Predicting Drug Effectiveness for Cancer Cell Lines

Parshva Shah
University of Alberta
Edmonton, Alberta, Canada
parshva@ualberta.ca

Aakash Durairaj
University of Alberta
Edmonton, Alberta, Canada
durairaj@ualberta.ca

Kevin Qian
University of Alberta
Edmonton, Alberta, Canada
cq4@ualberta.ca

Abstract

Cancer remains a leading cause of death across the world, with many subtypes lacking effective treatment options. Advances in cancer genomics and more comprehensive datasets, such as the Genomics of Drug Sensitivity in Cancer (GDSC2), provide opportunities to enhance predictive models for drug efficacy. In this project, we leverage genomic data and drug response profiles from the GDSC2 dataset to predict the effectiveness of specific drugs on diverse cancer cell lines. We aim to benchmark selected feature reduction alongside chosen models. By integrating gene expression and using advanced machine learning algorithms, our model achieves improved accuracy in identifying drug-sensitive and drug-resistant phenotypes. Our findings demonstrate the potential of combining genomic data and computational modeling to inform personalized cancer therapies. From this study, we were able to conclude a few key observation. Currently, there is a limitation in finding the most effective drug for cancer cell lines due to amount of data available. Although there is an abundance of genetic features per cell line, the amount of cell line that can be trained on is an inhibiting factor which results producing more error during training and validation. This caused the E-MTAB-3610 dataset producing the best results since this model had the most features and cell lines per drug. In terms of the most effective model, Elastic Net performed relatively well with an average mean squared error of 1.43 for the predicted ln(IC50) values. The neural network architectures did not perform as well as expected, largely do the limited data available for each drug. It had a difficult time predicting the IC50 values using the non-linearity of the genomic features.

1 Introduction

Predicting drug sensitivity across a wide range of cancer types and therapeutic compounds is a critical challenge in developing effective, personalized medicine. Current benchmarking studies do not use the most updated datasets and models when making predictions [5]. As research in cancer and genes grow, the complexity and diversity of genomic data grows alongside the research. This study aims to address this issue by conducting a comprehensive benchmarking study that evaluates both datasets and predictive models, providing insights into data quality, feature representation, and model performance.

Most models and datasets used for drug sensitivity prediction have limitations in accuracy and scalability, particularly in how genomic data is expanding in relation to research. By benchmarking datasets and leveraging newer machine learning models, this study aims to provide a benchmark that bridges the gap between genomic data and clinical decision-making, improving outcomes for cancer patients.

This work is essential because it contributes to the broader goal of optimizing cancer treatment by tailoring drug therapies using genomic profiles [2]. By benchmarking three key datasets: Genomics of Drug Sensitivity in Cancer (GDSC2) binary mutation data [17], GDSC2 genomic expression data (E-MTAB-3610) [11], and Cancer Cell Line Encyclopedia (CCLE) [4] —we not only evaluate the predictive potential of these datasets but also evaluate how well machine learning algorithms can model genomic data in relation to drug effectiveness. This study is unique as it focuses on evaluating the datasets available and application of current machine learning techniques.

Our contributions include the development of a benchmarking framework and the application of traditional machine learning modeling techniques such as Random Forests and Support Vector Regressors (SVR). Graphical Neural Networks (GNNs), a relatively new framework, has also been incorporated to assess genomic data. The incorporation of GNNs represents an innovative approach that leverages the structure of cell lines to enhance prediction. Feature reduction and selection techniques were also explored to reduce model training and evaluation time and improving model interpretability.

In this paper, a high-level overview of the methodologies, findings, and key insights are provided. The objectives of this study includes comparing the performance of predictive models across datasets, evaluating the datasets themselves for their representational quality, and identifying the most effective techniques for feature engineering and dimensionality reduction. Additionally, the potential of advanced neural network architectures, such as graphical neural networks and transformers, are investigated to improve prediction accuracy and clinical relevance.

The key findings of this paper includes:

1. A great availability of genomic feature data for each cancer cell line, but an overall limitation in the number of cancer cell line available for each drug, preventing machine learning models to accurate train on data
2. Pearson Correlation was the best feature selection technique alongside Elastic Net being the best model at predicting IC50 values
3. The neural network architecture provided unreliable results due to the limitation of the dataset

2 Background and Related Work

This section provides a brief background on cancer cell lines and the key metrics used in drug sensitivity prediction. It also reviews prior research, emphasizing the evolution of machine learning (ML) approaches in predicting drug responses and addressing their limitations.

2.1 Background

A cancer cell line refers to a population of cancerous cells maintained in a controlled laboratory environment, enabling extensive study over prolonged periods [12]. These lines are developed from tumors, typically through biopsy or surgery, and subsequently cultured under conditions fostering survival and proliferation [6]. Figure 1 illustrates this extraction and culture process.

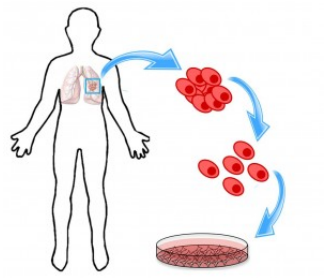

Figure 1: Cell Line Extraction Process [1]

Each cell line, including cancer cell lines, is differentiated by the genes that make up the cell line. A gene is a combination of deoxyribonucleic acid (DNA) that directs how the cell behaves [1]. It is responsible for making protein molecules and helping control the actions of other genes [1]. The genes and gene segments present in a cell line define a certain cancer cell line as seen in Figure 2. Cancer cell lines have mutations at a genomic level, which result in certain genes functioning differently. For example, a lung cancer cell line will contain gene segments that function differently than skin cancer cell lines, which function differently from a normal cell line. The functionality of a gene is measured by using genomic expression data. Gene expression data is the process of turning encoded information in a gene into a function [13]. It can be described as an indicator of the activity of a specific gene. The more involved a gene is in creating protein or being active, the higher the gene expression values will be. Each cell line is characterized by different genes being more active, therefore, gene expressions can and will be used to determine how the activity of a gene itself will cause a drug to be effective.

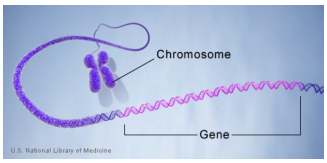

Figure 2: Gene/Gene Segments that Make a Cell Line [1]

The effectiveness of a drug on a cancer cell line is measured by using Half-maximal inhibitory concentration (IC50) values. IC50 values are "a measure of the concentration of a drug or compound required to inhibit a particular biological or biochemical process by 50%" [16]. The lower the IC50 value, the more effective a drug will be to a certain cell line since lower concentrations of that drug are needed to inhibit a biochemical process. This is a critical metric for developing personalized and effective treatments for cancer patients. We want to ensure that the right drug is being administered to the patient. Based on the patient, the effectiveness of a drug can vary, which is what this study is trying to optimize. Given a set of genomic sequences for a cancer cell line, what drug can be provided to result in the most effective outcome?

2.2 Related Work

Drug sensitivity prediction through ML has seen significant strides. Eckhart et al. [5] employed traditional ML models such as random forests and elastic nets, leveraging dimensionality reduction techniques like Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (MRMR). Using the GDSC2 dataset, they demonstrated that elastic nets combined with PCA reduced prediction errors effectively. However, their focus on a single dataset and limited exploration of alternative reduction techniques restricted broader applicability.

In contrast, Yang and Li [18] proposed the GPDRP framework, integrating Graph Neural Networks (GNNs) and Graph Transformers to analyze drug molecular graphs and cell line pathway scores. Despite its high predictive accuracy, this approach's complexity and computational intensity limit its practicality. Its reliance on specific representations, like molecular graphs, further narrows its generalizability.

This study builds on these efforts by aiming to balance interpretability and complexity. Incorporating datasets such as GDSC2, E-MTAB-3610, and CCLE, we provide a comprehensive evaluation of ML techniques. Beyond PCA, we explore advanced methods like Non-negative Matrix Factorization (NMF) and Pearson correlation, deepening the understanding of feature reduction's impact. Additionally, simpler GNN architectures are employed to model genomic relationships, reducing computational demands while maintaining effectiveness. A modular benchmarking pipeline has been developed for this study to ensure reproducibility and comparability. This pipeline systematically assesses models and dimensionality reduction techniques, emphasizing predictive performance, computational efficiency, and interpretability. While Eckhart et al. established a foundational benchmark and GPDRP highlighted the potential of multimodal approaches, this study demonstrates a simpler approach to these problems for applications in personalized medicine.

3 Approach

This study benchmarked three different datasets alongside three reduction techniques and five models to determine the mean square error for each technique and model. The three datasets that were used are: GDSC2, E-MTAB-3610, and Cancer Cell Line Encyclopedia (CCLE).

3.1 Datasets

IC50 values are used to evaluate the effectiveness of each drug in the cancer cell line. The IC50 values were sourced from the GDSC2 dataset [add source] where each combination of drug and cancer cell line has an IC50 value associated with them. Each combination also has other values associated with them that could have potentially been used to evaluate drug effectiveness such as z-score, area under

the curve (AUC), and root mean square error (RMSE). The ln $IC50$ measurements will be used to train, compare, and evaluate the models. The IC50 values can be big therefore we add a multiplicative scale by applying the natural logarithm function to the IC50 values.

3.1.1 *GDSC2.*

The first data source evaluated was the genetic mutation data from GDSC2. This data source contained 969 cell lines that could directly be mapped to the GDSC2 IC50 drug data [17]. Each cell line had 425 genes/gene segment identifiers which were used to define a cell. Each gene/gene segment had a binary matrix associated with the cell lines. The binary matrix was an indicator of whether that particular gene was mutated. All 969 cell lines could be identified by whether certain genes were mutated. This dataset gives a high-level idea as to determining if knowing a gene is mutated or not indicates the effectiveness of a drug.

3.1.2 *E-MTAB-3610.*

The second data source, E-MTAB-3610, provides transcriptional profiling of 1,000 human cancer cell lines as part of the Genomics of Drug Sensitivity in Cancer (GDSC) panel. This dataset is particularly valuable because it includes basal gene expression profiles derived from microarray data, offering a deeper look at the activity levels of genes across various cancer cell lines [11]. To process this data, we used the Affymetrix HT-HG-U133A v2 platform, which uses .cel files containing raw microarray measurements. Once the gene expression data was processed, it was merged with the IC50 values of the drug responses from the GDSC2 dataset. After processing the E-MTAB-3610 dataset, we obtained a comprehensive dataset containing 1,013 cancer cell lines, with each cell line characterized by 20,354 genes and their corresponding gene expression values.

3.1.3 *CCLE.*

The final dataset that will benchmarked is the CCLE dataset. This dataset contains 16,383 genomic features which provide many features to find the most correlated feature when determining IC50 values [4]. This data contains approximately 1,964 cancer cell lines, the wide range of cell line data providing more insight as to how different cell line structures affect drug effectiveness. The CCLE portal had both the gene expression files and a file to map their internal cell line identifier to their corresponding cell name. Once the cell line ID was mapped to its respective name, we combined it with the GDSC2 IC50 dataset to get the final dataset that is used for the models.

3.2 Feature Reduction and Selection Techniques

3.2.1 *Principle Component Analysis (PCA).*

PCA was another feature reduction technique to transform high-dimensional data into something more manageable. PCA is a linear transformation method that projects data into a lower-dimensional space while preserving as much variance as possible [3]. It achieves this by identifying the directions (principal components) along which the data varies the most. In this study, PCA fit and transform were applied to the training data and only fitting the validation and training data. PCA reduces overfitting by eliminating redundant features and reducing the training time. The number of components selects the principal components with the highest variance they explain [8].

3.2.2 *Non-negative Matrix Factorization (NMF).*

NMF is a reduction technique that is used by other studies to reduce the number of components, by creating another matrix that can be preserved. NMF has the following equation:

$$\mathbf{V} \approx \mathbf{WH}$$

where:

$\mathbf{V}$: Non-negative input matrix of size $m \times n$,

$\mathbf{W}$: Non-negative basis matrix of size $m \times r$,

$\mathbf{H}$: Non-negative coefficient matrix of size $r \times n$.

The basis matrix is the resulting reduced matrix with a rank of r, the reduced components that we want to have. Each column in the basis matrix correlate to a basis element that is the reduced components of n rank. H, the reconstruction matrix, is used to reconstruct W back to the original data points using linear combination [7]. We applied NMF by taking the raw input data and creating the basis and reconstruction matrix. The basis matrix is used as the new feature input space which is what the models will be trained on and the reconstruction matrix is used to determine the reconstruction error, a measure of how accurately the basis matrix can be transformed to the original input space.

3.2.3 *Pearson Coefficient Correlation (PCC).*

PCC is a statistical method that measures the linear relationship between two variables, producing values between -1 and 1. PCC was used to rank all the features (genes) in the dataset based on their correlation with the LN-IC50 values, identifying the most relevant features for predicting drug sensitivity [14]. By selecting genes with the strongest absolute correlations, we reduced the dataset's dimensionality while retaining biologically meaningful features. This process enhanced interpretability by focusing on genes likely driving drug response, improved computational efficiency by eliminating irrelevant or weakly correlated genes, and reduced the risk of overfitting. PCC's simplicity and speed made it an ideal choice for the large dataset, enabling the selection of a manageable number of highly predictive genes.

3.3 Models

3.3.1 *Random Forest.*

Random Forest Regression is an ensemble learning method that uses multiple decision trees to make predictions [10]. It operates by constructing a multitude of decision trees during training and outputting the mean prediction of the individual trees [10]. This approach reduces overfitting and improves predictive accuracy. Key strengths of Random Forest Regression include its ability to handle high-dimensional data, robustness to outliers, and minimal preprocessing requirements. We used 5-fold grid search CV to choose the hyperparameters and optimize for the negative mean squared error. The $n_{\text{estimators}}$ parameter, representing the number of trees in the forest, was varied as $n_{\text{estimators}} \in \{5, 10, 25, 50, 100\}$, enabling us to evaluate models ranging from computationally efficient to potentially more accurate with a higher number of trees. The *max depth* parameter, which controls the maximum depth of each decision tree, was explored as max depth $\in \{None, 10, 20, 40\}$ [10]. Limiting tree depth helps prevent overfitting, especially on smaller datasets. The *min samples split* determines the minimum number of samples

required to split a node, min samples split $\in \{2, 5, 10\}$ [10]. Larger values restrict tree growth and promote generalization.

3.3.2 *Support Vector Machines.*

In this project, we employed a Support Vector Regression (SVR) model to predict IC50 values. Unlike Random Forest (RF), which generates multiple decision trees and averages their predictions for regression, SVR focuses on finding a function that best approximates the relationship between features and target variables. A key feature of SVR is its use of kernels, such as the linear kernel and Radial Basis Function (RBF), which define how the data is transformed in the feature space [15]. For this study, we considered two kernel types: $n_{\text{kernel}} \in \{RBF, Linear\}$. Another important hyperparameter in SVR is the regularization parameter (C), which controls the trade-off between minimizing training error and ensuring good generalization [15]. Higher values of C focus on minimizing training error, while lower values allow for greater tolerance of errors, promoting better generalization. We explored $C \in \{0.1, 1, 10\}$ in the experiments. Lastly, we examined the effect of the Epsilon (ϵ) parameter, which defines the margin of tolerance for regression [15]. Epsilon sets the margin of tolerance around the true data points. We tested values of $\epsilon \in \{0.01, 0.1, 1\}$ to understand its impact on model performance.

3.3.3 *Elastic Net.*

Elastic Net (EN) is a regularized regression model that combines both L1 and L2 regularization. This makes it useful for high dimensional datasets with multicollinearity, as it balances feature selection and regularization using L1 and L2 regularization. EN has been used in other studies and found that it has performed well because of genomic data being non-linear [9]. In this study, we will tune both hyperparameters together to determine if they perform well for predicting IC50 values [5]. The model is tuned to optimize for two hyperparameters: α and λ. α relates to the overall weight of both regularization terms on the model. A larger α means greater L1 and L2 regularization resulting in a less complex model. λ relates to the trade-off between both L1 and L2 regularization. A larger value of λ results in a higher emphasis on L1 regularization and a smaller value results in a higher emphasis on L2 regularization. For this study, Elastic Net was particularly useful because it reduced the dimensionality of the dataset by creating sparse regularization weights, and setting irrelevant features to zero while retaining correlated features. The model handled correlated predictors and sparse solutions which is why it was chosen as a baseline model. We explored various combinations of α and λ to optimize IC50 prediction, ensuring the best feature selection and generalization.

3.3.4 *Fastforward Neural Network.*

Due to the non-linear nature of genomic expression data, multilayer perceptron (MLP) can model the relation between genomic expression data and the corresponding IC50 value for a particular drug [5]. Our MLP is a fastforward neural network with two dense, hidden layers and an output layer which is used to predict the IC50 value. The input layer will encode the input data to be passed through the neural network. The MLP model has two hidden layers, each with Rectified Line Unit (ReLu) activation function to better model the non-linear relations. The output later was an output neuron that would use linear regression to predict the IC50 value. The adam optimizer was used for stochastic-based gradient descent and the evaluation metric used between training was the mean absolute error, with the goal of decreasing the error. Both the epoch and the batch size were tuned to optimize for training time and decrease error generated by the model.

3.3.5 *Graph Neural Network.*

To predict drug response, we implemented a Graph Neural Network (GNN) model tailored to represent the structured relationships within cancer cell lines and their interactions with drugs. Each cell line was modeled as a graph where nodes represented gene expression values or mutations, and edges encoded relationships such as co-expression or known gene pathways. The model constructs graphs were constructed with node features, edges, and batch indices for grouping multiple graphs. A GCNConv layer aggregated information from neighboring nodes to compute node embeddings, which were pooled into a graph-level representation using a global mean pooling operation. This cell-line embedding was then concatenated with a drug feature vector to form a unified representation of drug-cell line pairs. The combined embedding was processed through a fully connected network (FCN) with dropout regularization and nonlinear activation to predict IC50 values. GNNs were ideal for this task, as they capture the complex structure of biological data, integrate features effectively, and offer interpretability through node and graph embeddings. This approach allowed us to use structured data for accurate drug response predictions in personalized cancer therapy [18]. Due to time constraints, minimal hyperparameter tuning was performed, and training was conducted for 8 epochs per model to balance performance with computational efficiency.

In contrast, the GPDRP model employs a more complex architecture with three GCNConv layers, global max pooling, and multiple dense layers for both graph and cell-line features, resulting in higher computational demands. While GPDRP processes drug and cell-line features independently through deeper pipelines before combining them, the model simplifies this by using a single GCNConv layer with global mean pooling to extract graph-level embeddings, which are directly concatenated with drug features [18]. This lightweight design reduces training time, mitigates overfitting, and ensures accessibility for researchers with limited computational resources, while still capturing essential relationships for drug sensitivity prediction. This simpler approach prioritizes scalability and interpretability, making it a practical alternative for broader applications.

4 Experiments

4.0.1 *Hardware.*

Bulk of results and models were trained using an AMD Ryzen 7 5700X3D CPU, RTX 3060 12GB, and 32GB of ram. Since the neural networks used are relatively light, it was also trained using CPU even though a GPU would train much faster. However, the GNN was trained using the GPU and it took a lot longer compared to the other models.

4.0.2 *Set Up.*

To systematically evaluate the datasets, feature reduction techniques, and predictive models, we implemented a modular pipeline

to train, validate, and benchmark performance. Each dataset was preprocessed to ensure consistency, remove noise, and address missing values, thereby enhancing the quality and reliability of the input data. Preprocessing steps included standardizing feature scales, normalizing gene expression values, and filtering out incomplete or irrelevant entries. Feature reduction techniques such as PCA, NMF, and PCC were applied to reduce dimensionality while retaining the most relevant information for predictive modeling.

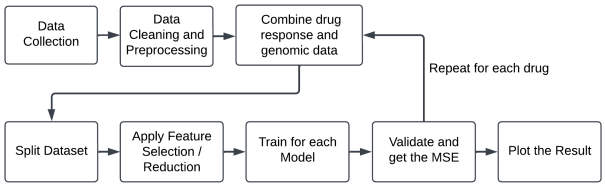

Figure 3: Pipeline of drug response prediction using genomic data

As shown in Figure 3, we first performed data collection. This involves retrieving bulk CSV files from GDSC2, E-MTAB-3610, and CCLE for which we build the datasets. Next, we perform data cleaning and preparation which involves combining the drug response (GDSC2) and the cell line expressions (Array Express) as well as filtering off drugs that don't meet the 500 cell lines requirement. The mean number of cell lines for each drug is about 721. 500 cell lines were chosen as the threshold since the majority of drugs will have enough data for us to train a reasonably good model. Afterward, the data was partitioned into 80% training, 10% validation, and 10% for testing to ensure unbiased model evaluation. Due to the limited data available, more data was used for training and using the 10% for validating our hyperparameters. Finally, the remaining 10% is used for testing the model's performance.

It would be too computationally expensive to explore every combination of feature reduction techniques and every model. Thus, we first need to find which feature selection/reduction technique is the best. We evaluated three different methods for up to 100 components, employing the Random Forest model for assessment. We picked the RF as the baseline performance since it generalizes well and is the fairest for all the reduction techniques. We apply one of the feature reduction methods to significantly reduce the dimensionality while retaining the most relevant information. We explored $k \in \{5, 10, 25, 50, 100\}$, where k is either the top k selected features (PCC) or the number of input features to transform the data into (PCA and NMF). We evaluate the performance using the best feature reduction technique across five different models.

The best feature selection technique is then explored in depth, $k \in \{5, 10, 25, 50, 100, 200, 400\}$ where k is the top k selected features (PCC). We explored hyperparameters specific to the given particular model using 5-fold grid search cross-validation. During cross-validation, each model was validated using the validation set, predicting a natural logarithm IC50 value outputting an MSE depending on how far off the model is. The steps from combining the drug data to validation are then repeated for each unique drug. Finally, results can be collected and analyzed.

5 Results

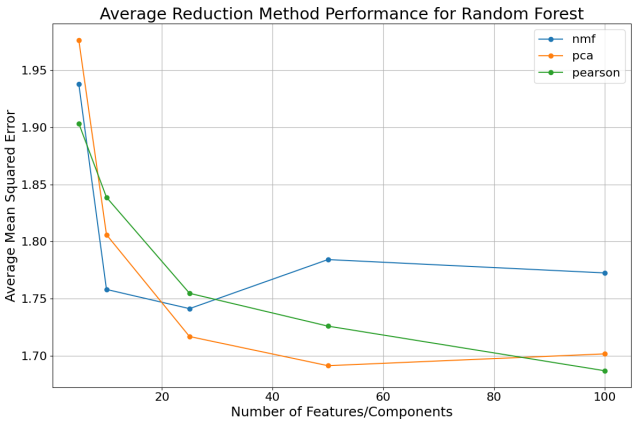

Figure 4: Average MSE of the Three Reduction Techniques using the Random Forest Model

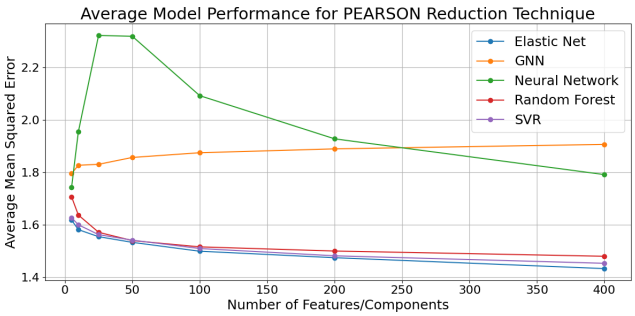

Figure 5: Average MSE across Five Different Models using PCC and E-MTAB-3610 dataset

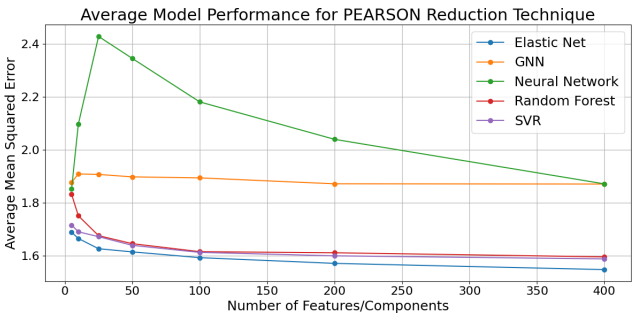

Figure 6: Average MSE Across Five Different Models using PCC and CCLE dataset

To find out which reduction technique is the best, we used a RF model to compare the accuracy between the three reduction technique. We trained for up to 100 components across all drugs in the

E-MTAB-3610 dataset. The results presented in figure 4 demonstrate the comparative performance of three feature reduction techniques—NMF, PCA, and Pearson correlation—when applied to Random Forest models. In figure 4, the performance using PCA and PCC are comparable. We ended up using PCC to further evaluate our models. However, PCC was slightly ahead and so we chose PCC to further evaluate our models. The results as shown in Figure 5 and 6 illustrates the performance of each of the tested models—EN, NN, RF, SVR, and GNN using the E-MTAB-3610 and the CCLE datasets respectively.

5.0.1 Discussion.

Across all methods, the average MSE tends to decrease with an increasing number of features or components. This aligns with the expectation that retaining more features generally preserves more of the original data structure and information. Pearson exhibits the lowest average MSE of 1.69 among the three dimensionality reduction methods, particularly at higher numbers of retained components such as k=100. This indicates that Pearson effectively captures the most important features while discarding noise, leading to better predictive performance. PCA performed well and achieved lowest average MSE from k=25 to k=50, with a respectable MSE of 1.70 at k=100. NMF consistently lags behind PCA and Pearson correlation in terms of Average MSE. Its performance stabilizes at a higher MSE value even as the number of features increases. This suggests that NMF may not effectively capture the essential data structure for Random Forest models in this context.

Model performance varied significantly, with Elastic Net (EN) outperforming other models due to its ability to balance feature selection and regularization, achieving an MSE of 1.43 on the E-MTAB-3610 dataset. While Random Forest (RF) and Support Vector Regression (SVR) delivered stable results, the limited dataset size hindered the effectiveness of neural network approaches, particularly the Graph Neural Network (GNN) and Multilayer Perceptron (MLP). The GNN's performance remained relatively stable throughout the training process as the number of features increased, likely due to its reliance on graph structures rather than raw feature dimensionality. Unlike traditional models, GNNs prioritize the relationships between nodes (genes or pathways) over individual feature values, which can limit the impact of added features on prediction accuracy. To enhance GNN performance, incorporating drug embedding data—such as molecular fingerprints or structural descriptors—could provide additional context about drug-cell line interactions, enabling the model to capture more nuanced relationships and improve predictions.

Our results, shown in Figures 5 and 6, indicate that model performance is similar across RF, SVR, and EN. We observe a consistent decrease in MSE for most models as the number of features increases across both the E-MTAB-3610 and CCLE datasets, reaching optimal performance at approximately 400 features. The NN performed poorly, attributable to the limited size of our dataset, which is insufficient for effectively training a deep learning model. The trend of the NN varied a lot, with the MSE increasing from 1.74 to 2.32 and then decreasing back down to 1.79. The GNN found limited success in both datasets, performing poorly in most drugs but performing very well in a few. Among the models tested, EN demonstrated the best performance across both datasets, achieving the lowest MSE of 1.43 at 400 input features using the E-MTAB-3610 dataset. The average accuracy using the E-MTAB-3610 was much better than the CCLE, given there were more cell lines and features available in comparison.

6 Conclusion

In this study, we evaluated three feature reduction techniques — PCA, NMF, and PCC—for drug response prediction using genomic data. By implementing a robust modular pipeline for data processing, feature reduction, and model evaluation, we were able to systematically analyze the performance of these techniques using a Random Forest model. The pipeline allowed for rigorous benchmarking by leveraging hyperparameter optimization and cross-validation while maintaining a focus on generalization across drugs. Our results indicate that PCA and Pearson correlation are superior to NMF in terms of predictive accuracy as measured by Average MSE. Specifically, PCC demonstrated consistently strong performance, particularly at higher numbers of selected components, achieving an MSE of 1.69 at k=100. PCA correlation showed competitive performance, particularly for smaller feature subsets, while NMF lagged behind, stabilizing at a relatively higher MSE. Our findings highlight the importance of selecting the appropriate feature reduction techniques when working with high-dimensional genomic data.

Moreover, this study underscores the potential of lightweight and interpretable models like Elastic Net in achieving accurate predictions while ensuring computational efficiency. The consistent performance of EN and PCC across diverse datasets demonstrates their robustness and applicability for personalized medicine. However, the limited dataset size constrained the effectiveness of more complex models, such as Neural Networks and GNNs, suggesting the need for larger, more diverse datasets to fully exploit their capabilities.

Future works opens the doors to extend this pipeline by leveraging machine learning frameworks and dimensionality reduction techniques to identify optimal drug therapies for individual cancer patients. By integrating multi-omics data, such as proteomics and transcriptomics, into these models, we can develop more comprehensive representations of cancer biology, enhancing prediction accuracy and personalized treatment strategies. These efforts will bridge computational advancements and clinical applications, advancing precision oncology.

Contributions

Parshva Shah, Aakash Durairaj and Kevin Qian designed and implemented this study. Parshva Shah worked on the implementation of the non-negative matrix factorization dimensionality reduction technique, the elastic net model, and multilayer perceptron architecture. Aakash Durairaj worked on preprocessing the GDSC dataset, the Pearson correlation reduction technique, and the implementation of the SVR, random forest, and GNN models. Kevin Qian worked on the implementation of PCA, all the model training, and gathering the results. Special thanks to Zhijie Wang and Lei Ma for supervising and reviewing the study.

References

[1] [n. d.]. Understanding Genetics: The Basics of Genes. https://medlineplus.gov/genetics/understanding/basics/gene/.

[2] S. Akhondzadeh. 2014. Personalized medicine: a tailor made medicine. *Avicenna journal of medical biotechnology* 6, 4 (2014), 191.

[3] J. Costello, L. Heiser, E. Georgii, et al. 2014. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology* 32 (2014), 1202–1212. https://doi.org/10.1038/nbt.2877

[4] Broad DepMap. 2024. Current DepMap Release data, including CRISPR Screens, PRISM Drug Screens, Copy Number, Mutation, Expression, and Fusions. *Figshare+* (2024). https://depmap.org/portal/data_page/

[5] Lea Eckhart, Kerstin Lenhof, Lisa-Marie Rolli, and Hans-Peter Lenhof. 2024. A comprehensive benchmarking of machine learning algorithms and dimensionality reduction methods for drug sensitivity prediction. *Briefings in Bioinformatics* 25, 4 (Jul 2024), bbae242. https://doi.org/10.1093/bib/bbae242

[6] M. Richter et al. [n. d.]. Cell and Developmental Biology Article. https://www.frontiersin.org/journals/cell-and-developmental-biology/articles/10.3389/fcell.2021.711381/full.

[7] Nicolas Gillis. 2014. The Why and How of Nonnegative Matrix Factorization. *arXiv* (2014). https://blog.acolyer.org/2019/02/18/the-why-and-how-of-nonnegative-matrix-factorization/

[8] Michael Greenacre, Patrick J. F. Groenen, Trevor Hastie, et al. 2022. Principal component analysis. *Nature Reviews Methods Primers* 2 (2022), 100. https://doi.org/10.1038/s43586-022-00184-w

[9] Alexander B. Keenan, Stephanie L. Jenkins, Kathleen M. Jagodnik, Simon Koplev, Ernest He, Daniel Torre, Zichen Wang, Aaron B. Dohlman, Michael C. Silverstein, Alexander Lachmann, and Avi Ma'ayan. 2012. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *BMC Proceedings* 6, S2 (2012), S10. https://doi.org/10.1186/1753-6561-6-S2-S10

[10] W. Koehrsen. 2018. Hyperparameter Tuning the Random Forest in Python. *Towards Data Science* (2018). Available: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74, Accessed: Dec. 11, 2024.

[11] Ultan McDermott. 2011. Gene expression analysis of 789 cancer cell lines using the Affymetrix HT-HG-U133A v2 platform. BioStudies, E-MTAB-783. Retrieved from https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-783.

[12] M. Moo-Young. 2019. *Comprehensive Biotechnology.* Pergamon, Amsterdam.

[13] National Human Genome Research Institute. 2024. Gene Expression - Genetics Glossary | National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Gene-Expression.

[14] Nikolay Pozdeyev, Minsoo Yoo, Robert Mackie, Robert E. Schweppe, Andrew C. Tan, and Bryan R. Haugen. 2016. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget* 7, 32 (2016), 51619–51625. https://doi.org/10.18632/oncotarget.10010

[15] Thanh Ngoc Tran, Quoc Dai Nguyen, and Binh Minh Lam. 2024. Impact of Kernel Functions on Support Vector Machine Models in Classification and Regression Problems. In *Proceedings of the International Conference on Sustainable Energy Technologies (ICSET 2023) (Green Energy and Technology)*. Springer, Singapore. https://doi.org/10.1007/978-981-97-1868-9_80

[16] D. I. Ugwu and J. Conradie. 2023. Anticancer properties of complexes derived from bidentate ligands. *Journal of Inorganic Biochemistry* 246 (Sept 2023), 112268. https://doi.org/doi:10.1016/j.jinorgbio.2023.112268

[17] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* 41, D1 (Jan 2013), D955–D961. https://doi.org/10.1093/nar/gks1111

[18] Yingke Yang and Peiluan Li. 2023. GPDRP: a multimodal framework for drug response prediction with graph transformer. *BMC Bioinformatics* 24, 484 (2023). https://doi.org/10.1186/s12859-023-05618-0