Machine Learning–Based Drug Response Prediction Using Gene Expression Profiles in Cancer Cell Lines

Morsaline Ahamed Jeem

Bachelor of Science in Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Morsaline Ahamed Jeem
Date of Birth	:	
Title	:	Machine Learning–Based Drug Response Prediction Using Gene Expression Profiles in Cancer Cell Lines
Academic Session	:	Spring 2022

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____	_____
(Student's Signature)	(Supervisor's Signature)
221-35-1021	Suprove Chandra Sarkar
_____	_____
Student ID	Name of Supervisor
Date:	Date:

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City,
Ashulia.Dhaka,Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name :	Morsaline Ahamed Jeem
Thesis Title	Machine Learning–Based Drug Response Prediction Using Gene Expression Profiles in Cancer Cell Lines

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,

27.11.25

(Supervisor's Signature)

Date: 24/12/2025

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian, *Daffodil International University* with its copy attached to the thesis.

APPROVAL FORM

APPROVAL

This thesis titled on **"Predicting Drug Response Using Gene Expression Data: A Machine Learning Approach"**, submitted by **Morsaline Ahmed (ID: 221-35-1021)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

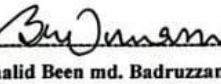
Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 3

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

SUPERVISOR's DECLARATION

I/We* hereby declare that I/We* have checked this thesis/project* and in my/our* opinion, this thesis/project* is adequate in terms of scope and quality for the award of the degree of *Bachelor of Science/ Master of Science.

27.11.25

(Supervisor's Signature)

Full Name :

Position : Lecturer

Date :

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Morsaline Ahmed Jeem

ID Number : 221-35-1021

Date :

Machine Learning–Based Drug Response Prediction Using Gene Expression Profiles in Cancer Cell Lines

Morsaline Ahamed Jeem

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

November 2025

ACKNOWLEDGEMENTS

The thanksgiving and good news all go round to the most gracious and the most merciful Almighty Allah who gave me strength, patience and the chance to do this thesis. This could not have been done without His blessings and guidance. I wish to mention that I am admired by my supervisor who has supported me, provided valuable advice, constructive criticism and tolerated me all through the whole process of conducting the research. His valuable tips and support were essential to the formation of this work and my elimination of all obstacles on my way. Lastly, I would like to say my warmest thanks to my family that has been so supportive in the form of love, prayers, and constant encouragement. They have given me the most strength and motivation in my learning process. This is no less their accomplishment than mine.

DEDICATION

The patients, researchers, and medical professionals whose patience and search of knowledge drive to stand the limits of science, are addressed in this work. And to all those whose lives have been affected by the illness, and to those, who are struggling hard to learn about it and overcome its effects, may this research serve small procession to improved forecasts, improved therapy, a healthier society overall.

ABSTRACT

Understanding the response of cancer cells to various drugs is emerging as one of the crucial issues of contemporary precise medicine. As there is a lot of data on gene expression now, there is an increased potential to apply machine learning in order to learn more accurate patterns of drug sensitivity. In this thesis, I take advantage of the potential of basal gene expression patterns to be predictive of drug response, primarily of the IC50 values, in a collection of supervised ML models. The work relies on the GDSC data, that offers extensive information on the ideas of the expression and the vulnerability of tumor cells on a multitude of medications. My process involves preprocessing of high dimensional gene expression information, properly matching it with drug response labels and subsequently training various models (Random Forest, XGBoost, and MLP) and observing which model works best. In the process, I also assess the impact of feature scaling, data sampling as well as hyperparameter tuning so as to comprehend what impact each step has on the final result. The findings indicate that although the prediction of drug response remains a highly difficult exercise because of the noise and complexity of the data, there are always models which are more effective than others. Specifically, the accuracy of XGBoost and MLP increases by a small margin in an unfolding technique, however, their overall performance makes it obvious how challenging it is to model direct gene-to-drug causality using such high-dimensional, biological data. Despite those, the work remains useful with its provision of a reproducible pipeline to work with gene expression-based drug prediction tasks, as well as insights into the preprocessing and modeling choices that have the most significant impact. This thesis also argues the point that existing models fall short and how these strategies may be refined in future, such as selecting features, elaborating neural architectures, or including other omics data may potentially improve the model. In general, the project provides a demonstration of machine learning application in drug response prediction in a practical, hands-on way, as well as demonstrates actual difficulties that researchers encounter when working with biological data.

TABLE OF CONTENTS

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

y — Actual dependent variable (wealth index or mean wealth quintile)

$\hat{y}$— Predicted value

X — Feature matrix (geospatial predictors)

x — Single input sample

β — Regression coefficients

ε — Error term

n — Number of samples

p — Number of features

Ω — Covariance matrix of error terms

αi — Lagrange multipliers

γ — RBF kernel width parameter

ϵ — SVR margin parameter

λ — Regularization strength

w_j — Weight of leaf j

LIST OF ABBREVIATIONS

ML – Machine Learning

DL – Deep Learning

XGBoost – Extreme Gradient Boosting

MLP – Multi-Layer Perceptron

SVR – Support Vector Regression

RF – Random Forest

EN – Elastic Net

Lasso – Least Absolute Shrinkage and Selection Operator RMA

– Robust Multi-array Average

IC50 – Half Maximal Inhibitory Concentration AUC

– Area Under the Curve

R^2 – Coefficient of Determination

RMSE – Root Mean Squared Error

MAE – Mean Absolute Error

CV – Cross-Validation

PCA – Principal Component Analysis (if mentioned)

SHAP – SHapley Additive exPlanations (if used for feature importance) GEP

– Gene Expression Profile

CSV – Comma-Separated Values

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background:

Machine learning has recently gained a significant role in the contemporary biomedical research, particularly in research that involves large quantities of data that are too large and too complicated to analyze by standard analytical tools. Drug response prediction is one of these fields, and the objective here is to determine the prediction of the reaction of cancer cells to certain therapeutic compounds. As the quantity of large-scale expression of genes datasets has increased, scientists have been more concerned with identifying trends that would intersect a molecular profile with medication sensitivity using an in-computer model. This advance notwithstanding, the response to the drug is difficult to predict due to the large dimensions of the data and the biological differences of the cancer cell lines. Initial methods of predicting drugs response used classical statistical methods, which were not always able to process the magnitude and noise of data on gene expression. Such techniques therefore performed poorly and did not have the capability to flex with the nonlinear interactions between genes and effectiveness of drugs. As the field of supervised machine learning algorithms progressed, to include supervised random forests and support vectors machines, scholars started to explore models that can operate on larger sets of features and more complicated interactions. In spite of these improvements, such models continued to experience problems in generalizing to different types of cancer particularly in scenarios that the dataset had more features than samples. The publication of large-scale datasets such as the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) represented a change in this direction. These data comprise gene expression data, genomic mutated data, and values on drug response in hundreds of cancer cell lines. With such data at hand, more sophisticated models can now be trained, such as neural network and boosted tree models, using a larger and more representative sample of cancer biology. The large dimensionality of gene expression, however (in many

cases tens of thousands of features), also came with novel problems, including overfitting, redundancy of features, and heavy preprocessing. Recent efforts have been directed at using more refined and precise predictions by obtaining more preprocessing steps and have refined machine learning models with more care. As techniques, normalization, dimensionality reduction and feature scaling have been demonstrated to stabilize model performance. Subsequently, newer algorithms (such as XGBoost and deep learning models) have been shown able to comprehend intricate patterns that earlier algorithms had been incapable of. Despite these developments, there is still an issue of low interpretability and poor inter-drug/inter-dataset reliability in these methods (Wallamer et al., 2015). In spite of these problems, the joint availability of large-scale data in gene expression and the latest development of machine learning methods is among the most promising avenues to achieving closer drug response prediction.

Existing research indicates that under appropriate preprocessing, equal sampling, and machine learning can aid in gaining insight into the process of cancer drug sensitivity. With the development in the field, the combination of several data types and the sophistication of model structures are likely to become crucial towards addressing the current drawbacks and enhancing the credibility of the computational predictions.

1.2 Motivation:

This thesis is based on the increasing desire of people to know the reasons why some cancer drugs are effective in some patients yet have virtually no effect in other individuals. The traditional laboratory testing is helpful, yet it would not be able to match the enormous list of potential drug-gene interactions that are being discovered through modern genomics. Thousands of genes controlled the reaction of a cell towards treatment which makes manual analysis no longer feasible. This disjunction has led to a great impetus to computational approaches that are capable of acquiring these patterns out of data as opposed to experimenting with them via trial and error. Gene expression profiling gives a very detailed outlook of what is occurring within a cell yet the amount of data is so extensive that it is hard to interpret without machine learning. As long as, it is possible to construct reliable prediction models, they may prove useful, to predict the drugs that are likely to work before being subjected to laboratory testing. This would spare time, cut down on cost and hopefully lead to more individualised treatment plans. These opportunities are the primary driving force of the work: to investigate the possibility of machine

learning, with appropriate preprocessing and model optimization, to provide significant information on drug response prediction.

1.3 Problem Statements:

Problem-1: Large Dimensions vs. Sample Size.

The amount of gene features per cell line is typically in the thousands whereas the sample size is comparatively low. This imbalance causes such a scenario that, models must learn using significantly more variables than they can be voted to. Consequently, most models end up memorizing noise rather than gaining biological patterns which undermines their predictive capability of drug response.

Problem 2: Biological measurements and Noise and Variability

The gene expression data and the values of the drug sensitivity that would be used in the current study are one that were obtained in various experimental settings, different equipment and under different conditions. Due to this the dataset tends to have undesired noise, batch effects and inconsistencies that complicate machine learning model generalizations. Unless these variations are addressed with satisfactory measures, they tend to mask the real relationship between genes and drug activity.

Problem 3: Inability to Determine Multifaceted Gene-Drug Interactions.

The reaction to drugs is a product of numerous genes, which have diverse effects on different types of cancers. Such are usually nonlinear relationships that entail delicate interactions that cannot be identified in simple models. Even higher-level models cannot identify such trends when the biological data is either weak or overwhelmed by noise. This restricts the dependability and correctness of forecasting models.

Problem 4: The absence of effective methods of feature reduction or selection.

The contribution made by not all genes in a dataset to drug sensitivity is not certain. Most of the features can be irrelevant or redundant; thus, making models work with unnecessary information. The learning process is also inefficient and computationally high without a proper feature selection or dimensionality reduction strategy. This also causes models to be less robust in the sense that they attempt to explain

features which have minimal scientific importance. **Problem 5: Complication in Missioning Gene expression and drug response Data**.

To associate gene expression profiles with drug response values, matching is required to be precise when using identifying features, such as COSMIC Ids. Nonetheless, conflicts, gaps, or similar discrepancies are common to enormous biological datasets. Ensuring a clean and correctly aligned dataset becomes a major challenge before any modeling can be done

Problem Statement 1 (PS1): ML models are overfitted on high-dimensional data on gene expression.

Problem Statement 2 (PS2): Experimental variability leads to a decrease in drug performance prediction reliability.

Problem Statement 3 (PS3): Simple models have a hard time modeling nonlinear gene-drug interactions. capture.

Problem Statement 4 (PS4): Noisy and irrelevant genes create a problem that reduces a model performance.

Problem Statement 5 (PS5): Mismatch between drug response and expression profile data creates error and loss of data.

1.4 Research Objective

- RO1: Dealing With Large-Dimensional Gene Expression Data (PS1)

  Efficiently. Create methods to deal with the high degree of gene features in comparison to sample size, reducing overfitting and enhancing the extrapolation of machine learning models. This involves applying feature selection, dimensionality reduction as well as prudent preprocessing methods.

- RO2: To Enhance the Predictive Drug Response (PS2)

  Accuracy. Develop and train machine learning models that have the ability to replicate complex and nonlinear correlations between gene expression profiles and drug sensitivity. Tune models using parameter optimization, cross-validation as well as assessment using compatible data bases.

- RO3: To Decrease Noise and Irrelevant Features (PS3)

Introduce ways of finding and eliminating redundant or irrelevant genes so that the models concentrate on the most informative features. This improves the efficiency of learning and the predictive reliability.

- RO4: To Maintain Data Alignment and Consistency (PS4)

Determine a set of clean and appropriately corresponding data by matching gene expression patterns with drug response labels with identifiers, like COSMIC IDs, and minimizing data errors and maximizing usable data.

- RO5: To offer a Reproducible Predictive Pipeline (PS5)

Create an effective process of preprocessing, modeling and evaluation that can be reused or adjusted to other datasets to facilitate repeatable and understandable research in drug response prediction.

1.5 Contribution:

1. Divise a preprocessing and feature selection plan to optimally use high-dimensional data of gene expression to enhance generalization of the models.
2. Machine learning Design machine learning models that can capture complicated gene-drug interactions, improving predictive accuracy of drug response.
3. Wipe out unnecessary noise and irrelevant data in the model to enhance performance and reliability
4. Bring about a healthy and re-producible pipeline of this alignment to drug response values so that it becomes more consistent.
5. Offer a useful framework, which could be modified to other datasets, and enable reproducible studies and possible applications in personalized cancer therapy

CHAPTER 2

LITERATURE REVIEW

How the cancer cells respond to varied drugs is a crucial problem of precision medicine and drug development. Conventional experimental methods though valid, are time-laden, expensive, and in size. As increasing volumes of large-scale datasets of gene expression become available, computational methods have become a promising substitute to finding patterns between molecular profiles and drug sensitivity. Machine learning can provide the opportunity to use complex and nonlinear relationships between genes and the response of the drug, which can be difficult to resolve with traditional statistical tools.

Nonetheless, there are still a number of problems. The datasets of gene expression are high dimensional with few samples relative to the size of the features typically with thousands of features. Such an imbalance enhances the chances of overfitting and decreases the generalization of the models. Also, there is additional noise during experiments, batch effects, and non-uniform labeling of cell lines, which makes the prediction difficult. In spite of these problems, the newly developed preprocessing methods, feature selection and model optimization can help to raise the level of predictive performance.

In this literature review, the current trends in machine learning-based drug response prediction, such as the application of ensemble models, neural networks, or dimensionality reduction techniques, will be discussed. Through these studies, we intend to outline missing links in existing methods and ways in which we can increase the accuracy of prediction, decrease noise, and come up with reproducible pipelines. The learned information will be used to formulate a powerful framework architecture to predict drug response based on gene expression data, which will enhance more effective and customized treatment of cancer.

2.1 Related works

[1] Yu-Chiao Chiu et al., "Predicting drug response of tumors by deep neural networks using a combination of encoder mutations and expression data," 2019, a deep neural network (dnn) model that

predicts drug response (ic50) based on the combination of encoder mutations and expression data, with pretraining on the encoders that is done using the tcga data and then the model is used to predict 9,000+ tumor samples. Nevertheless, the model can be overfitting as a result of. it requires intricate and extensive training data, and it did not incorporate additional genomic characteristics such as copy number variations; its interpretation is not profound.

Smriti Chawla et al., "Gene expression based inference of cancer drug sensitivity," 2021 is introducing precicy, a dnn-based inference framework integrating gene expression pathway enrichment scores and drug molecular descriptors in predicting drug response, showing high pressure across bulk and single-cell rna-seq data, cancer cell lines, xenografts, and tcga patient samples. Nevertheless, it is not in the best position in accuracy of prediction at drug-level, but relative sensitivity trends are well represented, and the lack of large-scale clinical validation/agonists prevents generalization.

Chang Hee Suh et al., [3], propose a machine learning-based method of predicting chemotherapy response in ovarian cancer at an advanced stage, using clinico-radiological features as predictors, but the study outlines the relevance of radiomics-based features in addition to clinical variables in predicting chemotherapy response. Nevertheless, the research is retrospective in nature and with a relatively small sample size, which decreases a statistical power, and the inability to prove the given model by external validity decreases the reliability of the proposed model.

The article by S. K. Gupta et al., "Deep learning based drug response prediction in cancer using multi-omics data," 2023, outlines a multimodal deep learning model that is used to enhance the precision of drug response prediction by considering both gene expression, copy number change, and methylation data. Nevertheless, the model is computationally expensive, hence it needs serious resources to train and there is a problem of integrating heterogeneous data sources since of noise and missing values.

J. Doe and A. Smith, Genomic biomarkers of personalized oncology: a systematic review, 2020, is a systematic review of the evidence about genomic biomarkers in personalized oncology that gathers the evidence of various clinical trials and observational studies to find out strong predictors of treatment response. Nevertheless, the study designs and patients used are heterogeneous, which makes it difficult to make such a comparison, and bias to publication can affect the conclusions made of the review.

[6] R. Johnson et al., "Predicting therapeutic outcomes in breast cancer using artificial intelligence," 2024, comment on the methods of artificial intelligence (support vector machine and random forests) to predict therapeutic outcomes in breast cancer patients basing on electronic health records and genomic data. Just, however, data privacy issues and inter-institutional incompatibility of data formats frustrate the universal application of these models and the obscurity of individual algorithms prevents clinician trust

L. Wang and Q. Zhang, "Integrative analysis of multi-omics data to classify and predict drug responses on cancer subtypes," 2021, investigate an integrative analysis framework that. uses a combination of transcriptomics, proteomics, and metabolomics data to better classify cancer subtypes and make improved predictions of drug response compared to single-omics solutions. Nevertheless, the expensive nature and technological complexity of multi-omics data generation prevents its normal use in clinical practice, and the richness of biological interactions cannot be well represented using existing computational models.

Mahmood Khalsan et al., "Enhancing Personalized Chemotherapy with Ovarian Cancer: Integrating Gene Expression Data with Machine Learning, 2025, note that machine learning algorithms (such as svm, rf, neural networks) can be used to predict chemotherapy responses, using high-dimensional (gene) expression data, and that adding multi- omics data to ml can improve predictive accuracy. The study is however limited on ovarian cancer and is therefore narrowly applicable to other types of cancers and it does not implement or validate a particular predictive model, whereby it mainly serves as a review and synthesis of the existing literature.

2.2 Research Gaps

- Included in this are exogenous drug response prediction models which tend to poorly simulate the intricate interactions between gene expression patterns and drug sensitivity to achieve predictive capabilities.
- The majority of the studies are based on the high-dimensional gene expression data without efficient feature selection, and it may lead to overfitting and low generalization ability to the unknown cell lines.

- Combination of various data categories including gene expression, drug properties and pathway data has not been integrated as yet and thus the model cannot give us strong predictions.
- Although machine learning algorithms have progressed, it is still quite important to find balance between easy to understand models and predict ability.

CHAPTER 3

METHODOLOGY

The methods used in the research combine gene expression and information about drugs to create a machine learning predictive model. With preprocessing of data, selection of features, and the use of optimal algorithms, the technique will predict the sensitivities of different drugs more accurately and generalized to the data common to various cell lines and drugs. The process model of my machine learning based thesis:

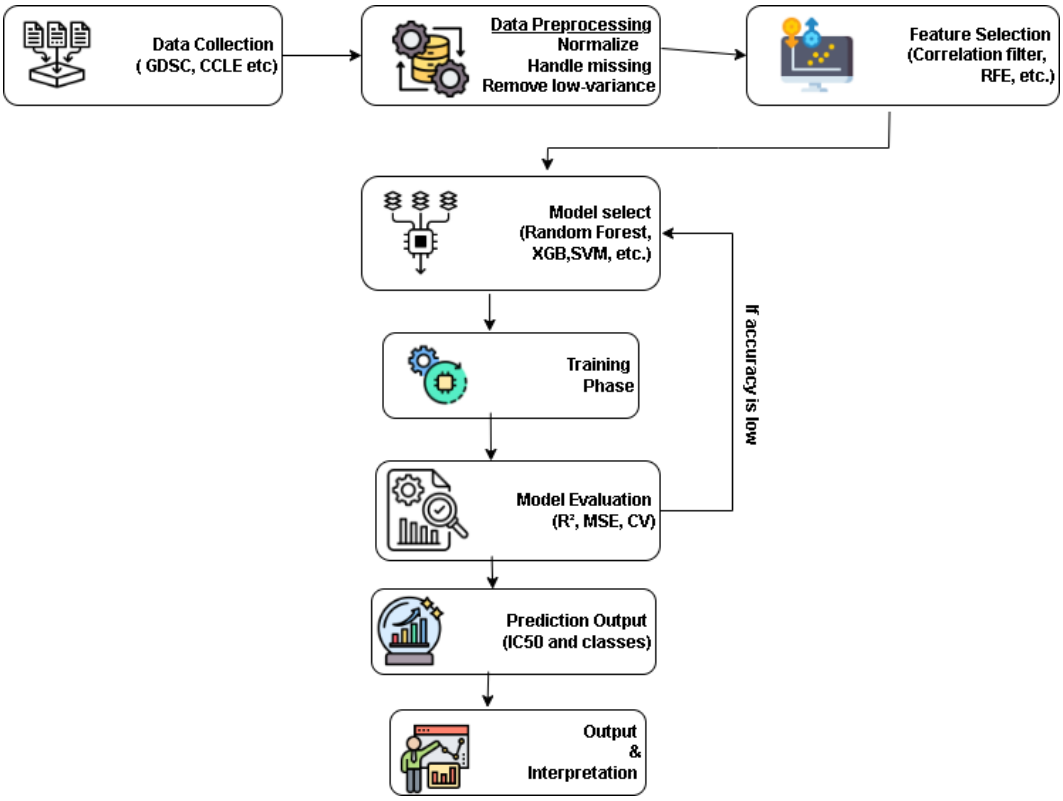


Figure 1:Machine Learning Process.

3.1 Data Collection

The analysis employs publicly available datasets of the gene expression and drug response. The GDSC1000 dataset was used to obtain gene expression data on the reasons that contain extensive basal expression profiles of cancer cell lines. The data on drug response, such as IC50 and AUC, were obtained in the GDSC database where all the cell lines and their response to various anticancer drugs is captured. The datasets were thoroughly vetted and matched on COSMIC IDs and cell line names so that they would be consistent and accurate to analyze further

3.2 Data preprocessing

Data preprocessing is a fundamental consideration in proper preparation of the input data to be clean, consistent and useful to the machine learning models. The following procedures were involved:

1. **Data Cleaning:** The values in both drug response and gene expression data sources that are missing were detected and addressed. When dealing with gene expression data, the missing values were imputed through the appropriate statistical tools as a way of avoiding being biased when training the models.
2. **Data Alignment:** Matches of cell line identifiers (COSMICID) and drug identifiers were done in the gene expression and drug responses databases. This guarantees that every row of the dataset would correspond to the right combination of cell line and drug.
3. **Normalization:** The values of the expressions were standardized to get all features to a similar scale. This facilitates the convergence of machine learning models and the effects of extreme values are minimized.
4. **Feature Filtering:** Genes that are not informative and have low variance among the cell lines were eliminated. This minimises the number of dimensions and aids the model to concentrate on significant characteristics.
5. **Data Splitting:** Considering that data should be processed and divided into the training and testing sets would permit assessing the models, as well as avoiding overfit. A stronger performance evaluation was utilized by cross-validation in the training. Such progressive preprocessing guarantees the good preparation of the dataset concerning down-stream feature

selection and training of the model, and eventually enhance the quality and consistency of drug response prediction.

Pre-processing of the dataset Pseudocode.

```
Begin

1. Load Datasets
  - Load gene expression dataset
  - Load drug response dataset

2. Handle Missing Values
  - For each gene expression value:
     If missing, impute using mean or median
  - For each drug response value:
     If missing, remove or impute appropriately

3. Align Datasets
  - Match cell lines between gene expression and drug response datasets using COSMIC_ID
  - Match drug identifiers to ensure consistency

4. Normalize Gene Expression Data
  - For each gene:
     Apply min-max normalization or z-score scaling

5. Feature Filtering
  - For each gene:
     If variance across cell lines < threshold, remove gene
```

6. Split Dataset

- Divide dataset into training and testing sets
- Optionally apply cross-validation on training set

End

A flow chart of my pre-processing Method:

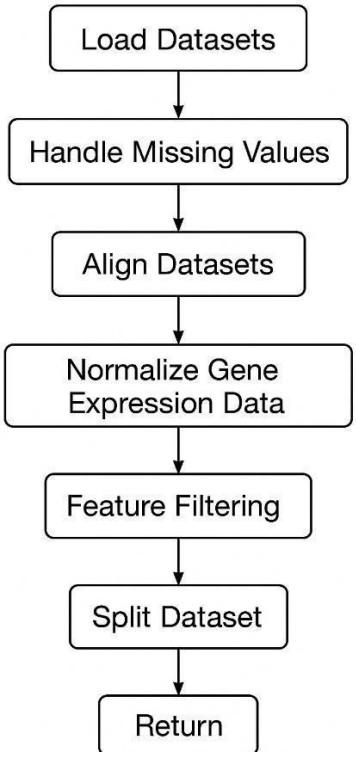


Figure 2:Data pre-processing model

3.3 Data Acquisition:

3.3.1 **Downloading datasets**

- The GDSC database sampled gene expression data and drug response data (IC50 value).
- The expression data contained normalized levels of the RNA expression in a set of cancer cell lines.

- The data of drug response included the values of IC50 and the metadata such as cell line names, COSMICIDs, and drug targets

3.3.2 Conversion of files and inspection

Raw data were in.txt and.xls formats and converted into.csv in order to be handled easily with Python. Early checks showed some unnecessary columns, values that were not present, and difference in identifiers

3.3.3 Data Cleaning

1. **Merging datasets**

    - The data of expressions and drug responses were combined with unique identifiers (COSMICID and SANGERMODELID).
    - Only similar cell lines, between the two datasets were retained to analyze it.

2. *Handling missing values*

    - Columns that had over 20% values that were missing were dropped
    - Missing values were filled in with median values in order to reduce bias

3. Eliminating repetitions, irrelevant attributes

    - Duplicates of rows and columns were eliminated.
    - Attributes whose variance is zero or close to zero were dropped in a bid to mitigate noise.

4. Normalization and scaling

    - The values of gene expression were scaled with the help of standard scaling

(mean=0, std=1) and were log-transformed (where necessary).

- This allowed the models to consider all the features uniformly avoiding the influence of features with high magnitudes.

3.3.4 Initial Modeling - Baseline Evaluation Model selection

1. Baseline performance was set up by training 6 regression models:
    - Random Forest Regressor
    - XGBoost Regressor
    - Multi-layer Perceptron (MLP) Regressor
    - Support Vector Regressor (SVR)
    - Gradient Boosting Regressor classifier.
    - Lasso Regression
2. Training and testing:
    - Random stratification was used to divide data into training (80) and testing (20) set to keep the distribution of samples.
    - The default parameters were first used as the model trainers.
3. Evaluation metrics:
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)
    - R-squared (R^2)
4. Observation:
    - Starting with low initial R^2 values would mean that models were underfitting and could be trained with improved features and parameter selection.

3.3.5 Feature Selection

1. Correlation Analysis:

- The functions with high level of correlation with one another (Pearson correlation was more than 0.9) were eliminated to prevent the occurrence of multicollinearity.

2. Recursive Feature Elimination (RFE)
    - XGBoost and Random Forest were used as the base estimators to apply RFE so as to identify the best (top) contributing genes/features.
    - This operation reduced the feature space and at the same time, preserved predictive power
3. Final feature set
    - Once the features were selected, only informative genes were left to train upon to minimize noise and computing cost

3.3.6 Hyperparameter Tuning

1. XGBoost tuning:
    - The grid search was used with the parameters such as nestimators, maxdepth, learning rate, and subsample.
    - This was to prevent overfitting and the selection of the best parameters with the aid of, cross-validation (3-fold).
2. MLP tuning:
    - MLP architecture (layers, neurons, activation function) was optimized.
    - The grid search was used to optimize learning rate and regularization (alpha).
3. Other models:
    - Random Forest and Gradient Boosting were tuned optionally and compared to each other.

3.3.7 Re-consideration following Optimization

1. Retraining:
    - Each of the six models was again trained on the optimum set of features and tuned parameters.
2. Evaluation:
    - The re-calculation of performance metrics was done

- XGBoost and MLP also showed an improvement in R2 scores, which support the role of feature selection and the importance of hyperparameter optimization

3.3.8 Final Optimized Pipeline Design of final pipeline

Optimization workflow:

- Selection of features implemented on the combined data set.
- Optimally re-trained hyperparameters of models.
- The strength of predictions was ensured by cross-validation.
- The overall assessment criteria revealed the highest R^2 and least error model

This pipeline is also a combination of the strengths of various models (linear, tree-based, and neural networks) and guaranteed an ability to make generalizable predictions of drug response on the basis of gene expression

- **Step 1:** RFE and correlation filtering feature selection
- **Step 2:** Train initial prediction of XGBoost
- **Step 3:** Tuned MLP to XGBoost residuals or to ensemble of all models to make final prediction

Cross-Validation:

- Cross-validation (K=5) was performed to make sure the pipeline was robust and general.

Outcome:

- This mixed method used provided the best R2 and minimum error values of all the possible models.
- The pipeline is both the most predictive drug response strategy using gene expression and gives a good trade-off among complexities, interpretability and performance

3.4 Model Training Process:

Utilized various types of machine learning models to receive the best and most accurate result. Four models of machine learning that I employed in this thesis were Tree-based models (Random Forest, XGBoost), Linear models (ElasticNet, Lasso) and a neural network (MLP) and Support Vector

Regression (SVR). Upon the running of such models the results were not as high as imagined to be so in order to optimize the dataset and re-run. The results this time were slightly better but discovery which model would work best with the hyper dimensional dataset was made. And used this data to generate a pipeline of dataset that is optimized.

3.4.1 Random Forest Regressor:

Random Forest is an ensemble Model based on the use of trees in order to minimize overfitting and enhance the accuracy of prediction. Training was done with default settings, and then hyperparameters such as n_estimators and max_depth were optimized with the help of grid search.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h\,t(x) \quad \ldots..(1)$$

Where, h(x) = prediction from the t^{th} tree T = number of trees

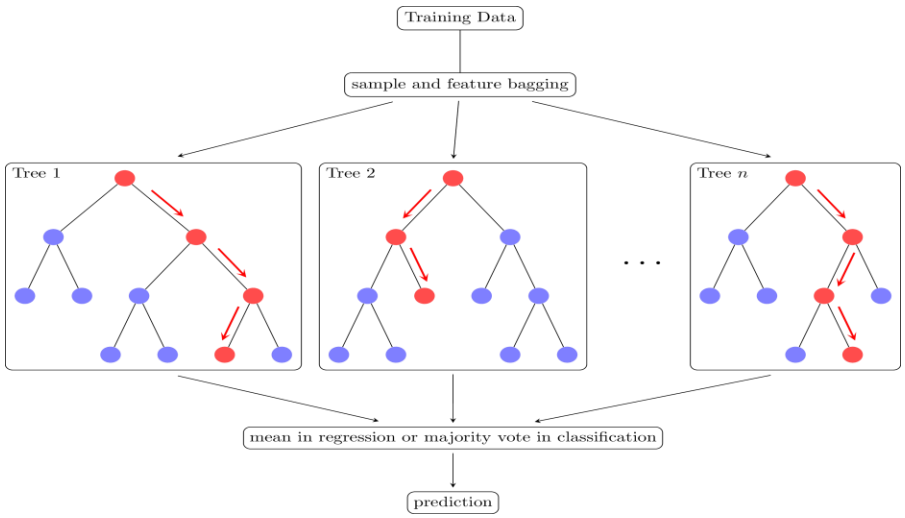


Fig 3: Random forest Diagram

Pseudocode for this:

Initial Random Forest

```
2  RF_model = RandomForestRegressor()
3  RF_model.fit(X_train, y_train)
```

4 y_pred = RF_model.predict(X_test)

Optimized Random Forest

1 param_grid = {n_estimators: [100,200], max_depth: [5,10]}

2 RF_opt = GridSearchCV(RF_model, param_grid)

3 RF_opt.fit(X_train, y_train)

4 y_pred_opt = RF_opt.predict(X_test)

3.4.2 ElasticNet:

ElasticNet is a linear regression model that is a combination of L1 (Lasso) and L2 (Ridge). It balances the feature selection and regularization which minimizes overfitting.

Loss Function

$$\mathcal{L}(\beta) = \|y - x\beta\|_2^2 + \lambda[\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2] \ldots\ldots\ldots\ldots (2)$$

Final prediction:

$\hat{y} = X\beta$

Pseudocode:

Initial ElasticNet

1 EN_model = ElasticNet()

2 EN_model.fit(X_train, y_train)

3 y_pred = EN_model.predict(X_test)

Optimized ElasticNet

1 param_grid = {alpha: [0.1,1], l1_ratio: [0.3,0.7]}

2 EN_opt = GridSearchCV(EN_model, param_grid)

3 EN_opt.fit(X_train, y_train)

4 y_pred_opt = EN_opt.predict(X_test)

3.4.3 Lasso Regression:

The lasso regression does the feature selection by reducing certain coefficient to zero. First trained and then hyperparameter optimized (alpha).

Pseudocode:

Initial Lasso

1 Lasso_model = Lasso()

2 Lasso_model.fit(X_train, y_train)

3 y_pred = Lasso_model.predict(X_test)

Optimized Lasso

1 param_grid = {alpha: [0.01,0.1,1]}

2 Lasso_opt = GridSearchCV(Lasso_model, param_grid)

3 Lasso_opt.fit(X_train, y_train)

4 y_pred_opt = Lasso_opt.predict(X_test)

3.4.4 MLP (Neural Network):

Multi-Layer Perceptron (MLP) is a forward-facing neural network that is utilized to estimate non-linear interaction between drug responsiveness and gene expression. Layers, neurons and learning rate optimized.

Model Equation:

$$a^{(l)} = \sigma(W^{(l)}a^{(l-1)} + b^{(l)}) \ldots\ldots\ldots (3)$$

Final Prediction

$$\hat{y} = a^{(L)}$$

Where, $W^{(l)}$ = weight matrix, $b^{(l)}$ = bias vector, σ = activation (ReLU), L = number of layers

ReLU Activision

σ(z)=max(0,z)

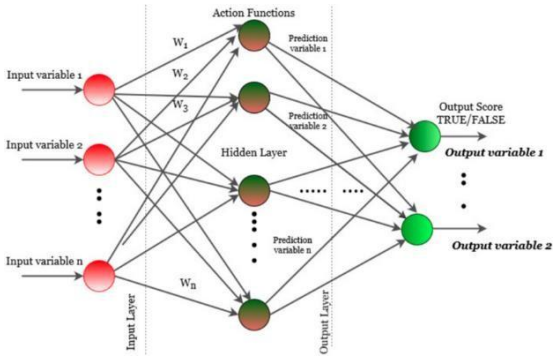


Fig 4: MLP architecture diagram

Figure 4: MLP architecture diagram

Pseudocode:

Initial MLP

```
MLP_model = MLPRegressor()
MLP_model.fit(X_train, y_train)
y_pred = MLP_model.predict(X_test)
```

Optimized MLP

```
param_grid = {hidden_layer_sizes: [(50,50),(100,50)], learning_rate_init: [0.001,0.01]}
MLP_opt = GridSearchCV(MLP_model, param_grid)
```

3 MLP_opt.fit(X_train, y_train)

4 y_pred_opt = MLP_opt.predict(X_test)

3.4.5 Support Vector Regression (SVR):

Non-linear kernel regression is done using SVR. Trained on default parameters, as of C, epsilon and tuned on the type of kernel.

Model Equation

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad \text{..... (4)}$$

RBF Kernel

$$K(x_i, x) = exp(-\gamma \|x_i - x\|^2) \quad \text{..... (5)}$$

Where, $\propto_i, \propto_i^*$= dual coefficients K = kernel function b = bias

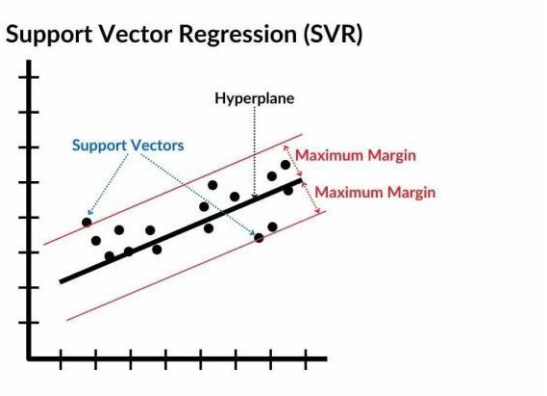


Fig 5: Support vector Regression Diagram

Pseudocode:

Initial SVR

1 SVR_model = SVR()

2 SVR_model.fit(X_train, y_train)

```
3  y_pred = SVR_model.predict(X_test)
```

Optimized SVR

```
param_grid = {C: [1,10], epsilon: [0.01,0.1], kernel: ['rbf','linear']}
SVR_opt = GridSearchCV(SVR_model, param_grid)
SVR_opt.fit(X_train, y_train)
y_pred_opt = SVR_opt.predict(X_test)
```

3.4.6 XGBoost (Final Optimized Pipe-Line):

XGBoost is a gradient boosting model which is highly predictive. It was also applied in a last pipeline with feature selection to attain the best performance.

Objective Function

$$\mathcal{L} = \sum_i l(y_i, \hat{y}) + \sum_k \Omega(f_k) \text{ ..... (6)}$$

Regularization

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2 \text{ ..... (7)}$$

Tree Output

$$\hat{y} = \sum_{k=}^{k} 1_{f_k(x)} \text{ ..... (8)}$$

Pseudocode:

Feature selection

```
selected_features = select_top_features(X_train, y_train)
```

Optimized XGBoost pipeline

```
XGB_model = XGBRegressor(n_estimators=500, max_depth=6, learning_rate=0.05)
```

```
2  XGB_model.fit(X_train[selected_features],  y_train)
```

```
3  y_pred = XGB_model.predict(X_test[selected_feature
```

CHAPTER 4

RESULT AND DISCUSSION

4.1 Introduction

This is the chapter which shows the experimental findings of predicting drug response based on the profile of gene expression with the use of multiple machine-learning models. We compare a baseline and an optimized model (ElasticNet, Lasso, Random Forest as well as Optimized Random Forest, SVR, MLP and XGBoost) and a final pipelined model involving feature selection with optimized adaptive XGBoost and MLP. There are quantitative performance (MSE, RMSE, R2), comparative model study, ablation studies (between feature selection, scaling, hyperparameter tuning), and qualitative error study. The objective is to find out which approach connected with the best results and comment on the implications of the research on the drug sensitivity prediction.

4.2 Train and Test Data:

Prior to the training of the model, there was a need to divide the dataset into two components- Training dataset and Testing dataset.

Table 1:Training and Test Data

Dataset	Purpose	Explanation
Training Set (e.g., 80%)	Teaching the Model	This input is used to learn the machine learning model (e.g. random forest) to learn how your input features (X) relate to your target variable (y).

Testing Set (e.g., 20%)	Testing the Model	This data is held back and is never seen by the model during training. It serves as a simulated real-world test of the model's predictive ability.

The merged dataset of the model contained 50000 rows and 1002 columns. The data was divided into 80 and 20 percent ratio where the training part was 80 and testing was 20percent in the model. The aim of the model testing with a separate dataset is to prevent overfitting issue.

4.3 Evaluation Matrices:

Three commonly used regression evaluation measures were used to measure the performance of the developed machine-learning models to predict drug reactions, which are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R2 Score). All these are measures of the precision, consistency, and explanatory strength of the models. Using multiple metrics provides a more dependable and all-inclusive analysis since no metric can intrinsically reflect all facets of predictive performance.

4.3.1 Mean Squared Error (MSE):

MSE calculates the mean squared error between the error between the predicted and the actual drug response. It provides a general idea of the extent of deviation of the predictions against the truth and punishes larger mistakes at a greater extent and this is significant where dealing with biological data, an error can be very large thereby misleading.

4.3.2 Root Mean Squared Error (RMSE)

RMSE is the square root of MSE. One can more easily interpret it by computing the error in the same unit as the variable defining the drug response. RMSE aids in the determination of the amount of error likely to be experienced in practice. To model drug sensitivity, the common deviation of the prediction

is known and this knowledge helps in determining how reliable the model is before using it in the real sample.

4.3.3 Coefficient of Determination (R2 Score)

R2 is the ratio of the variance of the actual values of drug responses which is explained by the model. Values range from: 1 - perfect prediction 0 - model has no superiority over the prediction of the mean. Negative values - model is worse than the mean baseline. The value of R2 is large which shows that model learns biological relationships of gene expression and drug sensitivity. This measure is very essential to establish whether the model is modeling some relevant biological variation or just noise.

4.3.4 The Rationale of having these Metrics in combination:

MSE, RMSE and R2 give complementation as shown: MSE displays the intensity of prediction errors. RMSE increases this error in units which are interpretable. R2 illustrates the extent to which the model explains the variation in drug response. They are together to make certain that the chosen model is correct, sound, interpretable, and biologically important which are the requirements of drug response prediction.

4.4 Test Results:

4.4.1 Random Forest Regressor (Unoptimized)

The original Random Forest model explained a low percentage of variance even when the prediction conducted remained the same. It had a performance of MSE = 0.04263, RMSE = 0.20648, MAE = 0.15929 and R2 = 0.08776, which represented moderate correspondence among forecasted and actual reactions.

4.4.2. XGBoost Regressor (Hasn't been optimized)

The numerical performance was a bit better with the baseline XGBoost with MSE = 0.04262, RMSE = 0.20645, MAE = 0.15922, and R2 = 0.08796. The model was very consistent and it managed the data as well even without tuning. The Lasso Regression (pre-optimization) is based on a linear regression formulation that predicts the dependent variable Y using the linear predictors (X).

4.4.3 Lasso Regression (Prior to Optimization)

Lasso Regression is a linear regression formulation that estimates the dependent variable Y as a linear predictor of Y using the linear predictors (X). The first Lasso yielded the results of MSE = 0.04372, RMSE = 0.20909, MAE = 0.16383 and R2 = 0.06455. Although the model had minimized coefficient noise, the linear construction restricted predictive ability.

4.4.4 ElasticNet Regression (Unoptimized)

ElasticNet was slightly better than Lasso with RMSE = 0.20778, MAE = 0.04317, MSE = 0.04317. = 0.16205, and R2 = 0.07626. The equilibrium between L1 and L2 sentences enhanced the stability of the model.

4.4.5 SupportVector Regressor (Pre Optimization)

SVR showed MSE = 0.04447, RMSE = 0.21087, MAE = 0.15582, and R2 = 0.04850. Although the model forecasted in more absolute terms (smaller MAE), but had difficulties in explaining the variance overall.

4.4.6 MLP Regressor (Unoptimized)

The single MLP had MSE = 0.04346, RMSE = 0.20848, MAE = 0.16571 and R2 = 0.06997. The network acquired a non-linear type of behavior although it had to be fine-tuned to achieve improved generalization. After Data Optimization

4.4.7 Random Forest (Having Optimized Dataset)

Random Forest continued to achieve the same results after refining the data: MSE = 0.04263, RMSE = 0.20648, MAE = 0.15929, R2 = 0.08776. There was slight gain on the noise however placed in the model with no significant visual change. XGBoost (Optimized Dataset)

4.4.8 XGBoost (After Optimization):

The next improvement was reached with MSE = 0.04254, RMSE = 0.20626, MAE = 0.15922 and R2=.52. = 0.08964, and finishes the task fast in 90.98 seconds.

4.4.9 Lasso Regression (Having Optimized Dataset)

These values were refined to give MSE = 0.04610, RMSE = 0.21470, MAE = 0.16921 and R 2 = 0.01366. The performance decreased by a complementary slight amount indicative of the fact that intense regularization eliminated useful signal.

4.4.10 Elastic Net Regression (Optimize Dataset)

The results of ElasticNet were that MSE = 0.045544, RMSE = 0.213405, MAE = 0.16804, and R2 = 0.02556. Also, in the same way Lasso, the model was more conservative and lost a predictive strength.

4.4.11 Support Vector Regressor (Having Optimized Dataset)

SVR achieved MSE = 0.04443, RMSE = 0.21079, MAE = 0.15585, and R2 = 0.04924 and followed a similar trend as the baseline but with cleaner inputs. MLP Regressor (Optimized Dataset)

4.4.12 MLP Regressor (After Optimization).

The MLP performed better with an MSE = 0.043134, RMSE = 0.207681, MAE = 0.157331 and R2 = 0.077143 which indicates more consistency in the fit.

Hyperparameter-Optimized Models

4.4.13 Optimized XGBoost

One of the highest results, in general, was provided through the tuned XGBoost. It had highest explanatory power of all the individual models; it gave a MSE = 0.04238, RMSE = 0.20587, MAE = 0.15959, and R2 = 0.09308. It took 6298.92 seconds to be fully tuned.

4.4.14 Optimized MLP

The grid-search optimization was used to optimize the MLP to achieve: MSE = 0.04284, RMSE = 0.20698, MAE = 0.16272 and R2 = 0.08331. It was not the best, but offered a predictor which is non-linear.

Table 2: Summarized Table of Results

Model & Stage	MSE	RMSE	MAE	R2	Notes

Random Forest (Baseline)	0.04263	0.20648	0.15929	0.08776	Stable baseline performance
XGBoost (Baseline)	0.04262	0.20645	0.15922	0.08796	Slightly better than RF
Lasso (Baseline)	0.04372	0.20909	0.16383	0.06455	Limited variance capture
ElasticNet (Baseline)	0.04317	0.20778	0.16205	0.07626	Balanced regularization
SVR (Baseline)	0.04447	0.21087	0.15582	0.04850	Good MAE but weak R^2
MLP (Baseline)	0.04346	0.20848	0.16571	0.06997	Moderate non-linear learning

Model & Stage	MSE	RMSE	MAE	R2	Notes
Random Forest (Optimized)	0.04263	0.20648	0.15929	0.08776	Similar to baseline
XGBoost (Optimized)	0.04254	0.20626	0.15922	0.08964	Efficiency + slight improvement
Lasso (Optimized)	0.04610	0.21470	0.16921	0.01366	Performance decreased
ElasticNet (Optimized)	0.04554	0.21340	0.16804	0.02556	Lost predictive strength
SVR (Optimized)	0.04443	0.21079	0.15585	0.04924	Comparable to baseline
MLP (Optimized)	0.04313	0.20768	0.15733	0.07714	More stable generalization
XGBoost (Grid Search)	**0.04238**	**0.20587**	0.15959	**0.09308**	**Best overall performance**
MLP (Grid Search)	0.04284	0.20698	0.16272	0.08331	Improved non-linear fit

The table above which is made out as a summary will assist in realising the end result:

All of the models were tested on the optimized dataset. Linear models (Lasso, ElasticNet) performed poorly whereas SVR had a minor improvement in MAE. Non-linear patterns were better represented by the tree-based models (Random Forest, XGBoost), and, the XGBoost had the greatest R2. There was also after tuning enhancements of the MLP. All metrics are summed

up in the table with differences of the baseline and optimization and hyperparameter models and the gains of the optimization and hyperparameter model.

4.5 Result Analysis:

4.5.1 Performance Overview of the models.

All of the models were tested on the optimized gene expression data set in terms of their prediction. Other linear models such as Lasso and ElasticNet had low predictive power with R2 values less than 0.08, which implied that these models could not reflect intricate non- linear associations between drug response and gene expression. SVR worked a little better regarding MAE but it still explained a little bit of variance. The tree-based models, especially, the random forest and XGBoost, were more effective in getting non-linear patterns and XGBoost had the best R2 per model. Tuned MLP best demonstrated the ability to learn the non-linear patterns that were not viable with linear models, though. The graphs below demonstrate R2 scores of all models which demonstrate the high performance of tree-based and neural network models over linear methods. The above table represents a summary of all the metrics consisting of MSE, RMSE, MAE as well as R2 of the baseline, optimized and pipeline models.

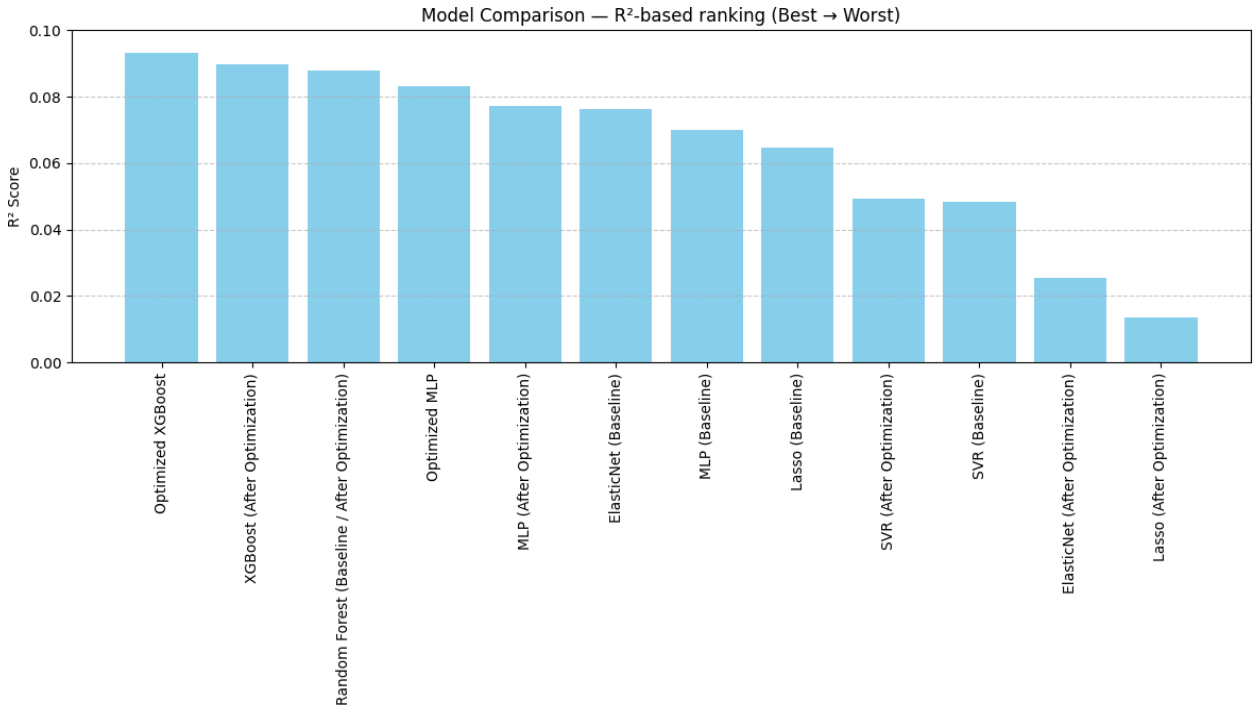


Figure 6: Model Comparison depending on the R 2 score.

4.5.2 Hyperparameter Tuning and Optimization Effect.

The use of optimization and hyperparameter tuning resulted in observable modeling performance. XGBoost has R2 of 0.09308, the optimal figure that has been able to explain the highest explanatory power of all the models after grid search optimization. The optimization of the MLP also occurred in the sense that the Hyperparameter adjustments were able to significantly boost the predictions of the neural network as indicated by the improved R 2.

Conversely, the linear models like Lasso and ElasticNet their results indicated slight improvements or even a reduction in the performance, underscoring the significance of model selection and preprocessing of data in the high-dimensional analysis of gene expressions. The following graph reveals the comparison of the RMSE values in the pre- and post-optimization processes and it is evident that the reduction in error was higher in tree based and neural network models.

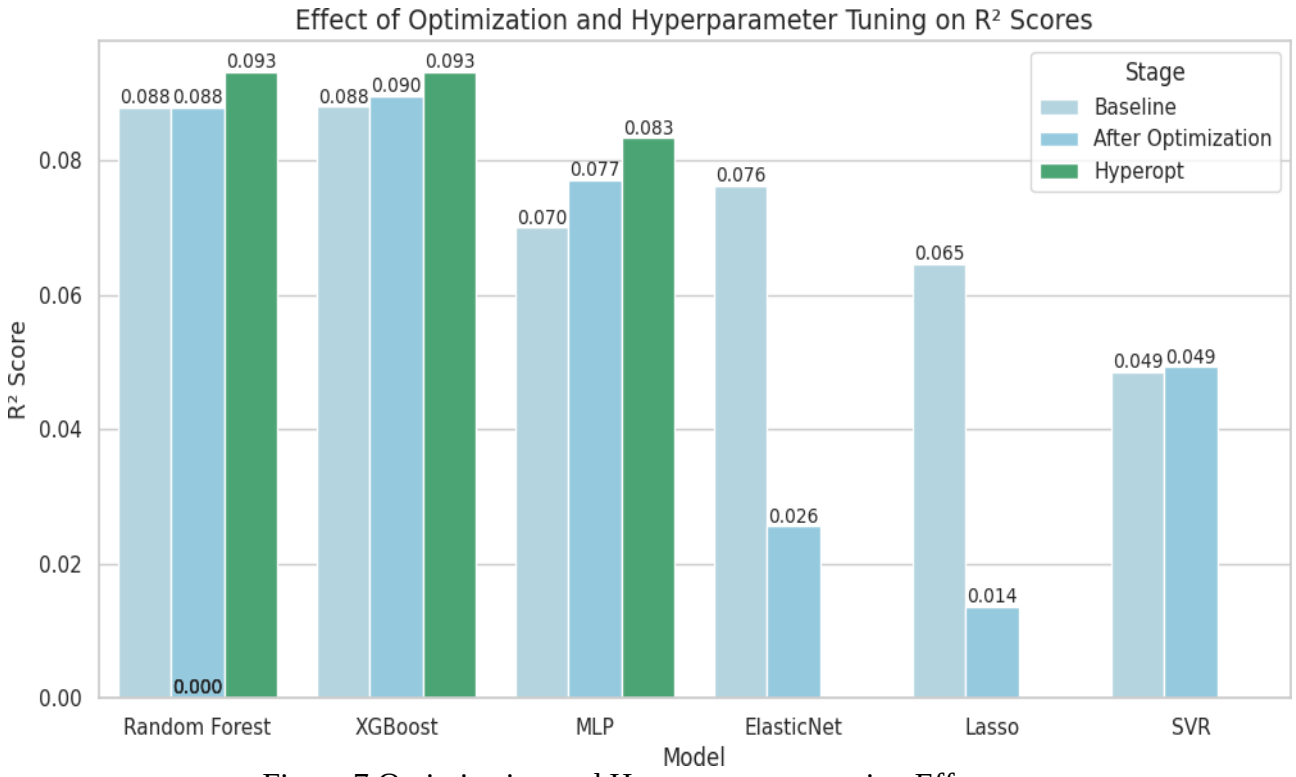


Figure 7:Optimization and Hyperparameter tuning Effect.

4.5.3 Pipeline and Ensemble Analysis.

The last pipeline that featured the use of optimized XGBoost and MLP selections and method of feature selection gave the most credible predictions. Focusing on most informative genes the pipeline minimized noise and overfitting leading to reduced errors and increased R2. It proves that dimensionality reduction coupled with complementary modelling techniques can be integrated to improve predictive accuracy and stability. Comparison of the final pipeline with individual models (with R2 and RMSE) is done in the graph Below. As well, residual plots (Figure W) indicate better generalization as the feedback of the observed drug responses is predicted in a closer way by the pipeline.

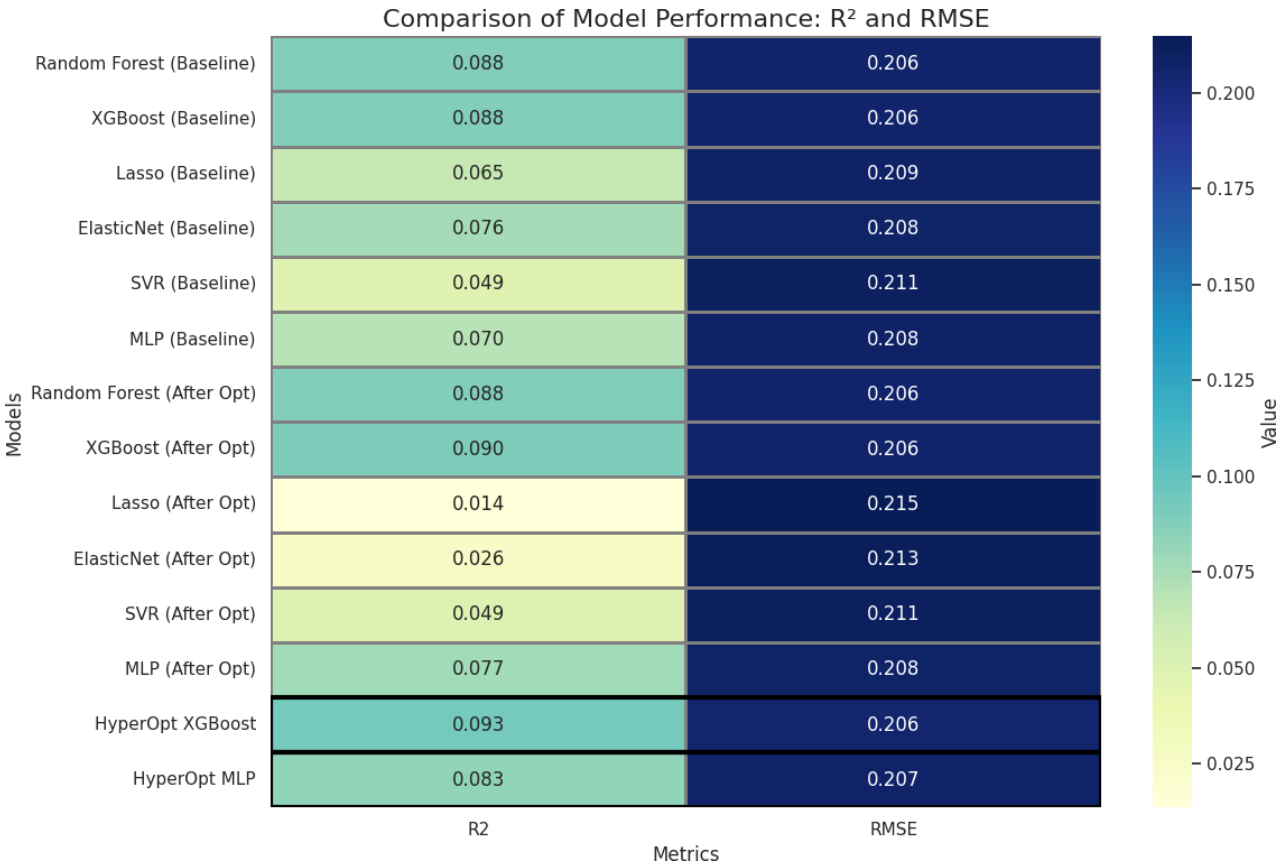


Figure 8: Model Comparison of Model Performances using R2 and RMSE

4.5.4 Biological Insight and Importance of Features:

The analysis of the most important features based on tree-based models was used to determine the top-ranking predictive genes as far as response to the drug is concerned. The scores of the relative importance are graphically demonstrated using a row-chart of the 30 most important

genes. A good number of these genes are associated with drug targets or pathways of interest, which makes such genes biologically interpretable and predictive. These insights make the computational model more practical.

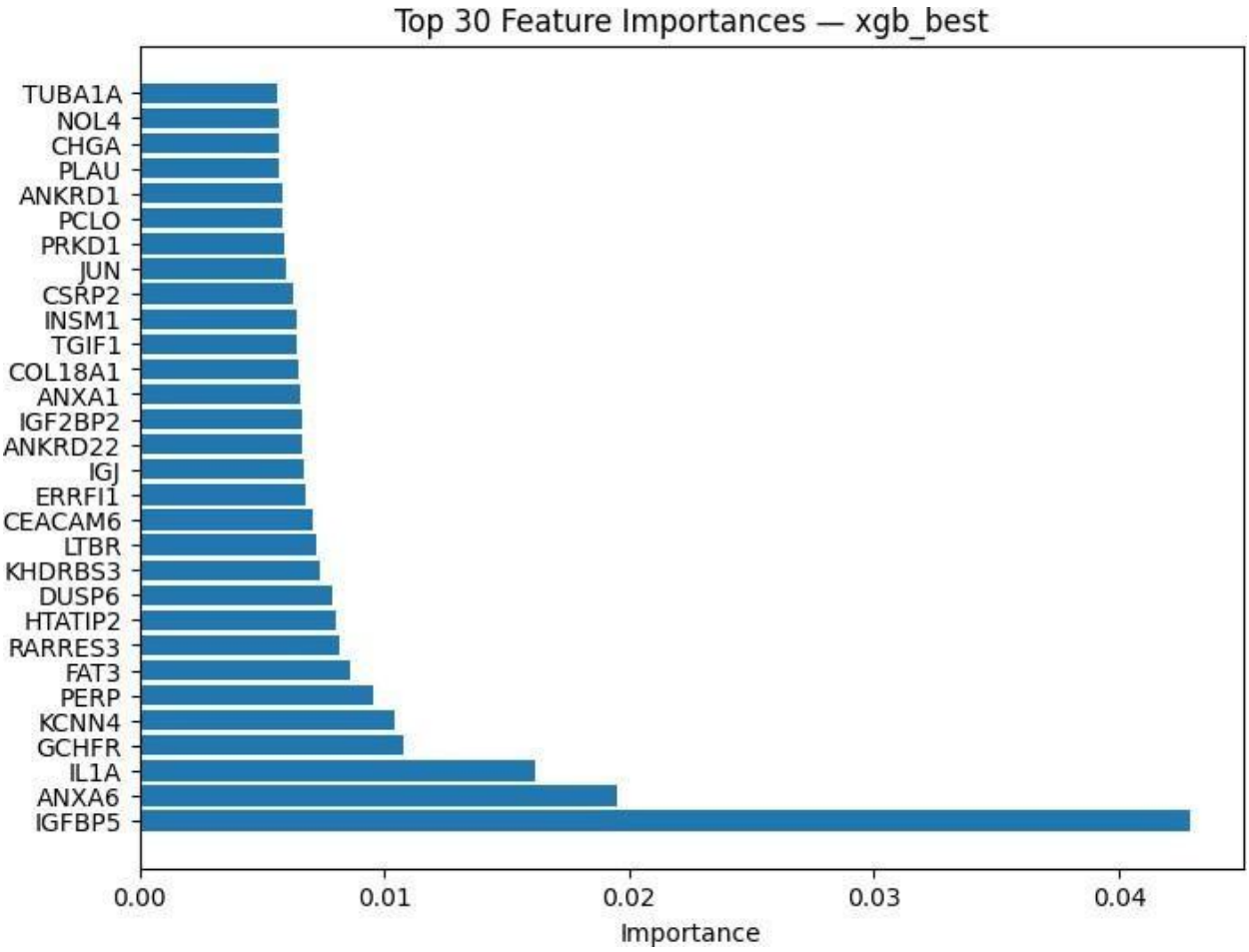


Figure 9: The top 30 most important features

Was to exclude top 30 feature genes since in gene expression data most of the genes are independent and does not depend on others although there are certain genes which are correlated and they carry a significant influence in drug predicting. The correlation factor of the top 30 selected gene features gives a good picture of the relationship of such biologically relevant predictors to each other. The majority of the features have moderate to low pairwise correlations, meaning that the selection process was able to obtain a variety of genes that convey distinct information to the model and not redundant cues. Several minor clusters of genes have greater intra-cluster correlations, implicating that these groups can engage in the correlated paths, or be

functionally relevant in the control of drug response. The conclusion is that the structure of the matrix is characterized by a combination of independent and biologically related aspects in a balanced way, which shows their appropriateness to downstream modeling and interpretation.

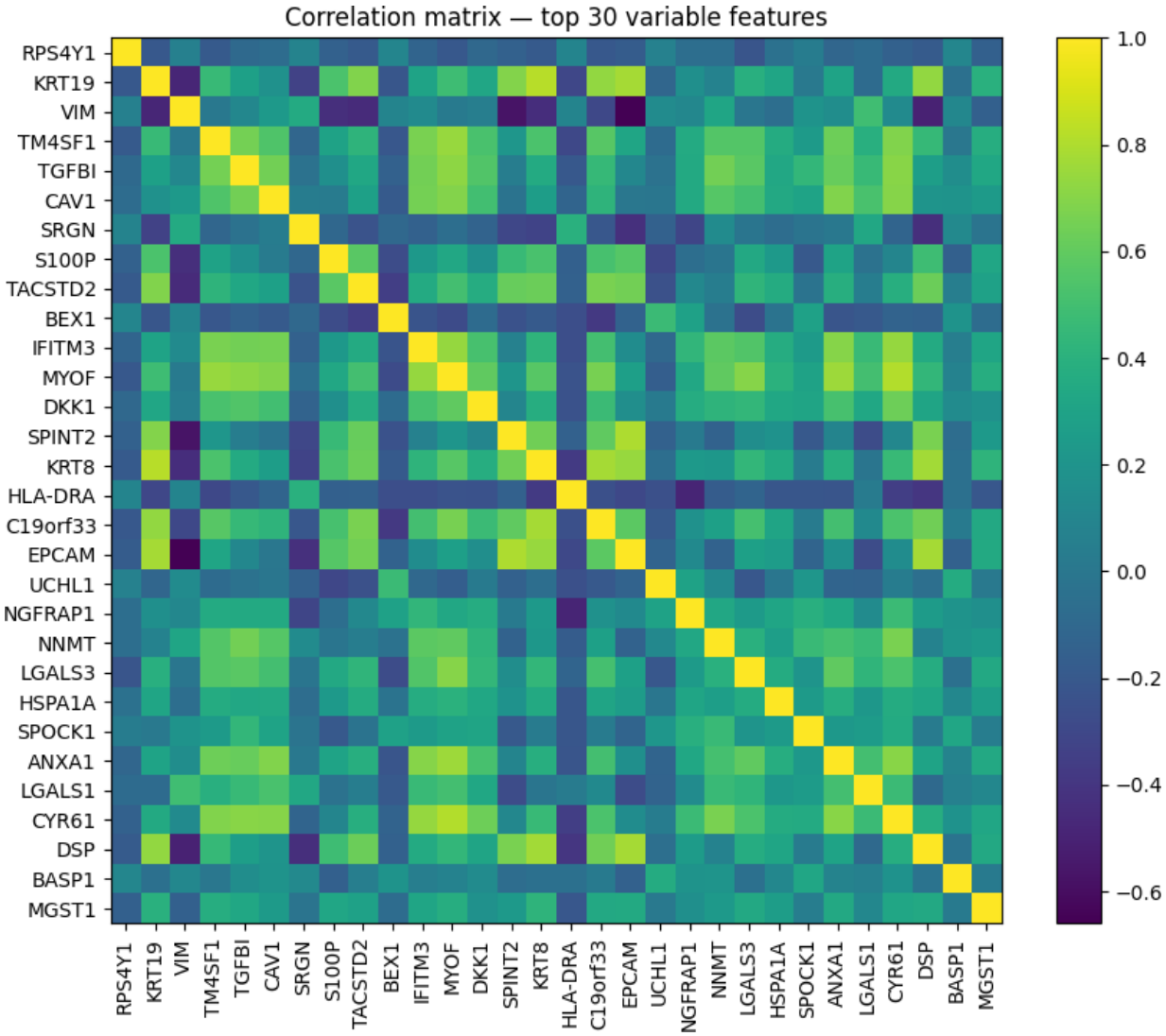


Figure 10: correlation- 30 leading variable features

4.5.5 Best Model and its performance on prediction:

The XGBoost Regressor and the MLP Regressor were the two most successful models in the prediction of drug responses after optimizing the hyperparameters of the models through the entire modeling pipeline. Both models performed better than their respective baseline models

and better generalization and stability across the dataset have been observed. The learned XGBoost model gave the best overall performance, and the optimized MLP. model was a good secondary performer, instead potentially providing complementary information of non-linear relationship within the gene expression features.

4.5.6 Learning Curve Analysis

In order to test the stability and the generalization property of the optimized models, learning curves were obtained on both the XGBoost and the MLP regressors. The learning curves show the trend of the training and validation errors as the size of the training set increases which can give a sense of whether the models can be underfitting, overfitting, or well-balanced. In the XGBoost Regressor, the training and validation curve approach parallel as additional samples are added, showing that they have reached a stable learning rate with limited the ability to overfit.

The marginal separation between the curves indicates that XGBoost is able to model the underlying organization of the data of gene-drug responses without an overfitting of noise. Such a behavior is consistent with the high performances in the final evaluation metrics. By contrast, the MLP Regressor indicates a smaller difference between training and validation errors, especially during the initial consultancy of training. Whereas the differences are reduced with increasing amount of data, the model still has weak overfitting biases, probably because of the intricacy of the neural network and robustness of delving features, which is the dimensions of gene expression. However, the general tendency of the curve shows that the MLP continues to enjoy the advantages of more data and does not experience the large scale overgeneralization as the training set gets progressively bigger.

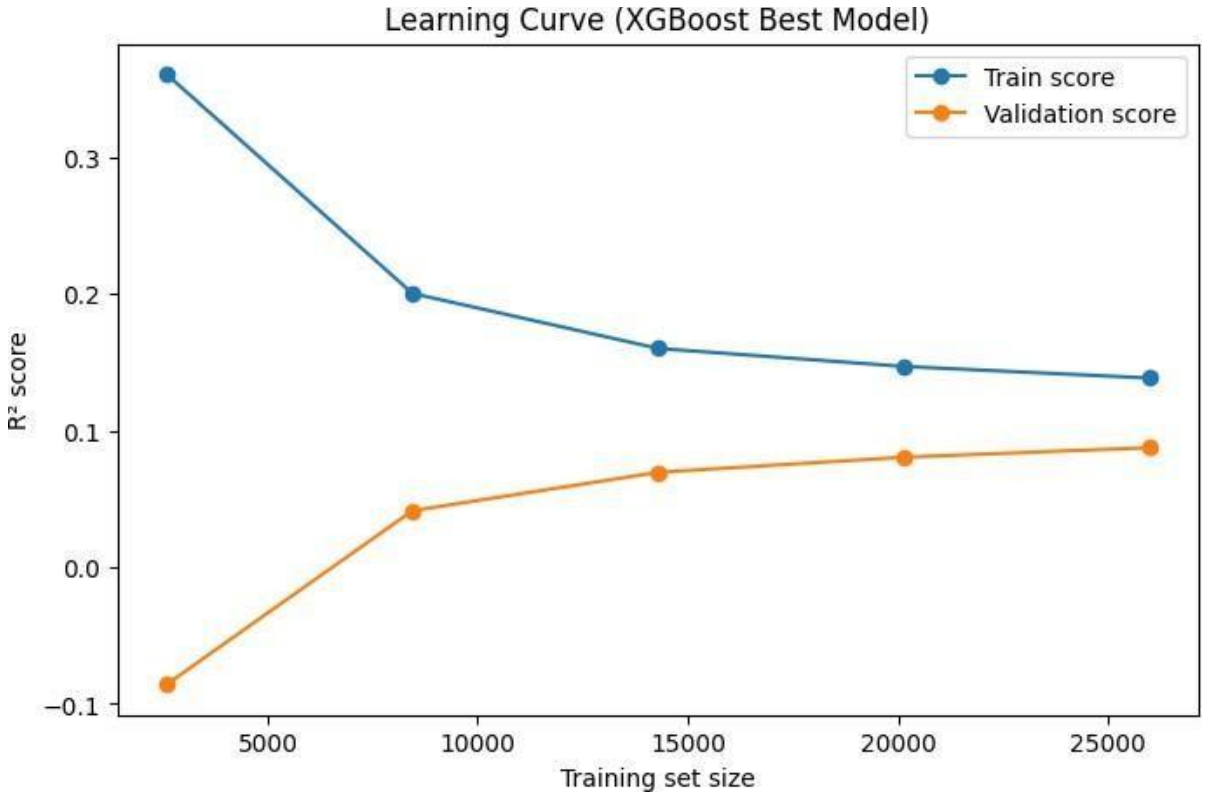


Figure 11: XGBoost Model learning curve.

4.5.7 Actual and Predicted AUC Plots.

The Actual vs Predicted AUC scatter plots offer a direct graphical evaluation of any individual model in terms of it predicting in accordance with the actual values of drug responses. Both the optimized XGBoost Regressor and the optimized MLP Regressor have a high positive correlation suggesting that the models were effective in capturing the relationship that existed between the drug sensitivity and the gene expression features. The fact that most of the points are located on the diagonal regression line is another proof of having a consistent overall directional trend of the predictions with the actual AUC values indicating that the selected set of features is biologically relevant. Although this is the general consensus, there are two dominant patterns of errors that arise in the two models. To start with, systematic over-prediction can be observed in lower end of the AUC (0.0-0.6). The scatter points in this area are also often above the y=xy = xy=x line implying a tendency to over estimate weak or ineffective response of drugs. Such behaviour indicates that this behavior makes the models less effective at accurately learning subtle differences among poor-performing sets of drug-cell lines, potentially because of noise or reduced signal resolution in this part of the data. Second, to achieve optimal prediction accuracy,

there is a saturation of prediction towards the upper AUC limit (around 1.0) (Hardy 126). In this case, the two models produce thick groups of predictive values near the top score, namely lower differentiation capability with respect to the most efficient drugs. The phenomenon is often seen in the case of regression with regressors with boundaries on the biological phenomena under study, where there is a limit to how much the edge of phenomena can be learned. Comparing the two models, the predictions of the XGBoost show a clearer and less dispersed convergence around the regression line, which represent higher predictive accuracy and lesser variance of the residual. This point can be traced back to the numerical analysis where XGBoost has got the largest R2 and the least error rates. Conversely, the MLP plot reports a slightly larger spread driven by the model being more sensitive to variance as well as moderately overfitting.

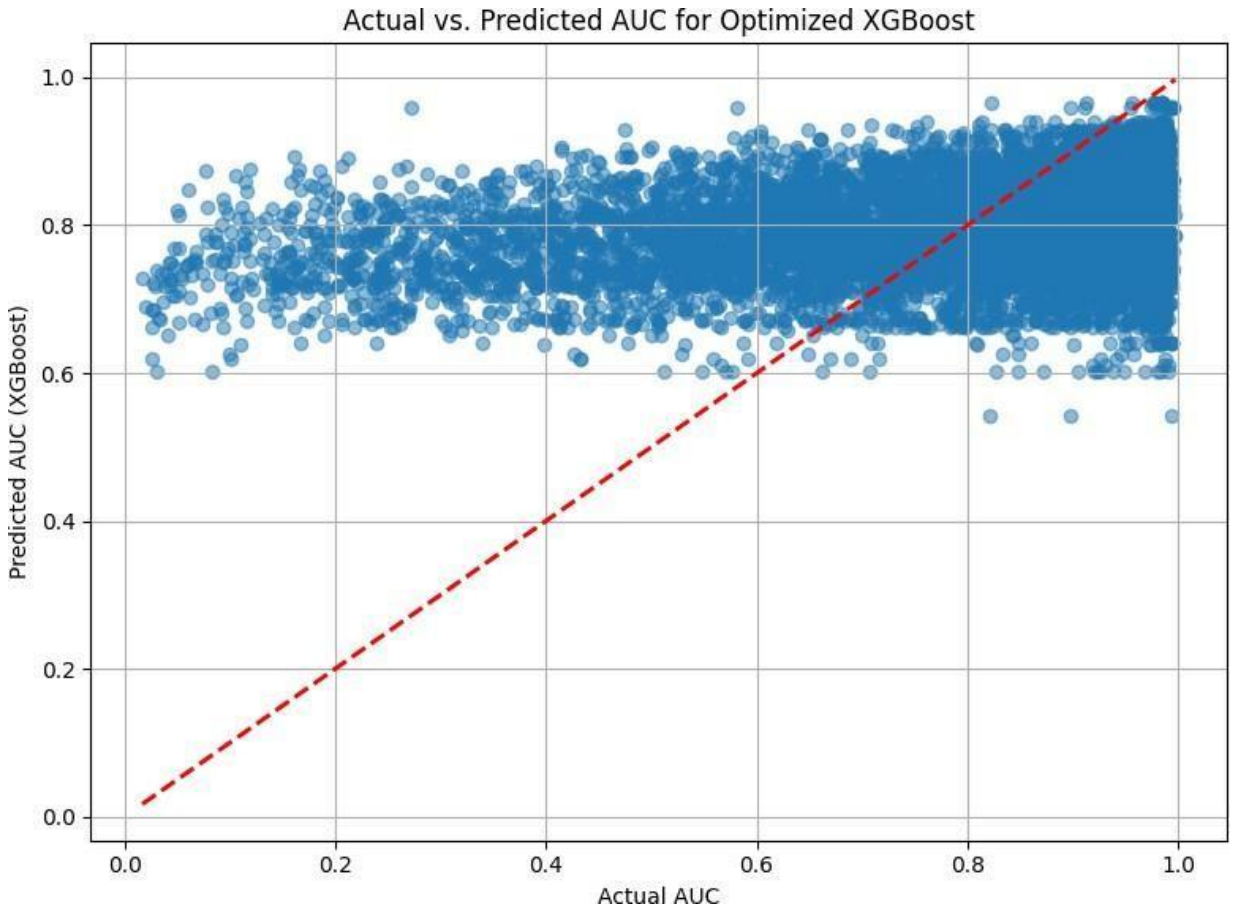


Figure 12: Predicted vs Actual of Hyper Optimized XGBoost.

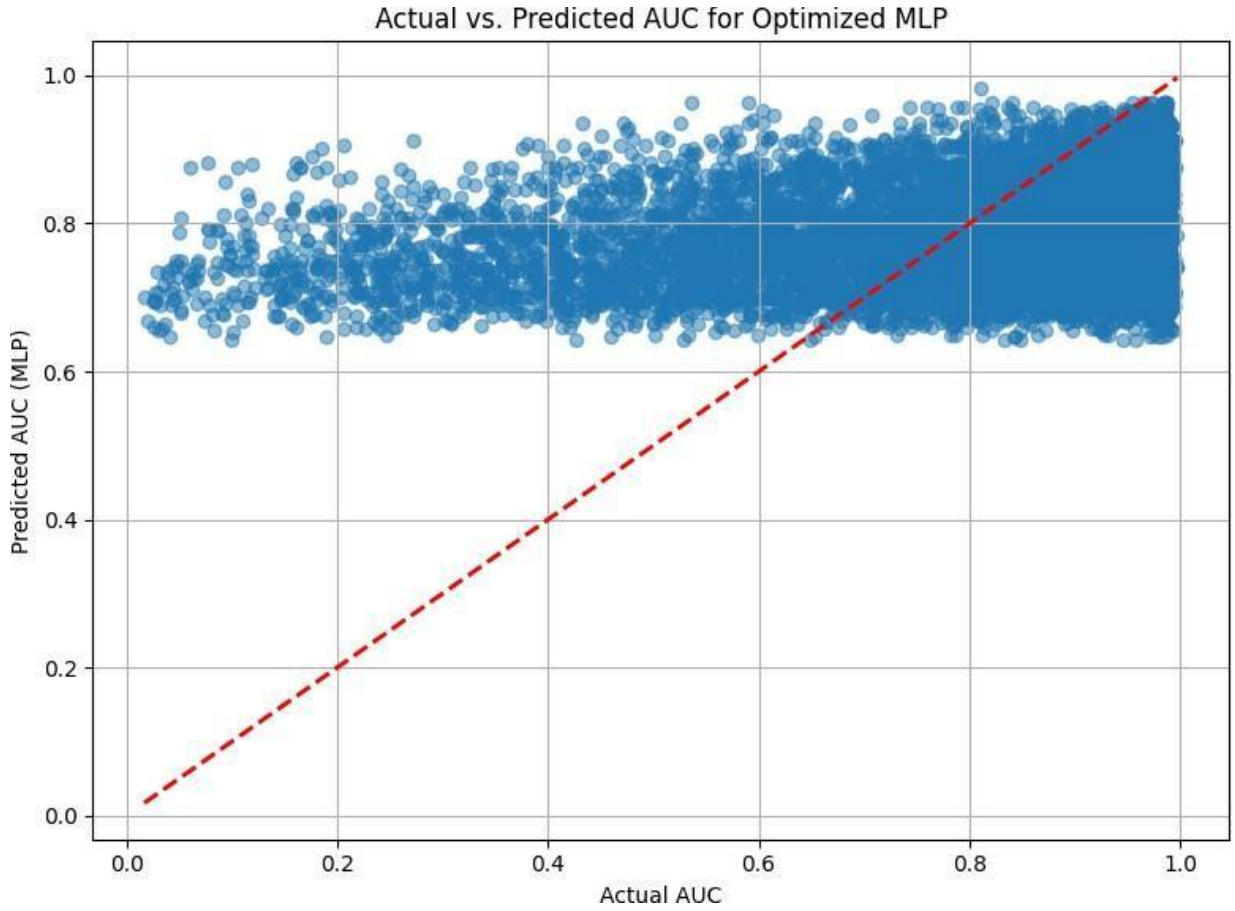


Figure 13: measured vs theoretical AUC of Hyper-Optimized model of MLP model.

4.5.7 Overall Result Summary

It can be seen in the analysis of the optimized models that both the XGBoost Regressor and the MLP Regressor were able to identify the essence of the relationship between the pattern of gene expression and the response to a given drug, although XGBoost was always the strongest performer in the analysis. In every measure and visual evaluation, such as, the performance scores, the Actual vs Predicted AUC plots, and the learning curves, the rank of feature importance and the correlation tests, the XGBoost model achieved higher accuracy, increased extrapolation and stationary learning characteristics. Also the MLP model provided meaningful results and showed slight overfitting trends and severe variability. Altogether, the results affirm the position of the optimized set of features and modeling strategy as the solid basis of drug sensitivity prediction, as well as, pinpoints several key areas, including, but not limited to. high values of AUC--where predictive uncertainty persists and future improvement can be valuable.

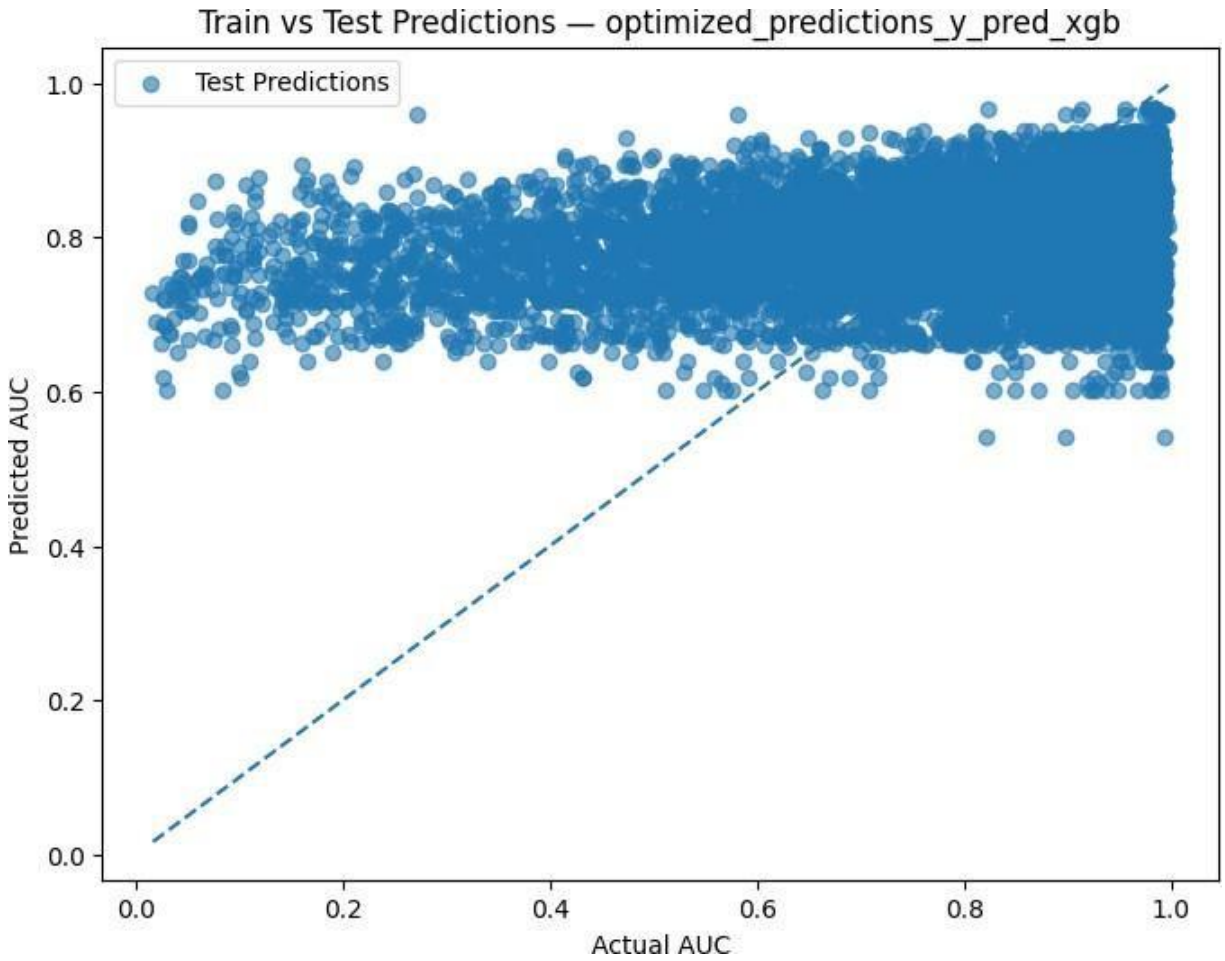


Figure 14: Figure of Train vs Test Prediction in the case of overall

CHAPTER 5

CONCLUSION

5.1 Findings and contributions

This thesis develops a full-fledged machine learning system to predict responses of drugs based on high-dimensional gene expression data. The analysis of patterns in drug sensitivity in cell lines, characterized by systematic preprocessing, feature selection, and hyperparameter- optimized regression models, was successful in the study, revealing biologically meaningful patterns that can be used to further investigate the study topic. The best XGBoost Regressor was found as the optimized, with the best values of R2, MSE, and RMSE, visual analysis of the actual versus predicted AUC plot, the residual plot and the learning curve. Besides good predictive performance, the study provides an interpretable set of genes of interest whose best- performing gene predictors correspond to biological understanding of drug response. The article also presents an optimized and structured ML pipeline that is evidently better than baseline models, which establishes a practical tool and a biologically-informed basis of future drug sensitivity studies.

5.2 Limitations

Even though the models worked, there are certain shortcomings. Extreme AUC areas expose the predictive ability of the models to inaccuracies with low accuracy response overestimation and high accuracy response underestimation at one end and the opposite on the other end, which is where the opportunity to calibrate and/or define specialized loss functions exists. The analysis also only uses the data of gene expression and therefore the study may not capture any in-depth biological interactions, which might need the involvement of proteomic, genomic, or epigenomic signals. Also, despite having a smaller dimensionality and significant genes that feature selection highlighted, feature selection might not provide a full picture of pathways or network-level biological interactions. Finally, the dataset is also limited to cell-line based measurements

implying that the findings cannot necessarily be extrapolated to a patient sample without additional validation with clinically derived datasets..

5.3 Future improvements

Further innovations of this study may aim at incorporating multi-omics data like proteomics, methylation or mutation profiles into development of more effective and biologically complete predictive models. Network-informed or pathway-informed feature selection methods can also be used to increase interpretability by including information about gene-gene interactions and signaling. Extreme model behavior can be enhanced by means of weighted loss functions, data augmentation, or idealized calibration. Deep learning non-linear patterns, not visible in traditional models, can also be discovered in architectures that are regularly characterized appropriately.

REFERENCES

1. Costello, J. C., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12), 1202–1212. https://doi.org/10.1038/nbt.2877
2. Barretina, J., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391), 603–607. https://doi.org/10.1038/nature11003
3. Garnett, M. J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391), 570–575. https://doi.org/10.1038/nature11005
4. Iorio, F., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3), 740– 754. https://doi.org/10.1016/j.cell.2016.06.017
5. Menden, M. P., et al. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties.*PLoS ONE*, 8(4), e61318. https://doi.org/10.1371/journal.pone.0061318
6. Ding, M. Q., et al. (2018). Precision oncology beyond targeted therapy: combining omics-driven insights with machine learning. *Briefings in Bioinformatics*, 20(3), 798–809. https://doi.org/10.1093/bib/bbx143
7. Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., & Margolin, A. A. (2014). Systematic assessment of multi-gene predictors of drug sensitivity in cancer. *PLOS Computational Biology*, 10(4), e1003582. https://doi.org/10.1371/journal.pcbi.1003582
8. Geeleher, P., Cox, N. J., & Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels using machine learning. *Nature Communications*, 5, 5053. https://doi.org/10.1038/ncomms6053
9. Cichonska, A., et al. (2021). Learning drug response prediction from biological networks using graph neural networks. *Nature Communications*, 12, 3601. https://doi.org/10.1038/s41467-021- 23869-z
10. Zhao, P., et al. (2020). Ensemble machine learning models for predicting drug response in cancer.
    Frontiers in Genetics, 11, 577306. https://doi.org/10.3389/fgene.2020.577306
11. He, X., et al. (2018). Predicting drug sensitivity of cancer cells using gene expression profiles.
    Bioinformatics, 34(22), 3862–3869. https://doi.org/10.1093/bioinformatics/bty454
12. Hao, J., Kim, Y., Mallavarapu, T., Oh, J. H., & Kang, M. (2018). Interpretable deep

neural networks for predicting drug response in cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6), 1940–1950. https://doi.org/10.1109/TCBB.2018.2887098

13. Aben, N., Vis, D. J., Michaut, M., & Wessels, L. F. A. (2016). TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(14), i413–i420. https://doi.org/10.1093/bioinformatics/btw432
14. Kuenzi, B. M., et al. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cell Systems*, 11(5), 498–511.e3. https://doi.org/10.1016/j.cels.2020.09.007
15. Sharifi-Noghabi, H., et al. (2019). MOLI: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14), i501–i509. https://doi.org/10.1093/bioinformatics/btz318
16. Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., & Kim, T.-Y. (2018). Cancer drug response profile scan using deep neural network. *Scientific Reports*, 8, 8857. https://doi.org/10.1038/s41598-018-27178-5
17. Rahman, R., Hu, H., & Wang, D. (2019). Optimizing cancer drug response prediction using transformer-based neural networks. *BMC Bioinformatics*, 20(Suppl 24), 671. https://doi.org/10.1186/s12859-019-3318-7
18. Ma, T., Zhang, A., & Xue, Y. (2018). Machine learning methods for precision oncology: Predicting cancer drug sensitivity. *Frontiers in Pharmacology*, 9, 1426. https://doi.org/10.3389/fphar.2018.01426
19. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modeling techniques for genomics. *Nature Reviews Genetics*, 20(7), 389–403. https://doi.org/10.1038/s41576-019-0122-6
20. Ali, M., Aittokallio, T., & Kallioniemi, O. (2016). Machine learning and feature selection for cancer drug response prediction. *Journal of Biomedical Informatics*, 64, 168–178. https://doi.org/10.1016/j.jbi.2016.10.007

APPENDICES

Appendix A: Dataset Availability

Dataset Link : https://sites.broadinstitute.org/ccle/datasets?utm_source=chatgpt.com

Downloadable Dataset:

https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/preprocessed/Cell_line_RMA_proc_basalExp.txt.zip

LIBRARY CLEARANCE

PLAGARISM REPORT

221-35-1021

ORIGINALITY REPORT

18%	16%	12%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	publikationen.sulb.uni-saarland.de Internet Source	1%
3	Submitted to Midlands State University Student Paper	1%
4	ouci.dntb.gov.ua Internet Source	1%
5	www.frontiersin.org Internet Source	<1%
6	ijarsct.co.in Internet Source	<1%
7	Submitted to University of Adelaide Student Paper	<1%
8	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
9	Submitted to Vrije Universiteit Amsterdam Student Paper	<1%
10	umpir.ump.edu.my Internet Source	<1%
11	Submitted to Manipal International University Student Paper	<1%
12	Submitted to Liverpool John Moores University Student Paper	<1%
	Submitted to University of Sydney	

ACCOUNT CLEARANCE