CAN DRUG EFFICACY PREDICTION BE ACHIEVED VIA MACHINE LEARNING?

Sai Vivek Kammari (A1807677)

Supervisor: Mohsen Dorraki, Dr. Zhibin Liao, Dr. Johan Verjans.

Master of Data Science Research Project (7097B)

Australian Institute for Machine Learning (AIML)
University of Adelaide
2021

Keywords

Cancer Cell Line Encyclopedia (CCLE), Cancer research, Cell Lines, Copy number variation, Drug response, Efficacy prediction, Genomics of Drug Sensitivity in Cancer (GDSC), Half-maximal inhibitory concentration (IC50), mRNA, Late integration, Machine learning, Mutation, Multi Omics, Prediction, Response, SMILES, STR Profiles, SNP Profiles,

Abstract

Motivation: Cancer care has evolved at a brisk pace during the last couple of decades. However, the survival rates of some of the most common cancers has not improved proportionally. Cancer drug resistance in patients is one of the prime reasons behind such a trend. This prompted the need to dig deeper into the evaluation of drug toxicity and efficacy in the pre-clinical set up to deliver successful cancer treatment for patients. Predicting the toxicity of a drug in the pre-clinical accurately will significantly improve the survival rate.

Results: We have conducted four major experiments during the duration of the project. For the first three experiments we have used the GDSC drug response dataset published by the GDSC project and predicted the efficacies using various boosting models. We have used the mRNA expressions and copy number variations as the cell line features and extracted the drug features using the SMILES and PaDelpy python library. The best results were achieved using the stacked regression model proposed in the study in all the experiments. The proposed model is a combination of XG Boost, LG Boost and CatBoost with LASSO as a meta learner. The efficacy predictions vary with the cell line features used as predictors. The multi omics representation of the cell lines using the mRNA and copy number variations produced the best predictions (R2: 0.852 & MSE: 1.119) followed by predictions achieved using the single omics representation of cell lines (R2: 0.851 & MSE: 0.130) followed by the predictions achieved using the STR and SNP profiles representation of the cell lines (R2: 0.812 & MSE: 1.418).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	**Explanation**
ANN	Artificial Neural Networks
CCLE	Cancer Cell Line Encyclopedia
DNA	Deoxyribonucleic Acid
GDSC	Genomics Of Drug Sensitivity in Cancer
IC50	Half-Maximal Inhibitory Concentration
LASSO	Least Absolute Shrinkage and Selection Operator
LGB	Light Gradient Boosting
mRNA	Messenger RNA
PDX	Patient-Derived Xenograf
SMILES	Simplified Molecular-Input Line-Entry System
SNP	Single Nucleotide Polymorphisms
STR	Short Tandem Repeats
XGB	Xtreme Gradient Boosting

Acknowledgements

First and foremost, I am extremely grateful to my supervisors, Mohsen Dorraki, Dr. Zhibin Liao and Dr. Johan Verjans for their invaluable advice, continuous support, and patience during my research study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. My sincere thanks especially to Mohsen Dorraki who has guided me right from the scratch of the research.

My gratitude extends to the AIML and SAHMRI for the support during the research. Lastly, I would like to thank the University of Adelaide for the studentship that allowed me to conduct this thesis

Chapter 1: Introduction

The Precision oncology is the most modern therapy in the field of cancer care and is perceived to be one of the most effective and innovative approach in cancer treatment. Precision oncology can be defined as developing a personalized therapy for an individual based on their genetic profile. The outcome of any cancer treatment depends on various parameters such as the drug's chemical composition, the genetical composition of the patient, external environmental conditions, pre-existing conditions etc. Most recent studies suggest that only a small minority of the patients are benefited from the targeted therapy or precision oncology, more specifically the percentage is around 5% of the total patients [1]. We can state a plethora of reasons owing to such a significantly low percentage of success however, we strongly believe that improved predictions of drug outcomes in a preclinical setup will lead to a significant improvement in the success rate.

In the most recent times, various studies of cancer have developed publicly available large datasets. These studies are majorly carried out on the cell lines or patient-derived xenograft (PDX) mice models [2]. Few of the publicly available datasets are the Genomics of Drug Sensitivity in Cancer (GDSC) [3] and Cancer Cell Line Encyclopedia (CCLE) [4]. These datasets consist of multi-omics representation of the cell lines or patients and their responses when treated with a range of drugs. These datasets help the researchers to further dig deeper and come up with more meaningful interpretations and computational methods to improve the drug response predictions.

The maximum impact or response a drug can produce in a patient is dubbed as the drug efficacy. To measure the drug efficacy, IC50 is the standard metric used. IC50, also known as the half-maximal inhibitory concentration, can also be defined as the concentration of the drug required to reduce a specific biological process or component by 50 percent [5]. IC50 is the general and approved metric in the pharma industry to measure the efficacy or potency (amount of the drug required to produce the desired effect) of a drug [6].

1.1 AIM/OBJECTIVE

The Objective of the research is to check if machine learning methods can be engaged to predict the efficacy of the anti-cancer drugs using the publicly available datasets such as the GDSC and CCLE.

1.2 MOTIVATION

Technology has evolved at a rapid pace during the last couple of decades and simultaneously the usage of technology in the field of medicine increased significantly. Medical Sciences has touched new avenues with increased usage of technology. Cancer research has moved in a new direction of predicting the drug outputs and drug efficacy in the pre-clinical stages of drug development. Despite of such huge developments in the cancer care, the improvement in the overall patient survival rates of some of the most common cancers is not significant. Studies which delved deep into the reasons holding back the patient survival rates even after introducing new developments into cancer treatment have highlighted that cancer drug resistance in patients [7] is one of the prime reasons behind such a trend. This prompted the need to dig deeper into the evaluation of drug toxicity and efficacy in the pre-clinical set up to deliver successful cancer treatment for patients. Also, unacceptable drug toxicity is one of the prime reasons why the drugs fail in the clinical trials or withdrawn from the market, and this involves huge cost, predicting the toxicity of a drug in the pre-clinical set up will save time, effort, and money.

On the drug development side, rapid developments in the modern medicine have brought us to the current stage where many of the diseases can be treated reducing the number of deaths caused. At the same time, the challenges in the field of drug discovery are also increasing at a significant pace. Eroom's law explains the observations or trend in the drug development, it asserts that, though there have been many developments in the fields of science and technology circumscribing the drug discovery process (such as DNA sequencing Bioinformatics, molecular biology) the efficiency of the drug discovery research and development has not been in increasing trend. Eroom's Law states that the number of new drug discoveries per $1 bn investment in research and development is following a declining trend [8].

In the current situation it is estimated that the pre-tax research and development investment for the development of each new drug is over $2.5 bn [9], this involves all the drugs which fail in the clinical trials. Simultaneous developments in the field of machine learning, bioinformatics and pharmacology have resulted in a situation where the availability of data has grown, and the power of machine learning algorithms improved beyond performing some basic tasks. This led to machine learning evolution in the drug discovery.

The Figure 1 represents the Eroom's law where number of new drugs discovered per $1 bn investment in the research and development is plotted on the y-axis and the x-axis represents the timeline. We can clearly notice from the figure that the number of drugs discovered per $1bn investment is declining at a compounded rate of 8.4% per year.

1.3 RESEARCH QUESTIONS

As discussed in the above sections, the primary objective of the research is to check if machine learning methods can be engaged to predict the efficacy of the anti-cancer drugs. We will develop a machine learning model which can predict the efficacy of a given drug and cell line pair.

1.4 THESIS OVERVIEW

The thesis is developed as the final part of the work done during the final year research project of the master's coursework and will be submitted for academic purpose. The thesis covers all the details of the project 'Can drug efficacy prediction be achieved using machine learning' across various sections of this report; beginning from the initial background research done, the detailed description of the datasets used for the project, the data pre-processing methodologies employed, detailed description of the machine learning methodologies explored and the model finalized, various experiments conducted during the research work and their results and the work planned for the future.

The thesis is developed as the final part of the work done during the final year research project of the master’s coursework and will be submitted for academic purpose. The thesis covers all the details of the project ‘Can drug efficacy prediction be achieved using machine learning’ across various sections of this report; beginning from the initial background research done, the detailed description of the datasets used for the project, the data pre-processing methodologies employed, detailed description of the machine learning methodologies explored and the model finalized, various experiments conducted during the research work and their results and the work planned for the future.

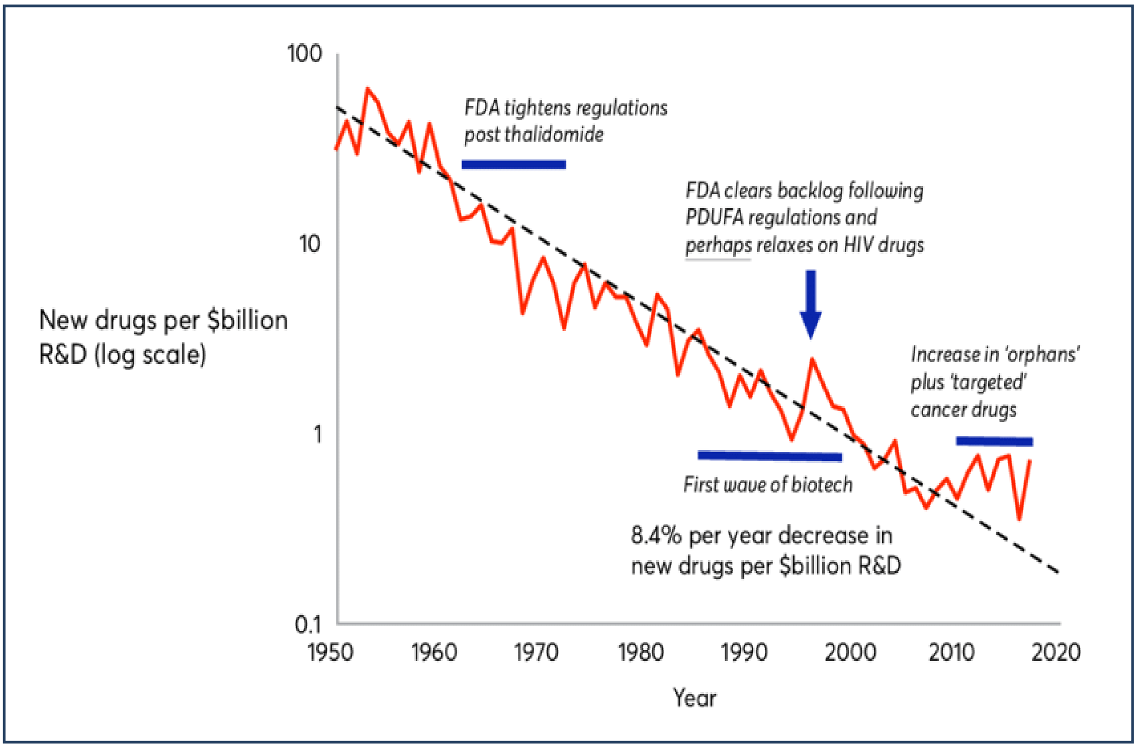


Figure 1: Eroom’s law [9]

Chapter 2: Literature Review

2.1 LITERATURE REVIEW

Machine learning became a part of the cancer research since mid-1980 [10]. Over the period, the number of clinical factors to be analysed increased significantly and it became a tough job for the doctors to integrate these factors with the data from different subsets of biomarkers. Machine learning provided them the ability to combine these datasets to analyse [11].

Cancer prediction and prognosis using machine learning started in the mid 2000's, prior to that, machine learning was majorly used for diagnosis or detection of cancer. In 2014, the study was conducted with its primary objective to review all the machine learning techniques such as Artificial neural networks (ANNs), Bayesian networks (BNs), Support vector machines (SVMs) and Decision trees (DTs) which were widely used in the detection of cancer, understanding the progress of the cancer in patients and classify the cancer patients into high risk and low risk which will help better clinical management of patients. The study found that the usage of support vector machine (SVM's) and semi supervised machine learning techniques using labelled and unlabelled data are gaining traction in predicting the outcomes and modelling cancer survival. The study also highlighted the need of research in the direction of constructing public database of cancer patients [12].

With the rapid advancements in technology, cancer research has moved in a new direction of predicting the drug outputs and drug efficacy in the pre-clinical stages of drug development. In 2012, in one of its early efforts to formulate and integrate drug efficacy prediction into preclinical setting, a study profiled 947 cancer cell lines at the genomic level and generated compound sensitivity data for 479 cell lines. The researchers used an automated response screening platform to generate eight-point dose– response curves for 24 anticancer drugs. The study used naive bayes classifier and the elastic net regression algorithms as predictive models on the drug sensitivity data [13]. By compiling the gene expressions, chromosomal copy number and massively parallel sequencing data from 947 human cancer cell lines, the research work generated something called as the Cancer Cell Line Encyclopedia (CCLE).

The study also published the drug responses of 479 cell lines when treated with 24 anti-cancer drugs. The study identified genetic, lineage and gene-expression based predictors. The study also highlighted that it has identified few more predictors in addition to the already known list of predictors of sensitivity and stated that IGFI receptor sensitivity is correlated with plasma cell lineage, MEK inhibitor efficacy in NRAS- mutant lines is correlated with the AHR expression and the sensitivity of topoisomerase inhibitors sensitivity is correlated to SLFN11 expression.

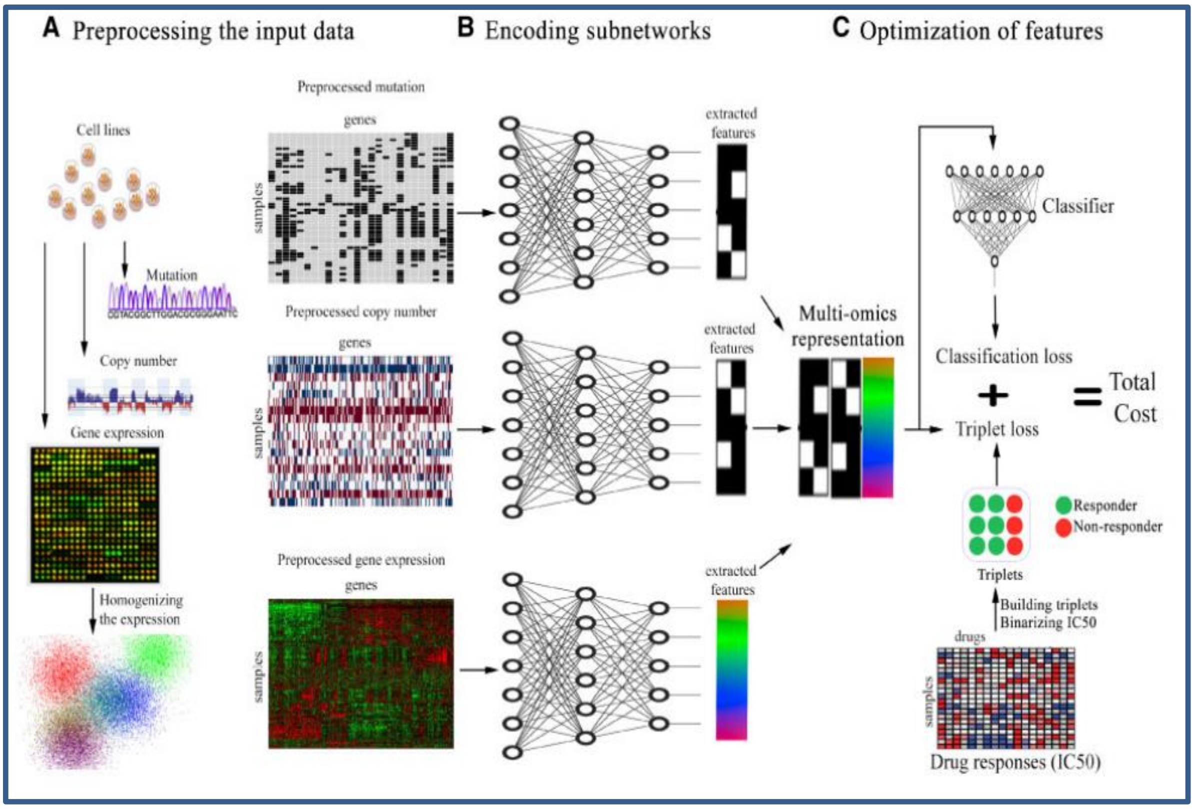
Many computational approaches were put forward to predict the drug efficacy and sensitivity prior to 2013. Some of these approaches were based on drug's chemical properties and some were based on genomic features. However, a study in 2013 attempted to predict cancer cell sensitivity to drugs integrating the approaches suggested earlier. The study used Genomics of Drug Sensitivity in Cancer (GDSC) screening data of 608 gnomically characterized cell lines in combination with chemical information of 111 drugs as data and a feed forward multi-layer perceptron with sigmoid activation function implemented in java for each drug to predict the IC50 curve [14]. The results published by the study can be considered as extraordinary at that point of time. The most important point is that the study claimed that it can predict the efficacy of a drug with unseen cell line with high confidence. The study also claimed that the results of the study go beyond virtual screening off drugs and can be used to predict responses of drugs out of the lab by using computational framework.

As the importance for precision medicine gathered prominence in cancer care, research in evaluation of drug efficacy and sensitivity continued and more sophisticated approaches were developed. In 2018, a study used a novel deep learning model dubbed Cancer Drug Response profile scan CDRscan) to predict the drug outcomes or response. CDRscan is an ensemble of 5 CNN's with different architecture and functionalities. The data used in the study is, structural profiles of 244 anticancer drugs and genomic profiles of 787 human cancer cell lines [15]. The study processed the genomic features separately and the molecular fingerprints of drugs separately using two step convolutional neural networks and later these two were merged using virtual docking technique the best results reported by the study are $R2 > 0.84$; $AUROC > 0.98$. As a proof of concept, the proposed CDRscan deep learning model was applied to 1,487 approved drugs and noted that 37 drugs (14 oncology and 23 non-oncology) have new potential cancer indications. The study used both GDSC and CCLP databases for training and the cell lines were represented my genetic mutation profiles (28,3278), the drugs were represented by 3072 molecular features extracted using the PaDEL library. CDRScan also claimed that it can be used to select the best suited drug using genomic profile of a patients by further clinical validation.

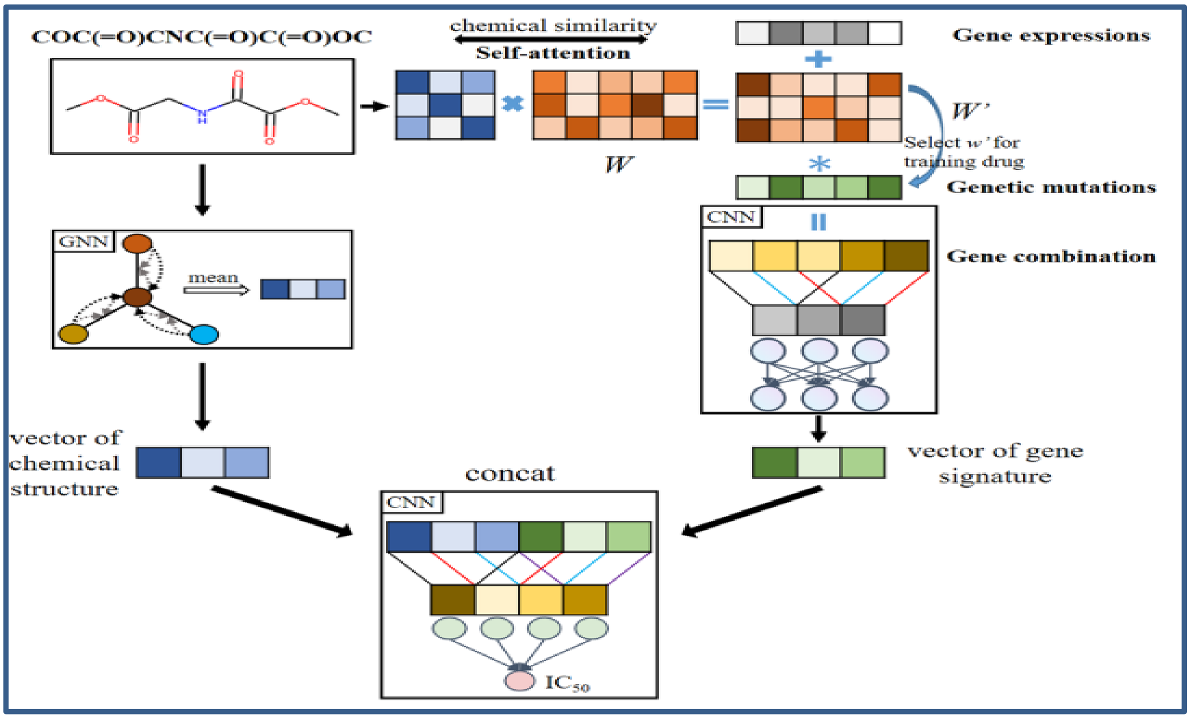
A recent study to predict the oncologic outcomes for drugs in randomized clinical trials used a combination of clinical trial, drug-related biomarker, and molecular profile information to make predictions using the random forest algorithm [16]. A variable dubbed Probability of Drug Sensitivity or PDS was defined as a numeric value of the degree of match between the molecular signature of the disease and drug regimens and their biomarkers. Overall, the study shows that PDS correlates with the clinical outcome. The study used random forest algorithm.

A study conducted in 2019 tried to increase the drug response prediction by integrating multi omics [17]. For this purpose, the study used Genomics of Drug Sensitivity in Cancer (GDSC) screening data as the training data and PDX Encyclopedia mice models, TCGA patients without the drug response and TCGA patients with the drug response datasets for validating the results produced. The study has developed a deep neural network machine learning model dubbed as MOLI (multi-omics late integration) for the purpose of integrating gene expression data, somatic mutation, copy number aberration datasets and predicting the drug response. The study found that the MOLI predicted with more accuracy in six out of seven external validation datasets and outperformed in all the seven external validation datasets when compared with the early integration using NMF (non-negative matrix factorization).

A recent research work dubbed ‘SWNet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures’ published its results and claimed to outperform all the existing models. The study used the Genomics of Drug Sensitivity in Cancer (GDSC) and The Cancer Cell Line Encyclopedia (CCLE) datasets. The research used SMILES of the drugs and obtained their Morgan fingerprints using the RDKit. The model integrates mRNA expression, genetic mutation and drugs chemical features using a multi task convolutional architecture.



a) MOLI



b) SWNet

Figure 2 Machine learning algorithms used in the previous research studies discussed in this paper

Chapter 3: Research Design

3.1 METHODOLOGY AND RESEARCH DESIGN

The research question can be answered by predictive modelling approach using python software. To measure the drug efficacy, we will consider the IC50 value. Note that, IC50 also known as the half-maximal inhibitory concentration is the general and approved metric in the pharma industry to measure the efficacy or potency (amount of the drug required to produce the desired effect) of a drug. IC50 can also be defined as the concentration of the drug required to reduce a specific biological process or component by 50 percent. 10uM(micro molar) is the general perceived benchmark for IC50 value i.e, in general up to 10 uM of any drug is accepted in plasma and above this value the drug starts to inhibit enzymes which are necessary. The machine learning model which we explore will predict the IC50 value and hence the efficacy of the anticancer drug.

Figure 3 shows the research design of the project. From the experiments in the laboratory, we get genomics data such as RNA sequencing, Copy number variation and Mutation. For obtaining the drug data we use SMILES ('simplified molecular input line-entry system is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings'[18] to extract the drug features using the PaDelpy library in python.

Each drug can be represented by 2756 features, of these 1875 are descriptors (1444 1D, 2D descriptors and 431 3D descriptors) and 881 fingerprints. In simple terms we can say drug descriptors and fingerprints are nothing but the numeric representation of drug's chemical properties and molecular structure [19].

Post obtaining the drug features and genomic data (mRNA, Copy number, Mutation) we integrate these two datasets to develop the integrated representation of the sample. The integration can be achieved using the following two techniques.

Early Integration: In this technique, all the available datasets are first concatenated, and this creates an integrated representation of the sample.

A

Late Integration: In this technique, some important features are learned from each dataset and these features are put together to represent each sample. This representation of each sample is then used in various machine learning models. These integrated representations of the samples are now used to predict the drug efficacy i.e IC50 values. The prediction is done by using various machine learning models. Figure 4 shows the various machine learning models or steps that have been followed during the project to obtain the objective.

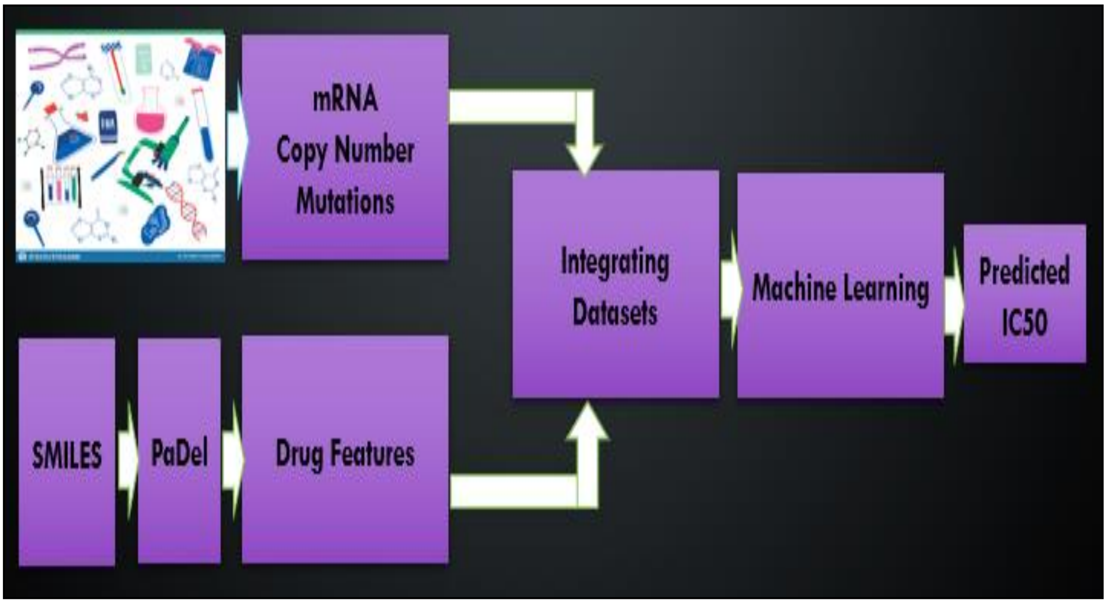


Figure 3: The research design of the project.

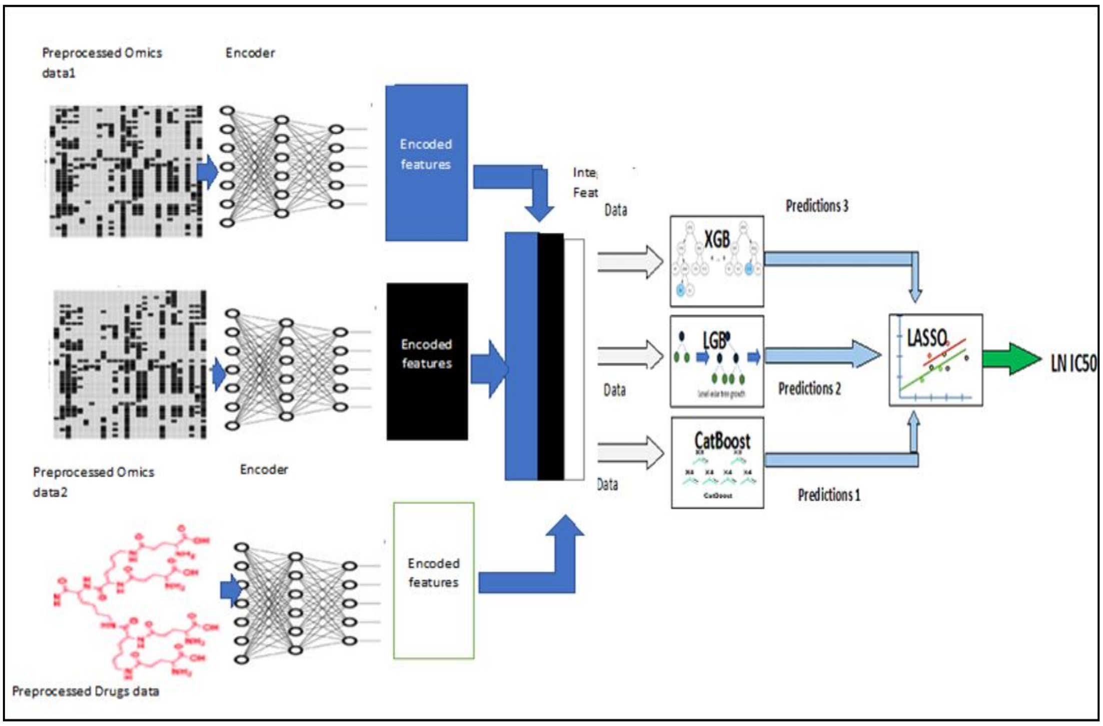


Figure 4: Multi Omics predictions architecture.

3.2 DATASET UNDER ANALYSIS

For answering the research question, we will be using the following datasets:
GDS cell lines dataset [3]: Cancer Genome Project at the Wellcome Sanger Institute (UK) and the Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center (USA) together fund The Genomics of Drug Sensitivity in Cancer (GDSC) project. The study publishes the data of around 1000 Cell lines and their response when treated with around 360 anti-cancer drugs. The study also publishes the Cell lines genomic data (mRNA, Mutation, Copy Number Variation & Methylation), STR profiles and SNP data. The dataset contains the data of around thousand cell lines of around 33 types. Cancer Cell Lines are cells which are allowed to continue dividing and growing under some specific conditions in a laboratory, these are primarily used for the cancer research purpose [20]

STR Profiles: Short Tandem Repeats (STR) Is the new DNA profiling standard currently used across the world. This is a faster and cheaper way of DNA profiling. In this technique, unlike the conventional RFLP technique, just specific portions of the junk DNA are studied. Technological advances in the field of molecular biology have led to the usage of STR DNA profiling to uniquely identify human cell lines.

SNP Data: Single nucleotide polymorphisms (SNP) is also one of the most common type of genetic variation between individuals. A SNP in simple terms is the variation of a single nucleotide between individuals. Hence these SNP’s can be used to distinguish minor differences both within a population and among different populations [21].

CCLE: The Cancer Cell Line Encyclopedia Project is a collaboration between the Broad Institute and the Novartis Institutes for Biomedical Research. The project was initiated in 2006 to characterize over 100 cell lines. The CCLE currently publishes data of around 479 cell lines and their mRNA expressions data. The study also published drug responses of 479 cell lines when treated with 24 anti cancer drugs. The drug response data consists of over 10,000 sample, each sample represents a tuple of drug, cell line and the response.

3.3 EXPLORATORY DATA ANALYSIS

The GDSC drug response dataset is considered as the base dataset which we used to train the models. The shape of GDSC drug response dataset was (310904, 19) and consists of several attributes to uniquely identify the cell line such as “COSMIC_ID”, “CELL_LINE_NAME” and also various attributes to uniquely identify the drugs such as “DRUG_ID”, “DRUG_NAME”. All the drugs belong to six different companies with company ID’s 1045, 1019, 1046, 1001, 1025 and 1009 (some of the drugs belonged to multiple companies). Other important attributes in the GDSC dataset were the drug response attributes “LN_IC50” and “AUC”. The GDSC project measures the IC50 value of every drug when used against a particular cell line and plots the IC50 curve by varying the concertation of a drug. Figure 6a, b,c shows the various exploratory plots used to understand the data distribution of GDSC drug response. A Heatmap has been plotted to understand the correlation between various features of the GDSC drug response dataset. We could not notice any significant correlation between any features of the GDSC drug response dataset

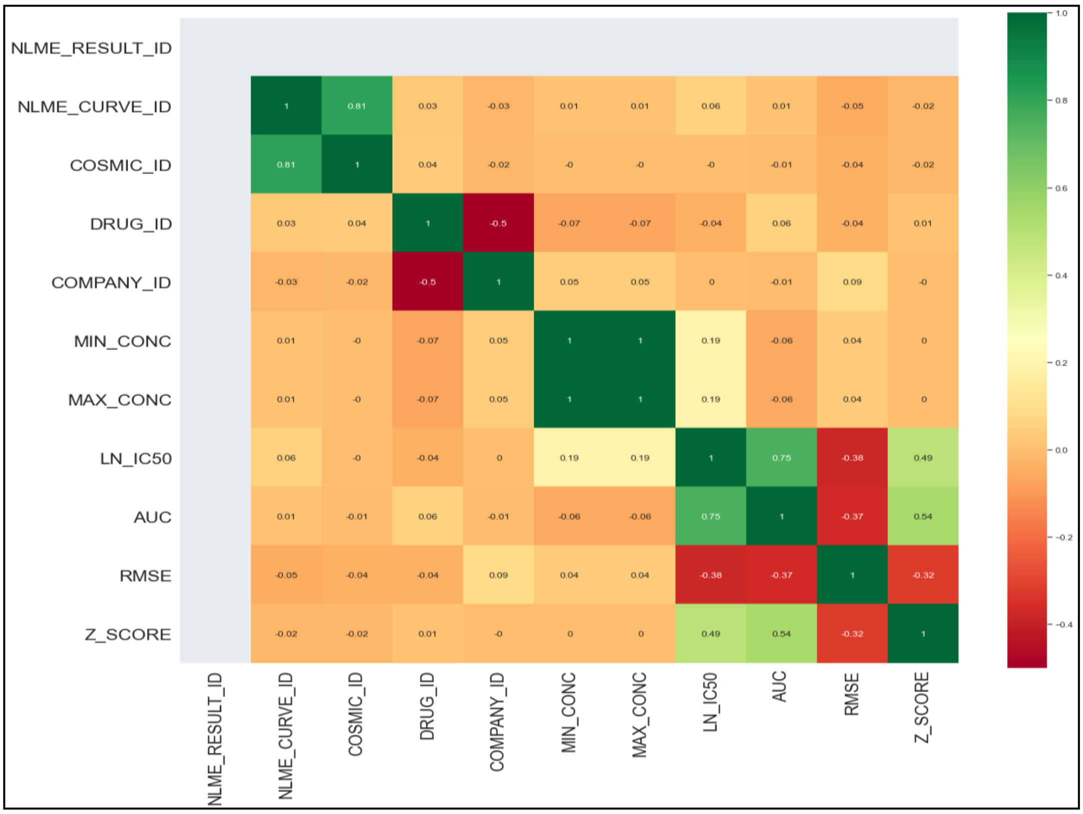


Figure 4: Heatmap of GDSC data

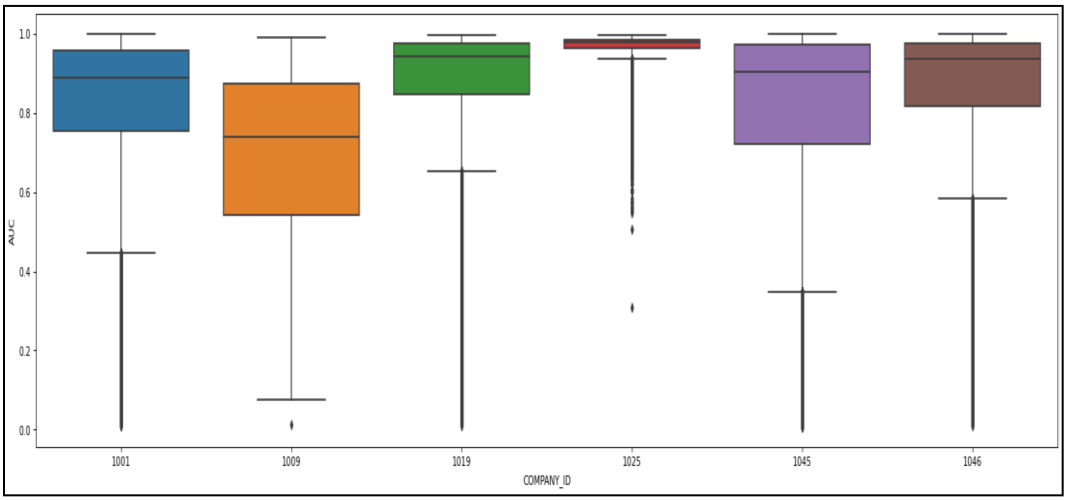


Figure 5: Piano plot of AUC values

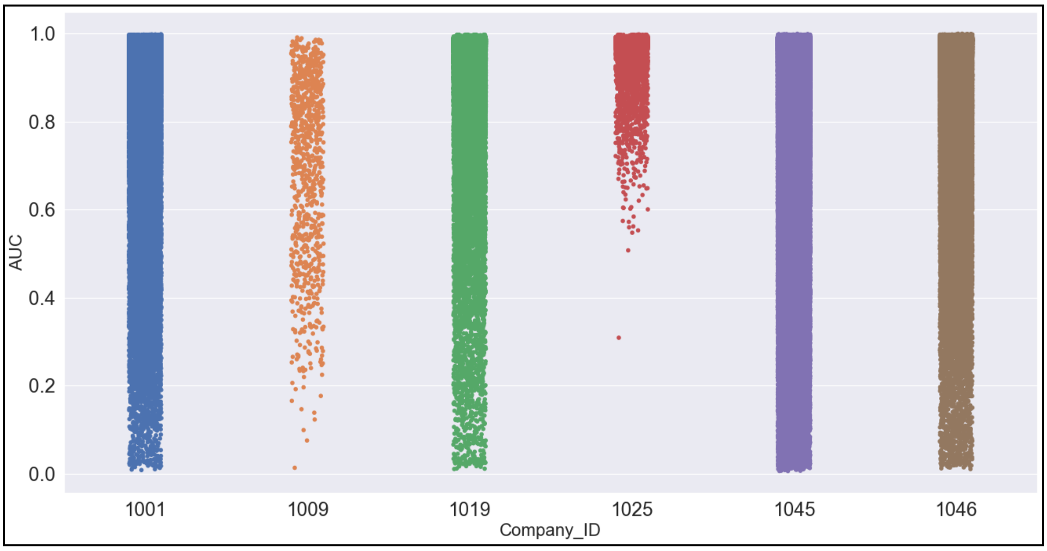


Figure 6: Box plot of AUC values for all the six different companies

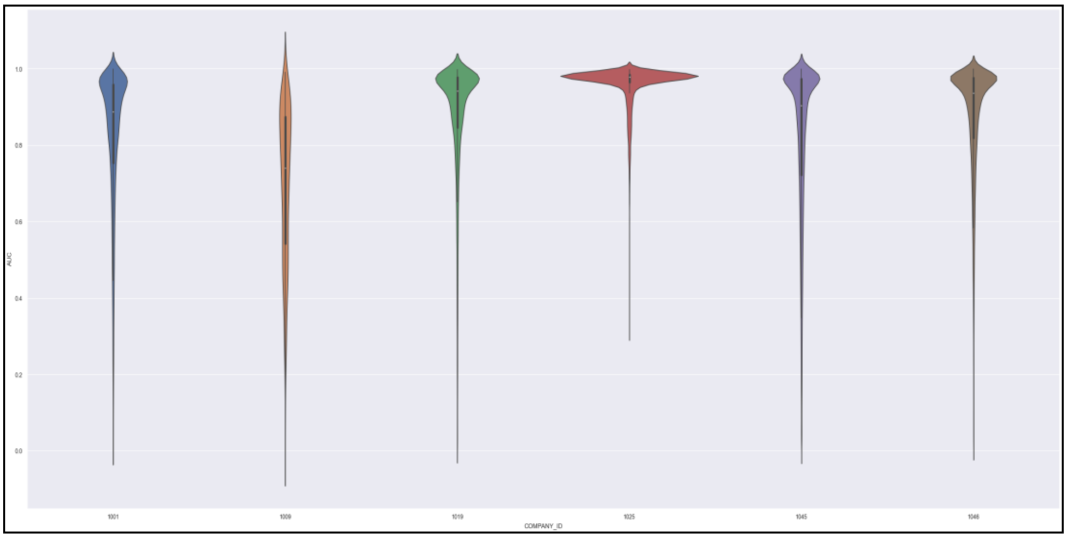


Figure 8: Strip plot of AUC values for all the six different companies

Figure 6,7 and 8 are the various plots to explore the dataset. We have plotted the AUC values of all the drugs of all the six companies to get deeper insights of the data distribution completion of each stage.

3.4 DATA PREPERATION

We have conducted four different experiments in our project and for the purpose of first three experiments. The first three experiments were conducted using the GDSC dataset and the fourth experiment was conducted using the CCLE dataset. The GDSC drug response dataset had 367 unique drugs when we used “DRUG_ID” as the unique identifier and 345 unique drugs when we use “DRUG_NAME” as the unique identifier. In all the experiments we have used the DRUG_ID as the unique identifier and assumed that each DRUG_ID represented a different drug. The GDSC drug response dataset had 987 cell lines with “CELL_LINE_NAME” and “COSMIC_ID” as unique identifiers.

Single Omics data: The main idea in the process of data preparation is to replace the cell lines and drugs with their respective features. In the GDSC dataset The final dataset will have records with drug features, cell line features and drug response. For representing the cell line features we have used mRNA expression dataset, so the idea is to replace every cell line in the GDSC drug response dataset with its respective mRNA expression. The mRNA expression dataset had 1047 cell lines and each cell line is represented by 37279 features, so the mRNA dataset shape was (37279,1047) where each row represents a gene and each column represents a cell line. Some basic pre processing steps such as filling the blank entries with zero were performed on the mRNA dataset. We have noted that the of the 987 cell lines in the GDSC drug response dataset, only 980 were also present in the mRNA dataset and hence we have considered these 980 cell lines for the experiments. We have first transposed the mRNA dataset to reshape it (1047,37279). Then reduced the size of the mRNA dataset from 297mb to 37mb by carefully checking and parsing the data types from float64 to float16, int64 to int 16 and uint8 where we could do, this conversion took more than four hours of processing. In the next step we developed an autoencoder using keras and tensor flow to reduce the dimensions of from 37,279 to 100 encoded or learned features.

In the next step we have processed the drugs data. We have obtained the SMILES of 223 drugs and used them to extract the features using the PaDelpy library in python. The drugs dataset has 2756 features and similar to the mRNA dataset, we have reduced the dimensions of using autoencoder to 50 dimensions. In the next step we have replaced the cell lines with learned 100 features of mRNA expression and the drugs with the 50 learned features in the GDSC data and now the single omics dataset is ready for predictive modelling.

Multi Omics data: For representing the cell line features we have used mRNA expression dataset and the copy number variations dataset, so the idea is to replace every cell line in the GDSC drug response dataset with its respective mRNA expression and copy number variation. During the preparation of single omics dataset we have processed the mRNA expression dataset and finally derived 100 learned features. Now, we will apply similar process to the copy number variations dataset and finally using similar autoencoders will learn 100 features. The copy number variations dataset published by GDSC has 987 cell lines and each cell line is represented with 24,502 features. The 24,502 dimensions of the copy number dataset are reduced to 100. Now, the 100 mRNA features and the 100 copy number features are concatenated to form the integrated multi omics representation of each cell line. In the next step we have replaced the cell lines with learned 200 features (of which 100 where mRNA expression features and 100 where copy number variation features) and the drugs with the 50 learned features in the GDSC data and now the multi omics dataset is ready for predictive modelling. in

The GDSC project also published the STR profiles and SNP data of the cell lines used the GDSC drug response dataset. The main idea of this data preparation task is to replace each cell line in the GDSC drug response dataset with STR profiles and SNP data (each cell line is represented by concatenating the STR profile and SNP data). We have first created an integrated representation of every cell line by concatenating its STR profile and SNP data. We have performed some basic pre-processing steps on this dataset. In the next step we have replaced the cell lines with the combined STR and SNP data. The integrated representation of the cell line now has 130 features. Out of these 130, 104 were categorical features and we have converted them using one hot encoding feature in python.

These 104 features were converted to 303 features. These new features and the drugs dataset with the unreduced dimensions were merged to build the integrated representation of GDSC records. The final dataset had 2,545 dimensions. In the next step we have reduced the dimensionality of the complete dataset to 150 features using autoencoders built using keras and tensorflow

.

3.5 PREDICTION ALGORITHM

Outline We have used some of the conventional machine learning predicting algorithms such as XG Boost regressor, Light Gradient Boosting and Cat Boost to predict the IC50 values. All the models we have explored belong to the class of boosting algorithms. Boosting is a technique in which a certain weak learning algorithm is scaled up to make the training error zero. In general, boosting is not a single algorithm it is a family of algorithms.

We have used all the three models on each of the datasets during the experiments and noted the results of all the models. In majority of the cases CatBoost performed better than the extreme gradient boosting and the Light gradient boosting algorithms. Furthermore, we have strived to improve the predictions by combining these three algorithms. For combining the three boosting algorithms we have followed two methods.

Averaged Model: In this methodology we have taken the three boosting models and trained them on the same train data and made three different predictions using the three boosting algorithms. The final predictions were calculated by simply taking the average of the three predictions.

Final Predictions = Average (XGB predictions, LGB predictions, CatBoost predictions)……………………………………………………………(3.1)

We have noticed that the averaged model outperformed all the three boosting algorithms in all the experiments.

Meta Learner: In this methodology, similar to the averaged model, we have trained all the three boosting algorithms on the same train data and made three different predictions using the three algorithms. The three predictions and the actual output were put together to form new data and a LASSO regressor was trained on this data and the final predictions made by the LASSO are taken as the final predictions.

Final Predictions = LASSO (XGB predictions, LGB predictions, CatBoost predictions)…………………………………………………………….(3.2)

the ethical considerations of the research and any [potential] problems and limitations (weaknesses), as well as any [anticipated or actual] threats to the validity of the results.

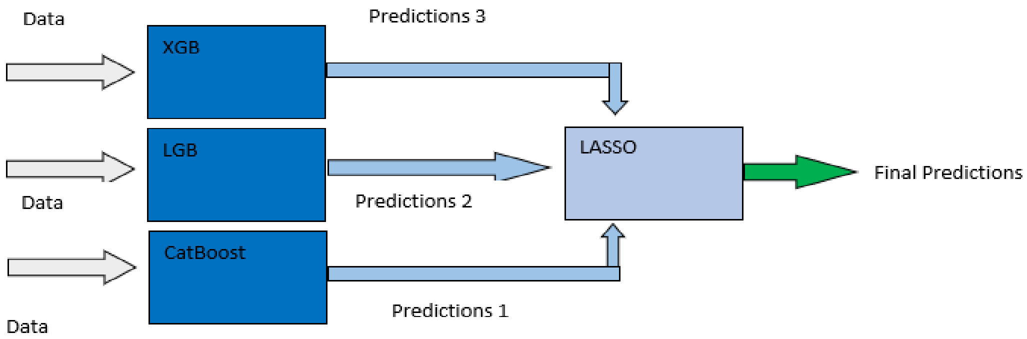


Figure 7 : Architecture combining the three boosting algorithms with a LASSO meta-learner.

In all the experiments we have noticed that the meta learner has outperformed all the previous models and hence we have chosen the meta learner as the final machine learning model for predicting the efficacy values.

Cross Validation: Cross validation is a resampling methodology which is crucial in generalizing the performance of a machine learning model. A machine learning model can perform well on a given test data set once it is trained however, we cannot generalize and say that the machine learning model will perform similarly on any given test data. In machine learning it is very important to generalize our final model, that is the performance metrics which we present at the end of the project should be in the same range when the model is used against any unseen test data.

In cross validation, the training and test datasets are randomly shuffled and the model's performance on each test set is measured. General rule is that the average of the model's performance can be stated as the general performance of the model. In our project we have used five-fold cross validation, that is the data is split into five equal segments and at every iteration one segment is chosen as the test data set and the remaining four segments are combined to form the train data set. In case of large datasets, we can still be sure of the model performance through using one validation dataset however, to strengthen our claims we have performed a fivefold cross validation.

3.6 EVALUATION METRICS

A good performing model should predict the values close to the actual observations. In case if we do not have sufficient information to make predictions, we will use the mean model which generally uses the mean of every predicted value. The machine learning model we use for making predictions should perform better than the mean predictive model. The predictions achieved by using various models and the performance of the predictive model can be measured by using various parameters. In our project we will be using the R Squared value and RMSE to evaluate the model's performance.

Before going to the evaluation metrics, first let us understand the Sum of Squares Total (SST) and Sum of Squares Error (SSE). SST gives a measure of how far the actual data is from the mean and similarly SSE gives a measure of how far the predicted values are from the actual observations.

R-Squared: R-Squared is the proportional improvement achieved via the regression model from the mean model. It is calculated by dividing the difference between STT and SSE (improvement in prediction via the regression model when compared to the mean model) by SST. R-Squared is not an absolute measure, it is a relative metric which ranges from 0 to 1. R-Squared value closer to 1 implies good predictions and 0 implies that the predictions has not improved from the mean model.

$$R^2 = 1 - \frac{SSres}{SStotal} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(3.3)}$$

MSE: Unlike R-Squared, Mean Squared Error on Prediction is an absolute measure of the model fit. It is an absolute value of how close the predictions of the model are to the actual data or observations. A predictive model which has lowest MSE is the best fit.

$$MSE \quad = 1/N \sum_{k=0}^{n} (yi - \ddot{y})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.4)}$$

yi is the actual value and ÿ is predicted value.

Chapter 4: Results

4.1 RESULTS OVERVIEW

In this project, we have used two major datasets namely the GDSC dataset and the CCLE dataset. GDSC is the primary dataset used to develop, train and test the machine learning model. In the next stage, we have implemented the finalized model using the CCLE dataset. Hence the results section is broadly classified into a) GDSC results and b) CCLE results. Furthermore, we have performed three experiments using the GDSC data to evaluate the model's performance in three different scenarios and hence the GDSC results has three subsections.

4.2 GDSC RESULTS

Single Omics: As described in the dataset section of the report, the GDSC dataset contains of data pertaining to the cancer cell line, drug, and drug response (LNIC50). Each cell line is represented by its gene expression (Each cell line is represented by 37, 279 genes) The resultant final dataset had 158 features. We have, split these samples into training and validation samples in 90:10 ratio.

From the Table1 and Table 2 we could notice that the lowest MSE(1.130) was achieved by the meta learner model and the best R2 value of 0.851 is also obtained for the meta learner model.

Multi Omics: In this experiment we have represented the cell line by combining two omics features those are the mRNA expression features and the copy number variations. The integrating of these datasets can be done in two ways, in the early integration technique, all the available datasets are first concatenated, and this creates an integrated representation of the sample. In late integration technique, some important features are learned from each dataset and these features are put together to represent each sample. This representation of each sample is then used in various machine learning models. We have used the early integration approach two combine the mRNA features and the Copy number features. The resultant final dataset had 258 features. We have, split these samples into training and validation samples in 90:10 ratio.

From the Table 3 and Table 4 we could notice that the lowest MSE(1.119) was achieved by the meta learner model and the best R2 value of 0.852 is also obtained for the meta learner model.

Another important observation was that the model performance almost remains at the same range for both multi omics and the single omics data representation of the cell lines. However, by using the multi omics data which uses mRNA expression and copy number variation data as features of cell lines, we could obtain better predictions.

STR & SNP profiles: In this experiment we have used the STR profiles, SNP data and Drug features extracted using the SMILES and PaDelpy library in python. The data preparation is described in detail in the section 4.4. We have used this dataset to predict the IC50 values using the same set of machine learning models. We have first created an integrated representation of every cell line by concatenating its STR profile, SNP data and MSI details published by the GDSC, this is similar to the early integration representation of cell lines for multi omics approach. The main idea behind this experiment was to gauge the amount of information is encoded in the given STR and SNP profiles of the cell lines.

From the below results tables we could notice that the results produced are still very significant, but the predictions are relatively weak when compared with the results obtained used the single omics or multi omics representations of the cell lines.

From the Table 5 and Table 6 we could notice that the best MSE(1.418) and the best R2 value of 0.812 are obtained by using the meta learner model.

4.3 CCLE RESULTS

The CCLE drug response dataset contains 10,853 samples, each sample is a tuple of the cell line name, drug name and the drug response. The main idea of the data preparation here is to replace the cell line with the mRNA expression (1468 genes) and the drugs with the encoded features. The final dataset had 150 features and 10,853 samples. We have split the data in 90:10 ratio for training and testing.

Table 1: Mean squared error of five fold cross validations using single omics data for various models

Model	*Five Fold Cross Validation Results (MSE)*					*Mean MSE*
XGBoost	1.195	1.161	1.193	1.178	1.187	**1.183**
LGBoost	1.235	1.198	1.238	1.224	1.227	**1.225**
CATBoost	1.160	1.129	1.163	1.148	1.156	**1.151**
Averaged Model	1.161	1.127	1.162	1.148	1.154	**1.151**
Meta Learner	1.141	1.108	1.138	1.129	1.135	**1.130**

Table 2: R squared of five fold cross validations using single omics data for various models

Model	*Five Fold Cross Validation Results* $((R^2)$					*Mean* R^2
XGBoost	0.842	0.846	0.842	0.846	0.844	**0.844**
LGBoost	0.836	0.841	0.836	0.840	0.838	**0.838**
CATBoost	0.846	0.851	0.846	0.850	0.848	**0.848**
Averaged Model	0.846	0.851	0.846	0.850	0.848	**0.848**
Meta Learner	0.849	0.853	0.849	0.853	0.850	**0.851**

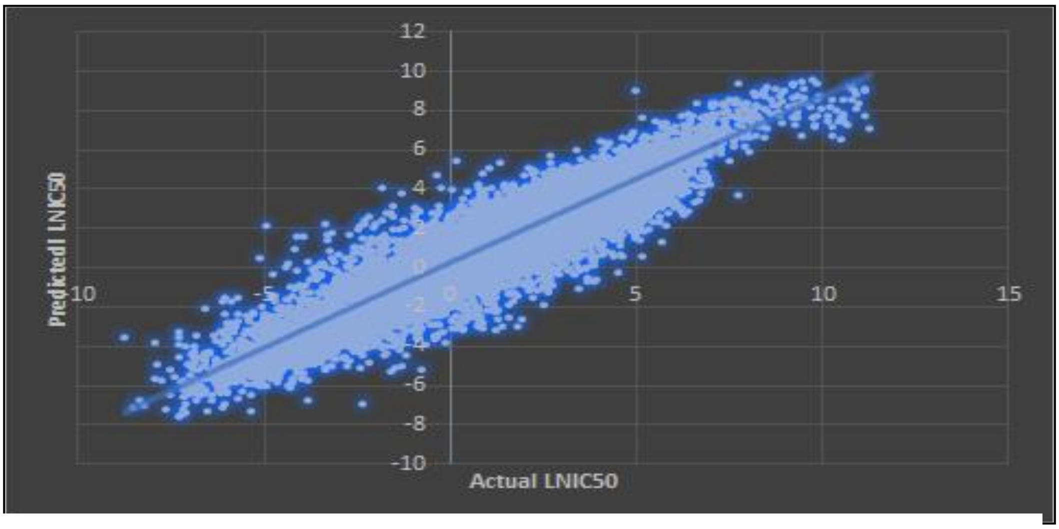


Figure 8: LN IC50 predictions using meta learner model using single omics data

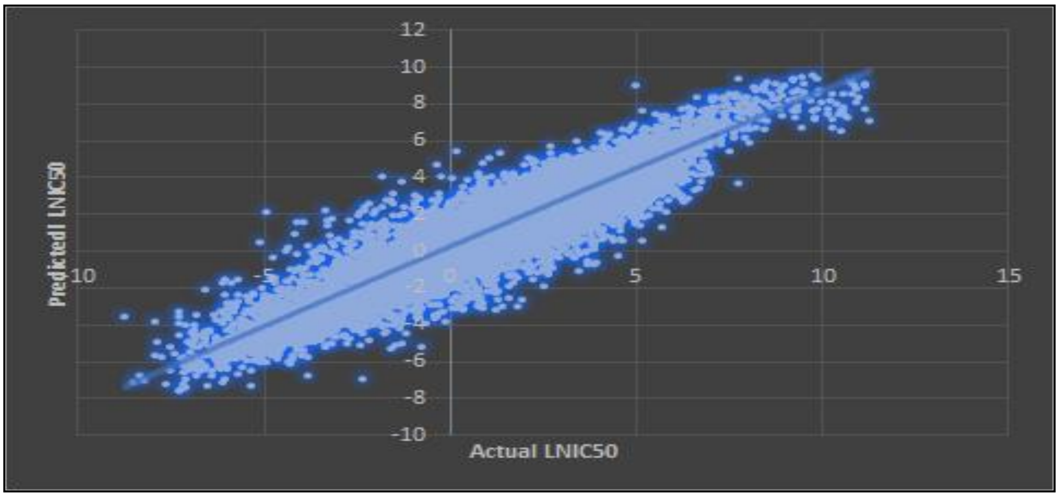


Figure 11: LN IC50 predictions using meta learner model using multi omics data.

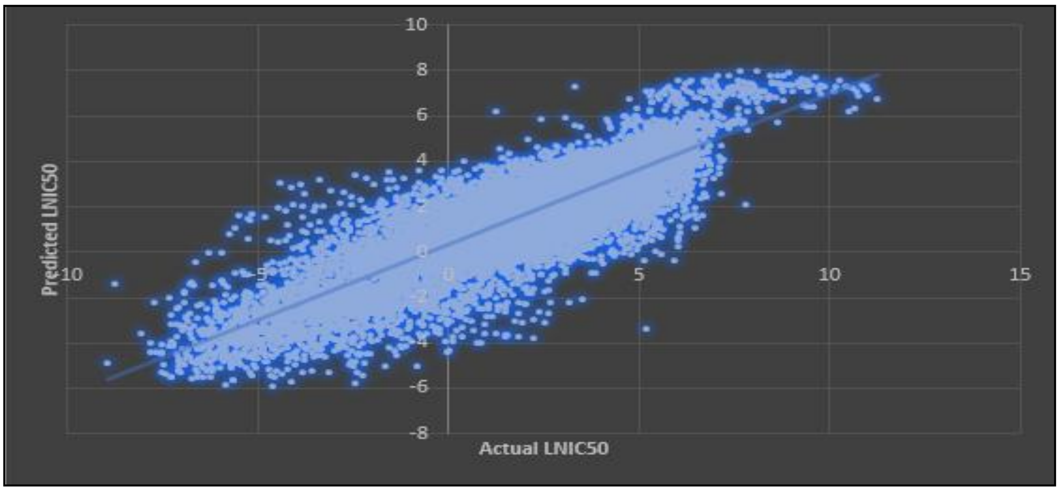


Figure 12: LN IC50 predictions using meta learner model using STR & SNP data.

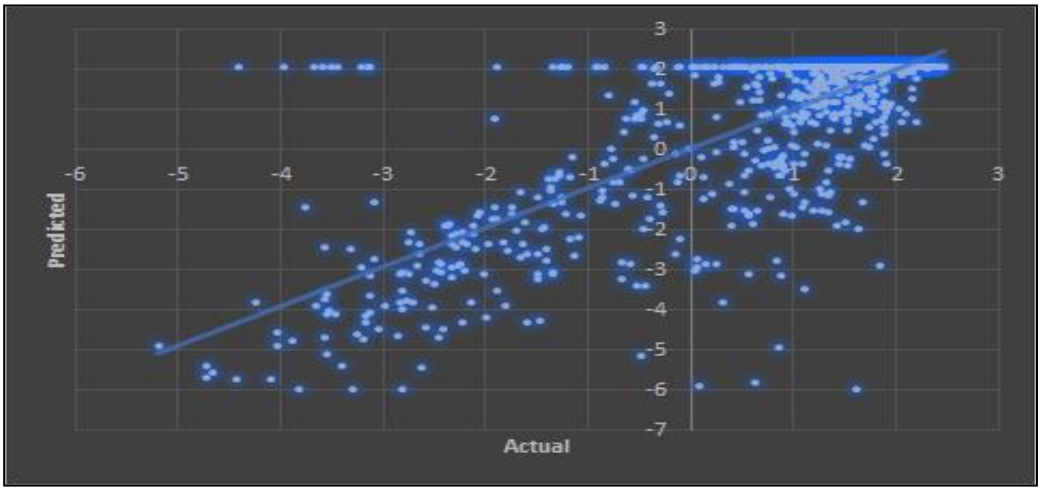


Figure 13: LN IC50 predictions using meta learner model using STR & SNP data.

Table 3: Mean squared error of five fold cross validations using multi omics data for various models

Model	*Five Fold Cross Validation Results (MSE)*					*Mean MSE*
XGBoost	1.186	1.165	1.165	1.172	1.192	**1.176**
LGBoost	1.231	1.203	1.222	1.221	1.233	**1.222**
CATBoost	1.154	1.123	1.137	1.139	1.152	**1.141**
Averaged Model	1.155	1.128	1.139	1.143	1.157	**1.144**
Meta Learner	1.134	1.104	1.115	1.117	1.126	**1.119**

Table 4: R squared of five fold cross validations using multi omics data for various models

Model	*Five Fold Cross Validation Results* ((R^2)					*Mean* R^2
XGBoost	0.841	0.846	0.846	0.846	0.844	**0.844**
LGBoost	0.835	0.840	0.839	0.839	0.839	**0.838**
CATBoost	0.845	0.851	0.850	0.850	0.849	**0.849**
Averaged Model	0.845	0.850	0.850	0.850	0.849	**0.849**
Meta Learner	0.848	0.854	0.853	0.853	0.853	**0.852**

Table 5: Mean squared error of five fold cross validations using STR & SNP data for various models

Model	*Five Fold Cross Validation Results (MSE)*					*Mean MSE*
XGBoost	1.420	1.449	1.418	1.421	1.426	**1.427**
LGBoost	1.478	1.505	1.483	1.487	1.491	**1.489**
CATBoost	1.461	1.499	1.466	1.461	1.469	**1.471**
Averaged Model	1.411	1.441	1.411	1.411	1.419	**1.418**
Meta Learner	1.415	1.447	1.418	1.419	1.425	**1.425**

4.4 RESULTS COMPARISION

We have selected six contemporary models and compared our best results against these six models. The six models are outlined briefly below. Towards the end of the discussion, we have presented a table comparing the results of all the studies. The data used in our project is very similar to the data used to evaluate the performance of the below stated models. SWNet research conducted in 2021 has done the work of comparing its work with other state of art models and published the comparative results. We have used these published comparison results to check where our model stands.

(1) SWNet[22]: The work presented a novel deep-learning model that integrates gene expression, genetic mutation, and chemical structure of compounds in a multi-task convolutional architecture. The datasets used were Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) datasets. The researcher has selected relevant cancer-related genes based on oncology genetics database and L1000 landmark genes and used their expression and mutations as genomic features in model training. They obtained the cheminformatics features for compounds from PubChem or ChEMBL. The paper also stated that combining gene expression, genetic mutation, and cheminformatics features greatly enhances the predictive performance.

(2) Kernelized Bayesian multi-task learning (KBMTL)[23]. KBMTL is a novel Bayesian algorithm that combines kernel-based non-linear dimensionality reduction and binary classification or regression.

(3) Similarity-regularized matrix factorization (SRMF)[24]. The study is focused around predicting the efficacy of the drugs using their chemical structures and the gene expression representation of cell lines. The similarity between drugs and the similarity between cell lines were computed using their chemical structures and gene expressions respectively and was used to regularize.

(4) Weighted Graph Regularized Matrix Factorization (WGRMF)[25]. The primary principle of the research was to generate the latent matrices of drugs and cell lines and use these in the prediction task. The research developed a p-nearest neighbour graph to sparisfy the similarity matrices of drugs and cell lines.

Table 6: R squared of five fold cross validations using STR & SNP data for various models

Model	*Five Fold Cross Validation Results* $((R^2)$					*Mean* R^2
XGBoost	0.812	0.809	0.813	0.814	0.809	**0.812**
LGBoost	0.804	0.802	0.805	0.805	0.801	**0.803**
CATBoost	0.806	0.803	0.807	0.809	0.804	**0.806**
Averaged Model	0.813	0.810	0.814	0.815	0.810	**0.813**
Meta Learner	0.812	0.810	0.813	0.814	0.810	**0.812**

Table 7: Mean squared error of five fold cross validations using CCLE data for various models

Model	*Five Fold Cross Validation Results (MSE)*					*Mean MSE*
XGBoost	1.574	1.535	1.629	1.550	1.540	**1.565**
LGBoost	1.560	1.567	1.597	1.481	1.425	**1.526**
CATBoost	1.251	1.325	1.331	1.237	1.243	**1.277**
Averaged Model	1.404	1.410	1.455	1.359	1.347	**1.395**
Meta Learner	1.217	1.274	1.291	1.221	1.194	**1.239**

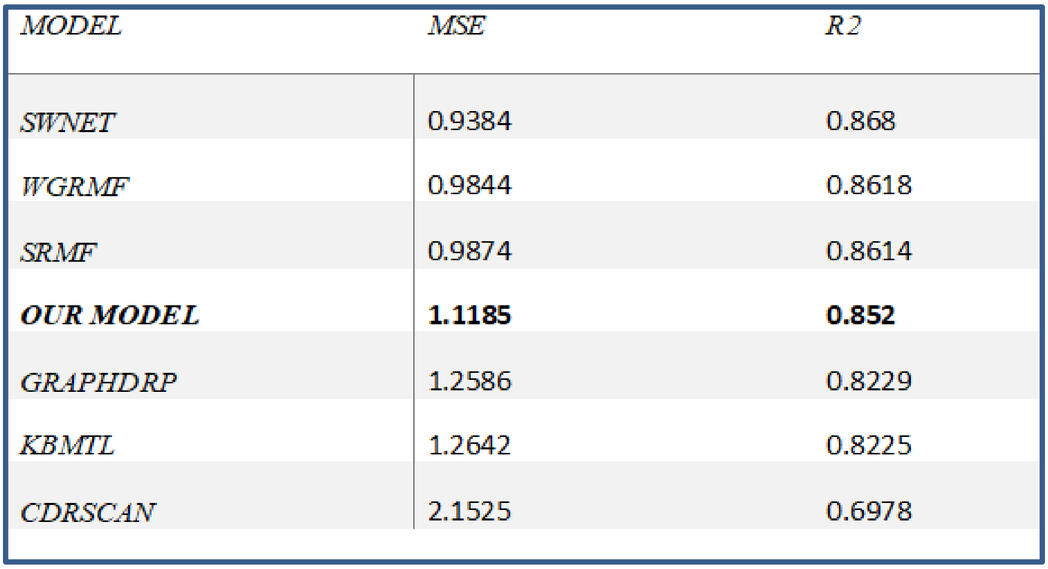
Table8: R squared error of five fold cross validations using CCLE data for various models

Model	*Five Fold Cross Validation Results (R2)*					*Mean R2*
XGBoost	0.608	0.601	0.604	0.587	0.554	**0.591**
LGBoost	0.612	0.593	0.611	0.605	0.587	**0.602**
CATBoost	0.689	0.656	0.676	0.670	0.640	**0.666**
Averaged Model	0.650	0.634	0.646	0.638	0.610	**0.635**
Meta Learner	0.697	0.669	0.686	0.674	0.654	**0.676**

(5) Cancer Drug Response profile scan (CDRscan). the study used a novel deep learning model dubbed Cancer Drug Response profile scan CDRscan) to predict the drug outcomes or response. CDRscan is an ensemble of 5 CNN's with different architecture and functionalities. The data used in the study is, structural profiles of 244 anticancer drugs and genomic profiles of 787 human cancer cell lines [10]. The study processed the genomic features separately and the molecular fingerprints of drugs separately using two step convolutional neural networks and later these two were merged using virtual docking technique.

(6) Graph convolutional network for drug response prediction (GraphDRP)[26]. The study used the molecular graph representation of the drugs to capture the bonds between the atoms and the cell lines were represented as binary vectors of genomic aberrations. Convolutional layers were employed for feature learning tasks of drugs and the cell lines. In the final step, the combination of drug and cell line features were used to predict the drug response.

Table 2: Comparison of the results from above mentioned studies using GDSC data



MODEL	*MSE*	*R2*
SWNET	0.9384	0.868
WGRMF	0.9844	0.8618
SRMF	0.9874	0.8614
OUR MODEL	**1.1185**	**0.852**
GRAPHDRP	1.2586	0.8229
KBMTL	1.2642	0.8225
CDRSCAN	2.1525	0.6978

From the Table 9 we could notice that the results produced by our research are highly significant and comparable to the contemporary research work.

Chapter 5: Findings

From the four experiments we have conducted during the project tenure we could notice some significant findings as listed below.

1) We have implemented some of the classical boosting algorithms such as XG boost, Light gradient boost and Catboost. All the three models produced significant results however, we have attempted to combine these three boosting models by taking the average of the predictions from each model as the final model and stacking the three boosting models using a LASSO meta learner. In all the four experiments stacked model outperformed all the other models.

2) We have attempted to evaluate if STR and SNP representation of the cell lines provide better information or the genetic (mRNA, copy number variation) representation provide better information for predicting the drug response. We could note that the genetic representation of the cell lines are more informative and assist in making better predictions.

3) Furthermore, we could obtain better results by late integrated representation of cell lines by using mRNA and copy number variations (multi omics representation).

4) In our project we have experimented using the single omics (mRNA expressions) and multi omics (mRNA expressions and copy number variations). However, the same architecture can be used for any number of omics datasets.

Chapter 6: Conclusions

The project primarily focused on developing a machine learning approach to predict the anti cancer drug efficacies using the publicly available datasets. To attain the objectives, we have used the GDSC and CCLE datasets and explore various classical boosting algorithms. Finally, we have combined XG Boost, CatBoost and LG Boost using a LASSO meta learner to attain better predictions. The model we have proposed has outperformed all the other models we have explored. The report presents the details of data exploration, preparation, machine learning implementation and results obtained in various sections.

Chapter 7: Future Work

The results which we presented in the report are very competitive and significant, we have also compared the results with various recent studies and noted that our results stand very close to the published results state of art research. However, we believe we can further improve the predictions by making some changes to the model architecture we have presented in the study. One such change is to develop a deep neural network for prediction and combine the predictions from this network with the current models. Currently we have combined three base models that are XG Boost, LG Boost and CatBoost.

From the examination of the results, we could notice that late integration of mRNA and copy number variation produced the best results. We could sense that by increasing the number of omics data, we can get better predictions. In the future we would like to extend the model architecture to combine other omics data such as mutation and methylations which are also very informative.

Bibliography

[1] Cheng, M.L. et al. (2018) Clinical tumour sequencing for precision oncology: time for a universal strategy. Nat. Rev. Cancer, 18, 527–582.

[2] Gao, H. et al. (2015) High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat. Med., 21, 1318–1325

[3] Iorio, F. et al. (2016) A landscape of pharmacogenomic interactions in cancer. Cell, 166, 740–754.

[4] Barretina, J. et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483, 603–607

[5] Sharifi-Noghabi, H, Zolotareva, O, Collins, CC & Ester, M 2019, 'MOLI: multi-omics late integration with deep neural networks for drug response prediction', Bioinformatics, vol. 35, no. 14, pp. i501–i509.

[6] Lee, J.-K. et al. (2018) Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. Nat. Genet., 50, 1399–1411.

[7] Wang, J, Wei, Q, Ye, J, Denduluri, SK, Wang, X, Mohammed, MK, He, T-C 2015, 'Insider information: Testing cancer drug sensitivity for personalized therapy', Genes & Diseases, vol. 2, no. 3, pp. 219–221.

[8] Ringel, MS, Scannell, JW, Baedeker, M & Schulze, U 2020, 'Breaking Eroom's Law', Nature Reviews. Drug Discovery, vol. 19, no. 12, pp. 833–834.

[9]'A decade in drug discovery: Nature Reviews Drug Discovery marks its tenth anniversary this month, providing an opportunity to reflect on the evolution of the landscape of drug research and development' 2012, Nature Reviews. Drug Discovery, vol. 11, no. 1, p. 3–.

[10] Cruz, JA & Wishart, DS 2006, 'Applications of Machine Learning in Cancer Prediction and Prognosis', Cancer Informatics, vol. 2, pp. 59– 77.

[11] Schperberg, AV, Boichard, A, Tsigelny, IF, Richard, SB & Kurzrock, R 2020, 'Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials', International Journal of Cancer, vol. 147, no. 9, pp. 2537–2549.

[12] Kourou, K, Exarchos, TP, Exarchos, KP, Karamouzis, MV & Fotiadis, DI 2015, 'Machine learning applications in cancer prognosis and prediction', Computational and Structural Biotechnology Journal, vol. 13, no. C, pp. 8–17.

[13] BARRETINA, J, CAPONIGRO, G, REDDY, A, LIU, M, MURRAY, L, BERGER, MF, … KOREJWA, A 2012, 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', Nature (London), vol. 483, no. 7391, pp. 603–607.

[14] Menden, MP, Iorio, F, Garnett, M, McDermott, U, Benes, CH, Ballester, PJ & Saez Rodriguez, J 2013, 'Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties', PloS One, vol. 8, no. 4, pp. e61318–e61318.

[15] Chang, Y, Park, H, Yang, H-J, Lee, S, Lee, KY, Kim, TS, … Shin, J-M 2018, 'Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature', Scientific Reports, vol. 8, no. 1, pp. 8857–11.

[16] Schperberg, AV, Boichard, A, Tsigelny, IF, Richard, SB & Kurzrock, R 2020, 'Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials', International Journal of Cancer, vol. 147, no. 9, pp. 2537–2549.

[17] Sharifi-Noghabi, H, Zolotareva, O, Collins, CC & Ester, M 2019, ‘MOLI: multi-omics late integration with deep neural networks for drug response prediction’, Bioinformatics, vol. 35, no. 14, pp. i501–i509.

[18] Öztürk, H, Ozkirimli, E & Özgür, A 2016, ‘A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction’, BMC Bioinformatics, vol. 17, no. 1, pp. 128–128.

[19] Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32: 1466–14

[20] Mirabelli, P, Coppola, L & Salvatore, M 2019, ‘Cancer Cell Lines Are Useful Model Systems for Medical Research’, Cancers, vol. 11, no. 8, p. 1098.

[21] Nkhoma, N, Shimelis, H, Laing, MD, Shayanowako, A & Mathew, I 2020, ‘Assessing the genetic diversity of cowpea [Vigna unguiculata (L.) Walp.

[22] 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, vol. 22, no. 1, pp. 1–434.

[23] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.

[24] Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.

[25] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. Cosmic: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47(D1):941–7.

[26] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2012;41(D1):955–61