

Where $Xscaled$ is the scaled value. original value $Xoriginal$ of the feature. is the $Xmin$ minimum value of data features. is the $Xmax$ maximum value in the feature data points.

$$\text{Xtransformed} = \text{Xscaled} \times (\text{max} - \text{min}) + \text{min}$$

Where from the Equation (2), $Xtransformed$ is the transformed value to a specific range. and $(max - min)$ is the desired range we use [0, 1] and the min is the minimum value, our minimum value is 0.

$$Y = \frac{X - Xmini}{Xmax - Xmin} \times (\max - \min) + \min$$

From the Equation (3), Y is the normalized value X is the original value $Xmin$ and $Xmax$ are the minimum and maximum values of the original data. $(max - min)$ range [0, 1].

We replace the absent values, generally known as (NaN), with the mean of the feature values to eliminate them. The feature's average of all data values is used to replace NaN values by applying the mean replacement technique. we calculate the mean for filling of NaN values.

$$\text{mean}i = \frac{\sum_{j=1}^{Ni} xij}{Ni}$$

Where mean i is the mean of the ith feature. xij is the jth value of the ith feature. Ni is the number of non-missing values in the ith feature from the Equation (4).

The correlations among ERLOTINIB and the PLP1 and ESM16 parameters are shown in a 3D scatter plot in Figure 9. The arrangement of data points and cluster features provides relationships and grouping among the data set.

Feature Selection

Our feature selection methodology combined principal component analysis (PCA) with RF-based feature significance to extract the most relevant features from the gene expression linked to lung cancer treatment response. This process has significance for reducing the dimensionality of the data because it was previously difficult to handle and resulted in the overfitting of models due to the abundance of gene expression variables. We started by performing PCA on the entire dataset to extract the largest variance within the data into fewer components. This procedure helped to transform the original high-dimensional data into a lower dimensional space. Much of the volatility in the gene expression data was explained once the dataset was reduced to 50 key elements. Following the PCA, we used an RF classifier to assess feature importance. The classifier was trained on the scaled dataset, and the importance of each feature was determined by analyzing its performance. We reduced the original dataset to the top 50 most important characteristics using the RF model's relevance scores. Figure 10 displays the top 10 influencing factors, which are PLP1, ESM1, PRSS1, SERPINB3, PHOX2B, KRT6A, ROBO2, Index, CYTIP, and PASD1.

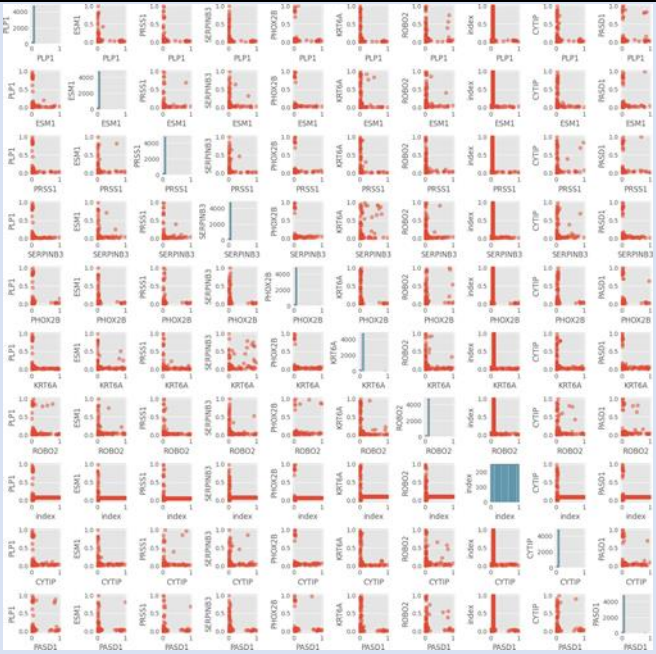

Figure 10: The top 10 features.

Figure 11 ranks the top 50 features based on their importance in a machine-learning model. The x-axis shows the importance and the y-axis names of the features. Dominant Features are ROBO2, PLP1, ESM1, and PRSS1 have higher importance and play a crucial role in the predictions.

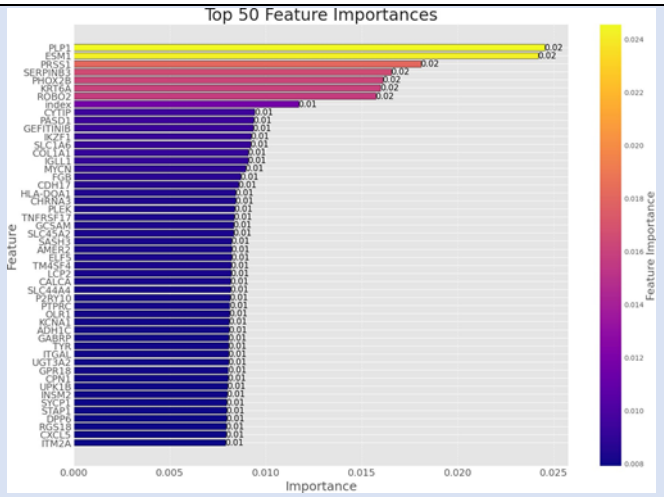

Figure 11: Most important feature with score.

Analysis of correlations in Figure 12 a relationship between two variables is represented by each square in the matrix, and the strength and direction of the link are shown by the color's intensity. White squares indicate no association at all, blue squares indicate a negative correlation and red squares indicate a positive correlation. A strong positive association can be seen in dark red. The relationship is moderately positive when it is light red. Bella, that is unimportant. There is a somewhat negative association seen in light blue. A considerable negative association exists with dark blue.

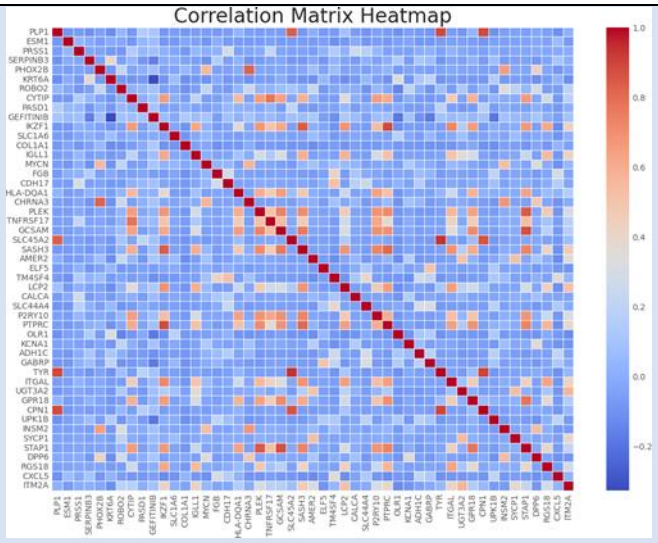

Figure 12: Correlation Matrix Heat map.

PCA of the association among erlotinib and gefitinib medication trends and Figure 13. Color variations and clustering arrangements indicate that these medications have varying effects on gene expression.

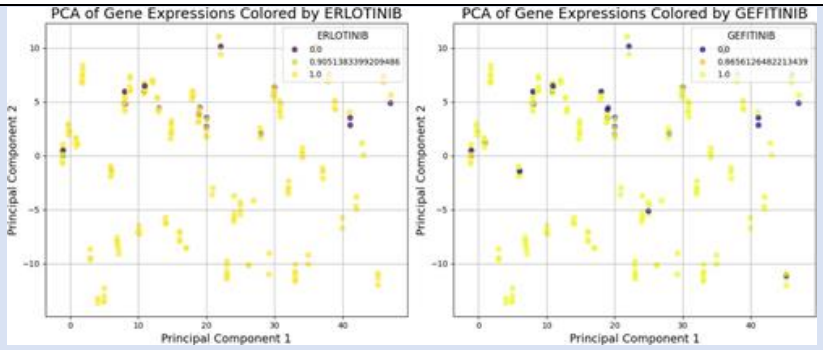

Figure 13: Analyzing PCA of ERLOTINIB and GEFITINIB.

ML & DL Models for Drug Response Prediction

We analyzed and predicted medication responses on lung cancer cell gene expression using ML and DL models. We employ LR, DT, RF, GB, and SVR models in machine learning for this. LSTM, NN, GNN, and ResNet-50 are used in DL. They work incredibly well at managing intricate relational data and deriving insightful patterns from inputs. Such models are chosen to make it easier to fully understand the molecular pathways in cells with lung cancer that are connected to the medications erlotinib and gefitinib.

Machine learning Models

For the GSE112274 dataset, we are predicting the response to Gefitinib drug response treatment. The model equations are as follows Objective Function Equation (5).

$$\text{LGefitinib}(\phi) = \sum_{i=1}^{507} \text{l}(\text{yi}, \hat{\text{y}}\text{i}) + \sum_{\text{k=1}}^{\text{K}} \Omega(\text{fk})$$

The objective function for Gefitinib. Summation over all 507 data points. Equation (6) The loss function measures the difference between the true value yi and the predicted value. The regularization term for the k-th tree fk. Summation of overall K trees in the model.

$$\sum_{\text{i=1}}^{507} \text{l}(\text{yi}, \hat{\text{y}}\text{i}) \hat{\text{y}}\text{i} \sum_{\text{k=1}}^{k} \Omega(\text{fk})$$

Loss Function for the Model Equation (7).

$$\text{l}Gefitinib(\text{yi}, \hat{\text{y}}\text{i}) = \frac{1}{2}(yi - \hat{y}\,i)^2$$

lGefitinib$(yi, \hat{y}\,i)$ in the loss function for Gefitinib. 12 is A constant factor for normalization. $(yi-\hat{y}i)2$ is the squared difference between the true value yi and the predicted value$\hat{y}\ i$. Regularization Term Equation (8).

$$\Omega\text{Gefitinib}(\text{f}) = \gamma\text{T} + \frac{1}{2}\lambda \sum_{\text{j=1}}^{\text{T}} \text{w}_\text{j}^2$$

ΩGefitinib(f) The regularization term for Gefitinib. γT is the complexity of the model, where γ is a regularization parameter and T is the number of trees in the Equation (8). $\frac{1}{2}\lambda \sum_{\text{j=1}}^{\text{T}} \text{w}_\text{j}^2$ is the sum of the squared weights wj of the trees, with λ being a regularization parameter.

$$\hat{\text{y}}_i^{(\text{m})} = \hat{\text{y}}_i^{(\text{m}-1)} + f_m(x_i)$$

Tree Structure $\hat{\text{y}}_i^{(\text{m})}$ is the prediction for the i-th data point at the m-th iteration Equation (9). $\hat{\text{y}}_i^{(\text{m}-1)}$ is the prediction for the i-th data point at the $(m-1)$-th iteration? $f_m(x_i)$: The m-th tree's prediction for the i-th data point.

Deep Learning Model

We use DL models, including LSTM, neural networks, graph neural networks, and graph attention neural networks (GANN), to predict the drug response for the lung cancer cell response. the neural network architectures that leverage the attention mechanism to perform node classification tasks on graph-structured data. Below are the equations and details of how these models are applied to both the GSE112274 and GSE149383 datasets.

Forget Gate: $f_t = \sigma(f_t)$
Input Gate: $i_t = \sigma(i_t)$
Cell State Update: $c_t = f_t oc_{t=1} + i_t oc_t'$
Output Gate: $o_t = \sigma(o_t)$
Hidden State Update: $h_t = o_t otanh(c_t)$
Candidate Cell State: $c_t' = tanh(u_t\, x_t + W_t\, h_{t-1})$
Final Output: $y_t = h_t$

Where σ is the sigmoid activation function respectively. shows the Equation (10) is the key equations of the LSTM model, for the f_t, i_t, o_t are the forget, - c_t is the cell state at time t, -x_t is the input at time t, and output gates cell state update, -ht is the hidden state at time t, - u_t, W_t are weight matrices, candidate cell state calculation, and final output generation.

Neural Network (NN) (11)

$y = f(X; \theta)$
$f(\cdot) = \sigma(W_1 X + b_1)$
$X' = \sigma(W_2 X' + b_2)$

Table 3: ResNet-50 Parameters.

Parameters	Value
Total Parameters	8,316,421
Trainable Parameters	2,771,905
Non-Trainable Parameters	704
Optimizer Parameters	5,543,812

Graph Neural Network (GCN) (12)

$$H^{(l+1)} = \sigma\left(\text{D}^{-\frac{1}{2}} AD^{-\frac{1}{2}} H^{(l)}\, \Theta^{(l)}\right)$$

Graph Attention Network (GAT) (13)

$$a_{ij} = LeakyReLU\left(\alpha_{ij}^{T}\, h_i \| h_j\right)$$

$$\alpha_{ij} = softmax(a_i)$$

$$h_i = \sum_{j \in N(i)} \alpha_{ij}\, W h_j$$

$$h_i^{(l+1)} = LeakyReLU\left(\sum_{j \in N(i)} \alpha_{ij}\, W h_j^{(l)}\right)$$

Where y is the output of the LSTM Model from the Equation 11. σ is the activation function used in all models $f(\cdot)$ represents the network architecture we use in all DL models W and b are weights which are the same in all models and biases are fine-tuned by the feedback loop. X is the input. $H^{(l)}$ is the hidden state at layer l remain different in all models. D, and A are degree matrix and adjacency matrix. α_{ij} are attention coefficients. $LeakyReLU$ is the leaky rectified linear unit activation function used in the NN models Equation 12 and Equation 13.

ResNet-50 Model

We fine-tuned the last 20 layers of ResNet-50 for the single-cell drug response prediction following the input layers total layers use is 6. Input Shape = (X_{train_scaled}. $shape$ [2]), with output dense layers following the other hyperparameters of:

Activation Functions= ReLU
Batch Normalization = Yes
Dropout Rate = 0.5
Optimization Algorithm = Adam
Learning Rate = 0.001
Batch Size = 32
Epochs = 100
Validation Split = 0.2
Early Stopping Patience = 10
Reduce LR Factor = 0.5
Min Learning Rate = 0.00001

Results

We analyzed several ML and DL models to forecast treatment response using single-cell RNA-seq data on lung cancer. For every model, the performance measures were evaluated, which included mean absolute error (MAE), root mean squared error (RMSE), and $R2$ score. Our findings are. The LR model has a mean absolute error of 0.0152, a root means square error of 0.0846, and an $R2$ score of 0.1593. The LR Model concluded comparatively low MAE and RMSE, and its $R2$ score suggested an average match between the predicted outcome and the actual data. The Decision Tree Regressor (DT) has an MAE of 0.0122, an RMSE of 0.1099., and a $R2$ score of -0.2688. Among all models, the Decision Tree Regressor had the least MAE; nevertheless, it also showed a high RMSE and a poor $R2$ score, showing low efficacy for the drug response prediction. RF with a moderate performance of MAE 0.0132, RMSE 0.0944, and $R2$ score close to zero (0.0643) concludes limited predictive power. Gradient Boosting Regressor (GB) from Table 4 showed similar performance to the RF slightly lower $R2$ score. Model Support Vector Regressor (SVR) performed poorly with the highest value of MAE and RMSE, a significantly negative $R2$ score. The best model among all ML models is the DT, which performs better gaming than all models. The neural network for single-cell lung cancer prediction gives better performance compared to the LSTM but a relatively high value of 0.4126 of MAE and 1.1726 of RMSE, with a substantially negative $R2$ score of -143.4178.

Table 4: Performance Metrics of Machine Learning Regression Models.

Model	MAE	RMSE	R2 Score
LR	0.015166	0.084622	0.159252
DT Regressor	0.012168	0.109916	-0.268844
RF Regressor	0.013182	0.094387	0.064344
GB Regressor	0.012639	0.101226	-0.076152
SV Regressor	0.091570	0.118818	-0.482702

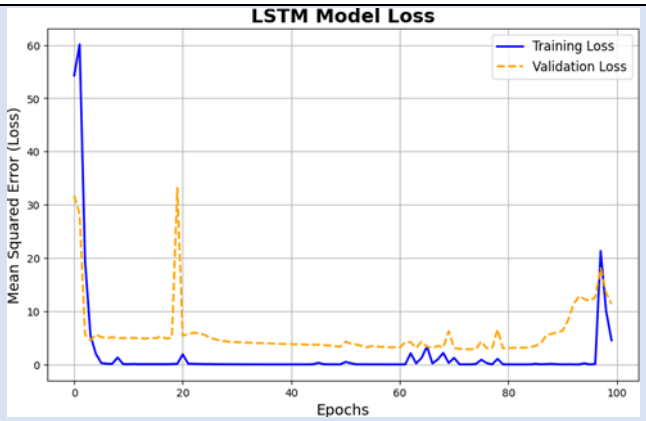

Figure 14: GNN.

The LSTM Model drug response data is in the table. With the lowest *R*2 score and high MAE and RMSE among the all models from Table 5, it is less effective for the drug response prediction. The lowest MAE of the three metrics indicates that the LSTM model's predictions are frequently quite close to the actual values, but the RMSE is higher than the MAE. On the other hand, the *R*2 score indicates that the model is not the perfect prediction for the drug response, as also shown from the training and testing performance of improvement for the data variation. Figure 15 shows the performance measures of the model indication model loss curve during the training phase model learns accurately but fails during testing.

Table 5: Performance Metrics of DL Models.

Model	MAE	RMSE	R2 Score
LSTM	0.4643	3.7906	-1508.0187
Neural Network	0.4126	1.1726	-143.4178
GNN	0.0654	0.3416	-11.2562
ResNet-50	0.0163	0.0976	-0.0014

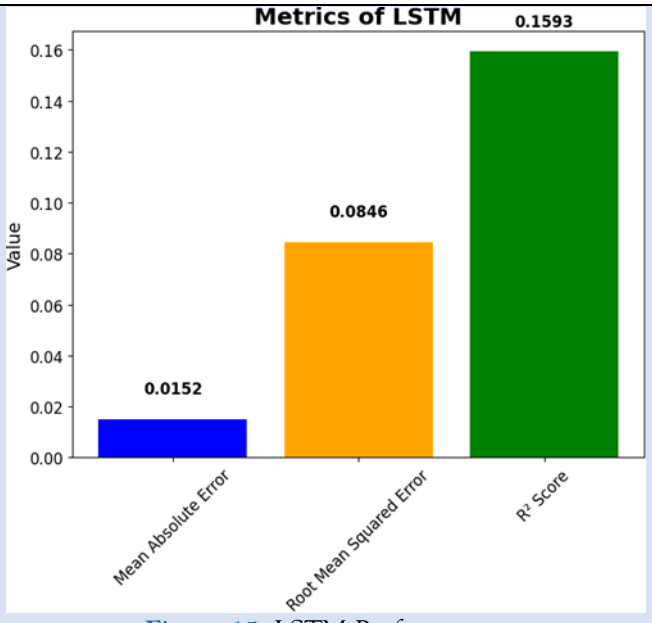

Figure 15: LSTM Performance.

Although their poor *R*2 score, graph neural networks outperform DL and ML models in terms of MAE and RMSE. The training as well as validation loss curve for the GNN model are shown in Figure16. The number of iterations is shown on the x-axis, while the level loss is shown on the y-axis of the graph. The line-colored orange represents a validation loss result, and the line of blue represents the result of training loss. The model's beginning training losses drop off sharply, and this is promptly followed by an unstable period. the outcome of GNNs' complex and the significance of fitting the model to the graph's architecture. The approach has stabilized and is not overfitted, as seen by the convergence of the training

and validation losses. It seems that the general loss estimation is rather large when compared to other models. This may suggest that the chosen framework and hyperparameters are subpar or that the GNN is having difficulty identifying deeper patterns within the data.

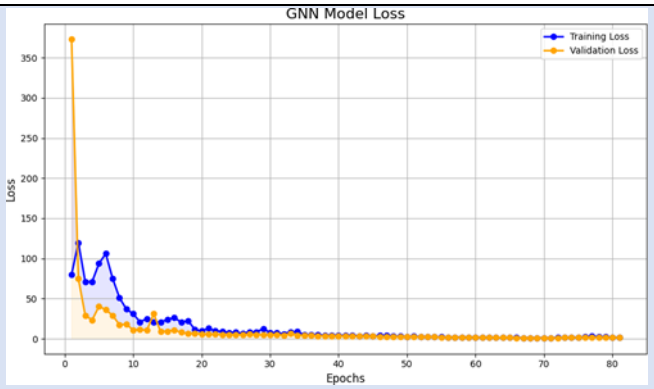

Figure 16: GNN Model Loss Curve.

The training and prediction losses of the ResNet-50 algorithm are shown in Figure 17. The horizontal axis displays the number of iterations of the model we train on; we use 100 iterations for each model and early stop optimization to stop the model training when reaching the most optimized evaluation metrics, while the vertical y-axis shows the loss of training and validation. The blue line, which decreases as the model learns from the training data, represents the learning loss. The orange line evaluation loss of the model performance on test data. The result of $R2$ values is close to zero, which is a fair prediction between actual and forecasted values indicated by $R2$ values of -0.0014. The model does not seem flawless and does not adapt effectively to the new data, since training and prediction losses seem to be constant. In general, ResNet-50 produced the best outcomes in terms of generalization and prediction accuracy for the future of cancer cell reactions given the drug.

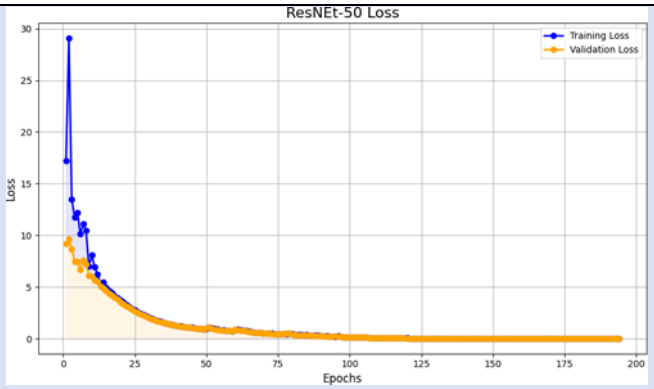

Figure 17: ResNeT 50.

To sum up our results, the role of ideas is demonstrated by early forecasting of drug response by fine-tuning the ReNet-50 model with total parameters of 83 million, including the trainable parameters of 27 million. The models outperform all the DL and ML, having the lowest RMSE and MAE as well as the best $R2$ score, which is comparatively decent. However, the low performance of LSTM and neural network models indicates that they might not be appropriate for the current task and dataset. These results demonstrate the possibility of enhancing and broadening the range of therapeutic approaches for predicting drug responses in lung cancer. The results show that different models perform differently. The results of LR show an $R2$ score of 0.1593, RMSE of 0.0846, and MAE of 0.0152. The decision tree regressor performed poorly with an $R2$ score of -0.2688, RMSE of 0.1099, and MAE of 0.0122. The gradient boosting regressor and the RF regressor performed comparably, with $R2$ scores of -0.0762 and 0.0643, respectively. The least efficient regression model is the SVR, with an $R2$ score of -0.4827 and an MAE of 0.0916. On the other hand, advanced models show varying degrees of effectiveness. LSTMs, neural networks, and graph neural networks show higher errors and lower $R2$ scores compared to classical models. ResNet-50's $R2$ score is -0.0014, RMSE is

0.0976, MAE is 0.0163, and RMSE is 0.0976, showing significantly better performance.

Conclusion

Our studies highlight that the fine-tuning ResNet-50 model has the potential to predict drug responses for single- cell RNA big data in a more sophisticated evaluation metric with less computational power than the other complex approaches of ML and DL. This study lays the groundwork for future implementations focused on treatment strategies while offering significant new insights into the effectiveness of drug response forecasting.

References

1. Xie, J., Zhang, Z., Li, Y., Rao, J., Yang, Y. (2024). Interpretable Drug Response Prediction through Molecule Structure-aware and Knowledge-Guided Visible Neural Network. *bioRxiv*.
2. Yang, Y., Li, P. (2023). GPDRP: A Multimodal Framework for Drug Response Prediction with Graph Transformer. *BMC Bioinformatics*, 24(1):484.
3. Wei, D., Liu, C., Zheng, X., Li, Y. (2019). Comprehensive Anticancer Drug Response Prediction Based on A Simple Cell Line-Drug Complex Network Model. *BMC Bioinformatics*, 20:1-15.
4. Lao, C., Zheng, P., Chen, H., Liu, Q., An, F., et al. (2024). DeepAEG: A Model for Predicting Cancer Drug Response Based on Data Enhancement and Edge-Collaborative Update Strategies. *BMC Bioinformatics*, 25(1):105.
5. Shin, B., Park, S., Kang, K., Ho, J. C. (2019). Self-Attention-Based Molecule Representation for Predicting Drug-Target Interaction. In Machine Learning for Healthcare Conference. *PMLR*. 230-248.
6. Nguyen, G. T., Vu, H. D., Le, D. H. (2021). Integrating Molecular Graph Data of Drugs and Multipleomic Data of Cell Lines for Drug Response Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(2):710-717.
7. Chu, T., Nguyen, T. T., Hai, B. D., Nguyen, Q. H., Nguyen, T. (2022). Graph Transformer for Drug Response Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1065-1072.
8. Nguyen, T., Nguyen, G. T., Nguyen, T., Le, D. H. (2021). Graph Convolutional Networks for Drug Response Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):146-154.
9. Wang, X., Wen, Y., Zhang, Y., Dai, C., Yang, Y., et al. (2024). A Hierarchical Attention Network Integrating Multi-Scale Relationship for Drug Response Prediction. *Information Fusion*, 110:102485.
10. Cui, C., Ding, X., Wang, D., Chen, L., Xiao, F., et al. (2021). Drug Repurposing Against Breast Cancer by Integrating Drug-Exposure Expression Profiles and Drug-Drug Links Based on Graph Neural Network. *Bioinformatics*, 37(18):2930-2937.
11. Chen, Y., Zhang, L. (2022). How Much Can Deep Learning Improve Prediction of The Responses to Drugs in Cancer Cell Lines? *Briefings in Bioinformatics*, 23(1):bbab378.
12. Liu, P., Li, H., Li, S., Leung, K. S. (2019). Improving Prediction of Phenotypic Drug Response on Cancer Cell Lines Using Deep Convolutional Network. *BMC Bioinformatics*, 20:1-14.
13. Althubaiti, S., Kulmanov, M., Liu, Y., Gkoutos, G. V., Schofield, P., et al. (2021). DeepMOCCA: A Pan-Cancer Prognostic Model Identifies Personalized Prognostic Markers Through Graph Attention and Multi-Omics Data Integration. *BioRxiv*.
14. Zheng, Y., Conrad, R. D., Green, E. J., Burks, E. J., Betke, M., et al. (2024). Graph Attention-Based Fusion of Pathology Images and Gene Expression for Prediction of Cancer Survival. *IEEE Transactions on Medical Imaging*.
15. Liu, X., Song, C., Huang, F., Fu, H., Xiao, W., et al. (2022). GraphCDR: A Graph Neural Network Method with Contrastive Learning for Cancer Drug Response Prediction. *Briefings in Bioinformatics*, 23(1):bbab457.
16. Zhao, X., Wu, J., Zhao, X., Yin, M. (2023). Multi-View Contrastive Heterogeneous Graph Attention Network for LncRNA-Disease

Association Prediction. *Briefings in Bioinformatics*, 24(1):bbac548.

17. Zhao, L., Qi, X., Chen, Y., Qiao, Y., Bu, D., et al. (2023). Biological Knowledge Graph-Guided Investigation of Immune Therapy Response in Cancer with Graph Neural Network. *Briefings in Bioinformatics*, 24(2):bbad023.
18. 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(12):btad734.
19. Partin, A., Brettin, T. S., Zhu, Y., Narykov, O., Clyde, A., et al. (2023). Deep Learning Methods for Drug Response Prediction in Cancer: Predominant and Emerging Trends. *Frontiers in Medicine*, 10:1086097.
20. Niu, Y., Song, C., Gong, Y., Zhang, W. (2022). MiRNA-Drug Resistance Association Prediction Through the Attentive Multimodal Graph Convolutional Network. *Frontiers in Pharmacology*, 12:799108.
21. Inoue, Y., Lee, H., Fu, T., Luna, A. (2024). drGAT: Attention-Guided Gene Assessment of Drug Response Utilizing a Drug-Cell-Gene Heterogeneous Network. *ArXiv*.
22. Peng, W., Wu, R., Dai, W., Yu, N. (2023). Identifying Cancer Driver Genes Based on Multi-View Heterogeneous Graph Convolutional Network and Self-Attention Mechanism. *BMC Bioinformatics*, 24(1):16.
23. Liang, M., Liu, X., Chen, Q., Zeng, B., Wang, L. (2024). NMGMDA: A Computational Model for Predicting Potential Microbe-Drug Associations Based on Minimize Matrix Nuclear Norm and Graph Attention Network. *Scient Repo*, 14(1):650.
24. Stanfield, Z., Coşkun, M., Koyutürk, M. (2017). Drug Response Prediction as A Link Prediction Problem. *Scientific Reports*, 7(1):40321.
25. Wang, L. J., Ning, M., Nayak, T., Kasper, M. J., Monga, S. P., et al. (2024). shinyDeepDR: A User-Friendly R Shiny App for Predicting Anti-Cancer Drug Response Using Deep Learning. *Patterns*, 5(2).
26. Lin, C. X., Guan, Y., Li, H. D. (2024). Artificial Intelligence Approaches for Molecular Representation in Drug Response Prediction. *Current Opinion in Structural Biology*, 84:102747.
27. Vasanthakumari, P., Zhu, Y., Brettin, T., Partin, A., Shukla, M., et al. (2024). A Comprehensive Investigation of Active Learning Strategies for Conducting Anti-Cancer Drug Screening. *Cancers*, 16(3):530.
28. Wang, L. J., Ning, M., Nayak, T., Kasper, M. J., Monga, S. P., et al. (2024). shinyDeepDR: A User-Friendly R Shiny App for Predicting Anti-Cancer Drug Response Using Deep Learning. *Patterns*, 5(2).
29. Chen, J., Wang, X., Ma, A., Wang, Q. E., Liu, B., et al. (2022). Deep Transfer Learning of Cancer Drug Responses by Integrating Bulk and Single-Cell RNA-Seq Data. *Nature Communications*, 13(1):6494.
30. Humbert, P. O., Pryjda, T. Z., Pranjic, B., Farrell, A., Fujikura, K., et al. (2022). TSPAN6 is a Suppressor of Ras-Driven Cancer. *Oncogene*, 41(14):2095-2105.
31. Zhao, Y., Li, M. C., Konaté, M. M., Chen, L., Das, B., et al. (2021). TPM, FPKM, Or Normalized Counts? A Comparative Study of Quantification Measures for The Analysis of RNA-Seq Data from The NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, 19(1):269.
32. Tolppanen, A. M., Nevalainen, T., Kolehmainen, M., Seitsonen, S., Immonen, I., et al. (2009). Single Nucleotide Polymorphisms of The Tenomodulin Gene (TNMD) in Age-Related Macular Degeneration. *Molecular Vision*, 15:762.

Cite this article: Dhekra S., Xing H. (2025). Enhancing Lung Cancer Drug Prediction: From Traditional Methods to Deep Learning, *Clinical Case Reports and Studies*, BioRes Scientia Publishers. 9(1):1-16. DOI: 10.59657/2993-0863.brs.25.184

Article History: **Received:** September 09, 2024 | **Accepted:** December 20, 2024 | **Published:** January 08, 2025