Using machine learning to understand the link between gene essentiality, gene expression and the chemosensitivity of cancer cells.

Kuhle Mcinga

MSc (Med) in Bioinformatics

Division of Computational Biology

Department of Integrative Biomedical Sciences

Faculty of Health Sciences

Supervisor: Dr Musalula Sinkala

Co-supervisor: Associate Professor Darren Martin

2023

Declaration

I *Kuhle Mcinga*, hereby declare that the work on which this dissertation is based on my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signed by candidate

11 October 2023

Acknowledgements

Thank you, God.

I also want to thank my family—Toko Mcinga, Nobulungisa Mcinga, Mlibo Mcinga, and Bulela Mcinga—for being my biggest supporters.

I am also grateful to Nosipho Mabizela and Sisipho Sathula. They have supported me, listened to me complain about my project, and even though they are busy with their own projects, they have helped me and celebrated my success.

I am also grateful to my supervisors, Dr Musalula Sinkala and Prof. Darren Martin for their insights and for allowing me to run this project in my own direction.

I also want to extend my gratitude to the UCT Master's Research Scholarship and the National Research Fund for funding this project.

Lastly, I want to share a quote from esteemed rapper Snoop Dogg: "I would like to thank me for believing in me, I want to thank me for doing this hard work. I want to thank me for never quitting and I want to thank me for being me at all times." These words connect with my journey.

Abstract

The emergence of pharmacogenomics databases has presented unique opportunities to leverage machine learning in precision medicine, particularly in drug response prediction. In this thesis, an in-depth investigation is conducted on carefully curated and integrated breast cancer focused datasets from the GDSC (Genomics of Drug Sensitivity in Cancer) and Achilles (CRISPR derived) project databases. Specifically, machine learning techniques are employed to accurately predict the drug responses of cancer cells, laying the groundwork for personalised treatment strategies. Through rigorous training of machine learning models, drug-response classifiers were devised that demonstrated remarkable predictive capabilities, with the best performing classifier achieving an F1-score of 0.86 and an AUC of 0.85, indicating its effectiveness in drug response prediction. Training these models on GDSC and Achilles datasets encompassing various drug IC50 values, ensured generalization of the models across different drugs and cell lines. A model explainability technique, the permutation feature method, was employed to identify the most influential genes contributing to drug response predictions. This provided deeper insights into the molecular mechanisms governing drug sensitivities in breast cancer. Pathway enrichment and kinase enrichment analyses were used, to yield additional insights into key signalling pathways and a list of potential therapeutic targets for treating breast cancer was created. To gain a broader perspective, gene expression changes were investigated in both drug-sensitive and drug-resistant cell lines by integrating the GDSC and LINCS (Library of Integrated Network-Based Cellular Signatures) datasets and the GDSC and Cancer Cell Line Encyclopaedia (CCLE) datasets to uncover potential breast cancer specific features that could be targeted for precision medicine applications. Overall, these integrative analyses using machine learning enabled the accurate prediction of drug responses and uncovered novel therapeutic avenues for personalized treatment approaches in breast cancer patients.

Contents

List of abbreviations

ANN: Artificial Neural Network

AUC: Area Under the Curve

CCLE: Cancer Cell Line Encyclopaedia

CTRP: Cancer Therapeutic Response Portal

CTD2: Cancer Targe Discovery and Development

DNN: Deep Neural Networks

GDSC: Genomics of Drug Sensitivity

GBM: Gradient Boosted Machine

IC50: Half maximal inhibitory concentration

KNN: K-Nearest Neighbour

KEA: Kinase Enrichment Analysis

KEGG: Kyoto Encyclopaedia of Gene and Genomes

LINCS: Library of Integrated Network-Based Cellular Signatures

MAE: Mean Absolute Error

MRMR: Maximum relevance Minimum Redundancy

PEA: Pathway Enrichment Analysis

R^2: Coefficient of Determination

RF: Random Forest

RMSE: Root Mean Square Error

ROC: Receiver Operating Curve

ROS: Reactive Oxygen Species

svmRadial: Support Vector Machine Radial

UCT HPC: University of Cape Town High-performance computing

1. Literature Review

1.2 Introduction

Worldwide, about 10 million people were impacted by cancer in 2020, making it one of the leading causes of human mortality [1]. With over 2.3 million cases of breast cancer cases reported, breast cancer surpassed lung cancer as the single most diagnosed cancer worldwide [1]. Accordingly, breast cancer was the leading cause of death in women, with the highest incidence rates reported in those between the ages of 45 and 65 [2].

Cancer can develop when normal cells undergo genetic alterations affecting the intracellular signalling pathways that control various cellular processes. While these alterations generally differ between tumours of different tissues, they can also differ between different tumours of the same tissue [3,4]. Since the response of tumours to drugs is influenced by genetic changes that impact gene expression, even among individuals with the same type of cancer, there is significant variability in how their tumours react to anti-cancer drugs [4].

The promise of personalised medicine is the development of treatment regimens that, by targeting the cancer cells of a particular patient based on the specific genetic peculiarities of those cells, it should be possible to kill the cells without any collateral harm to the patient. Of particular interest in this regard is the identification of genetic and/or other molecular characteristics of cancer cells that are responsible for specific clinical outcomes [5]. The combination of these tumour-specific characteristics could be utilised to predict the impact(s) of any given drug on a both the healthy and tumour cells of a patient. Specifically, establishing the existence of statistically significant associations between the presence of particular genetic/molecular characteristic of tumour cells and the degree of sensitivity displayed by these cells to particular drugs establishes the groundwork for tailoring drug therapies according to

the genomic contexts of individual patients [6]. Recent research has examined the mutation profiles, DNA methylation statuses, copy number abnormalities, and gene expression patterns of a vast number of cancer cell lines in the presence and absence of hundreds of different compounds to facilitate the identification of genomic markers related to both the development of cancers and the efficacy of different potential anti-cancer drugs [7–10]. Due to the abundance of these drug perturbation and genetic datasets, computational systems analyses are essential for assisting clinical judgments. Using machine learning and artificial intelligence techniques (Figure 1), such analyses promise the inference of actionable insights into the molecular mechanisms underlying cancer progression in individual patients.

This review discusses (1) the dysregulation of signalling pathways (2) the significance of essential "driver" genes in the development of cancer; (3) the sources of the datasets used to train machine learning models; (4) the machine learning algorithms that can be used to predict the drug responses of cancer cell lines; and (5) the methods used to assess and evaluate the predictive power of models.

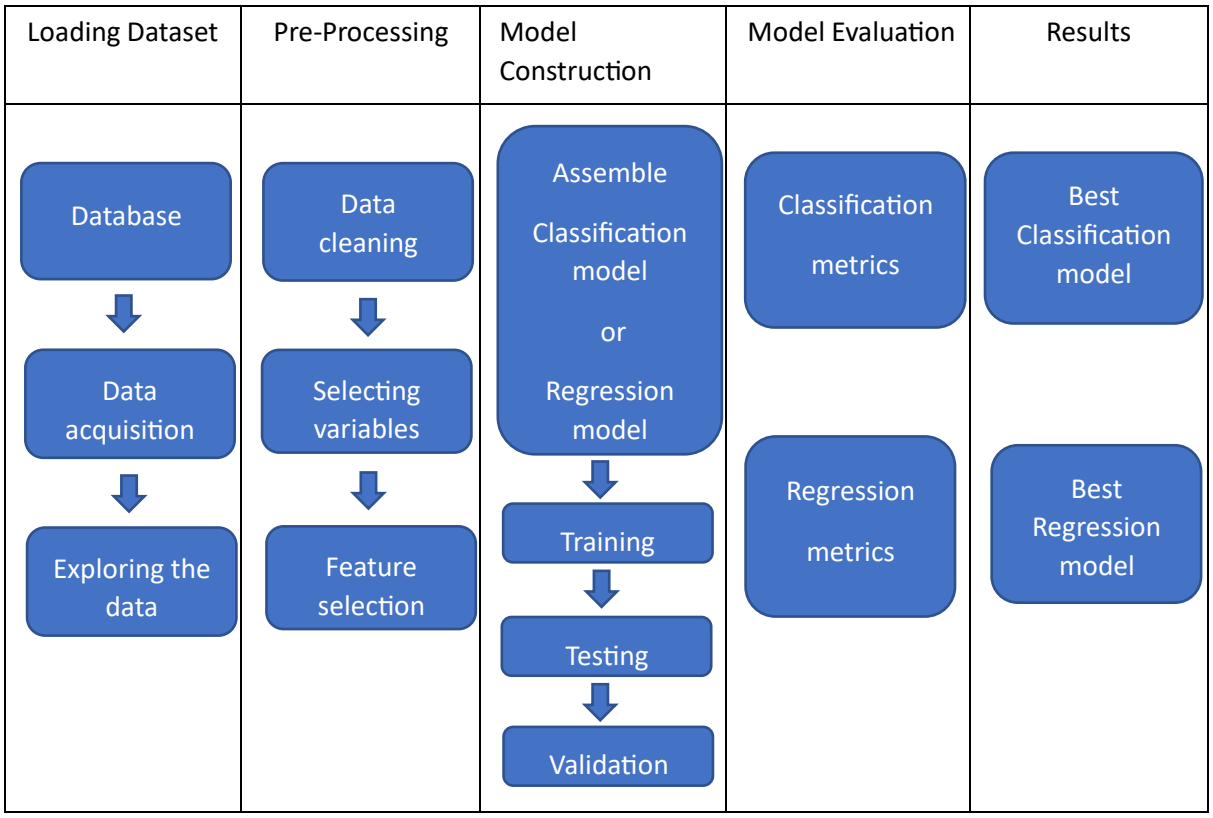

Figure 1: Process flow diagram of anti-cancer drug response prediction using machine learning algorithm.

1.3 Dysregulation signalling pathways lead to breast cancer.

Although breast cancer is more common in women, it can also affect men. Ageing, family history, hormonal variables, and lifestyle factors such as obesity and physical inactivity are among the top risk factors for breast cancer [11,12]. Normal human development is carefully controlled by complex signalling pathways that allow cells to communicate with each other and with their surroundings. Breast cancer develops because dysregulation of these signalling pathways allows cells to escape the mechanisms that control their survival and proliferation (Figure 2) [13]. Among the most important of these dysregulated pathways during cancer development is the DNA repair pathway: The mutations most frequently causing disruptions to this pathway are found in the genes BRCA1, BRCA2, TP53, and PTEN [14]. Alterations in these genes and other DNA repair pathway associated genes are frequently linked to sporadic

cancers and familial breast cancers, which respectively account for 85% and 5-10% of all instances of breast cancer [12,15]. Given the enormous degrees of genetic diversity displayed by breast tumour cells from different patients and the size and complexity of the available multi-omics cancer datasets, it is necessary to use machine learning to identifying pathways that, when dysregulated, promote oncogenesis. Understanding the many genetic and metabolic routes that culminate oncogenesis is particularly important because of its potential utility in devising new anti-cancer drugs.

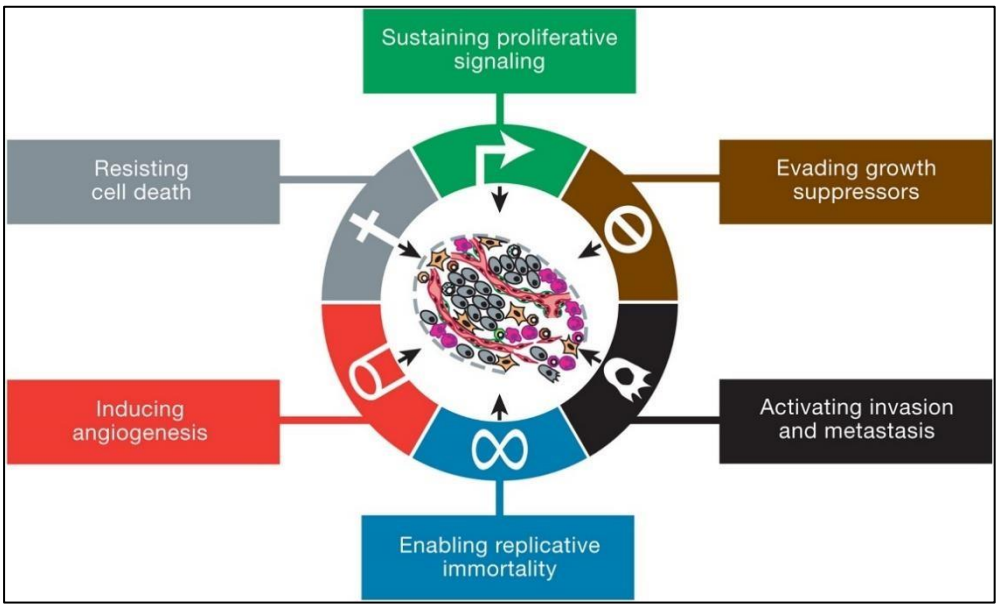

Figure 2: The formation and progression of cancer cells involve dynamic processes. Prior to drug treatment, cancer cells display diverse characteristics such as resistance to apoptosis, uncontrolled cell growth, evasion of growth inhibitory signals, and other abnormal cellular behaviours. These processes play a significant role in the initiation and persistence of cancer. Gaining insights into the underlying molecular mechanisms driving these cellular changes is essential for developing targeted therapeutic strategies.

1.4 Essential genes as anti-cancer drug targets

There are many sequencing projects where thousands of cancer cell genomes have been sequenced and released to the public via databases such as cBioPortal, Gene Expression Omnibus and National Cancer Institute (NCI) Genomic Data Commons [16–18]. The current

frontier in cancer research is to correctly grasp the functions of encoded genes so that we can mine actionable information from rapidly growing cancer molecular feature datasets. Essential genes (i.e., those genes that are necessary for the survival of cancer cells) are of particular interest in the context of cancer progression [19]. The Achilles project, among others, has performed high throughput gene knockdown and gene knockout experiments on immortalized cancer cell lines to identify potential drug targets [19]. Although we might ideally perform such experiments on the tumours of every cancer patient these knockdown and knockout experiments are simply too expensive and too labour-intensive. However, with this experimental data available for an array of different cancer cell lines, we can cheaply leverage and generalize this data using computational approaches such as machine learning to identify the essential genes of other unanalysed tumour cells, including those that have been freshly biopsied from patients. With a clear shortlist of the essential genes found in a patient's tumour cells we could productively use prior knowledge from gene knockdown and gene knockout experiments to reconstruct the relevant gene regulatory networks associated with oncogenesis in that patient and, using prior knowledge from drug susceptibility experiments, identify the best anticancer therapy(ies) to treat a patient's specific tumour.

1.5 Data and Machine Learning

Machine learning is a sub-field within the field of artificial intelligence aimed at training computer systems using past examples of situations/patterns of data (i.e., past "experiences") so that they can learn to accurately identify future versions of these same situations/patterns of data [20,21].Various machine learning algorithms can be used to train computer systems, such that the trained systems can then be set-up and quickly integrated into, for example, standard medical operational procedures to help patients and healthcare providers make diagnostic and/or treatment decisions [20].

Three categories of machine learning techniques exist: supervised learning, unsupervised learning, and reinforced learning [22]. This review focuses on supervised learning because it is the basis of most AI systems that are presently deployed in healthcare settings [23,24]. In precision medicine, supervised learning refers to methods in which a machine learning model is trained on a variety of variables (commonly molecular features such as gene transcript levels or amounts of translated protein but can also include, for example, the HIV status, age or sex and lifestyle choices of a patient) that are connected to a known outcome (commonly a diseased state). For instance, a model can be taught to connect a patient's characteristics (such as tumour size, form, and invasiveness) to a specific state (benign or malignant) [25]. After the model has been trained, it can be used to make predictions on a different dataset. Even though machine learning in healthcare is a very active research area, the overwhelming majority of health data that has so-far been gathered throughout the world has never been productively used to create predictive models that are useful in a clinical environment [26,27].

1.6 Feature selection for machine learning.

Careful selection of the tumour molecular features that are used to train a learning algorithm is commonly a key step for many machine learning approaches. Feature selection approaches (including dimensionality reduction approaches) have therefore become a rich area of research [28,29]. For example, in gene expression studies, feature selection assists in discovering the subset of genes with associated mutation, transcription and methylation on profile data that are the most informative predictors in classification models [30]. Feature selection helps to reduce the computational resources of any machine learning task by (a) reducing the amount of data that needs to be considered by the learning algorithm; (b) reducing the training time of the algorithm; and (c) enhancing the generalizability of trained models by reducing the possibility of overfitting models to the data [31].

The purpose of feature selection is to generate datasets that are compact and maximally informative [32]. A variety of different machine learning approaches can be used for feature selection including naive bayes, random forests, k-means clustering, hierarchical clustering, and density-based clustering. For instance, to predict breast cancer recurrence, Jerez-Aragones et al. used a decision tree model to find key features [33]. Although using feature selection to reduce the dimensionality of the dataset can be challenging, it can be crucial in that it enables the training of machine learning models that are simple enough, both for humans to comprehend, and to yield some insights into the actual cellular processes that underly the associations between molecular features and disease states: information that can be crucial for discovering new therapies [29].

Three feature selection techniques exist: filter-based techniques, embedding techniques, and wrapper techniques [30,32]. Filter-based methods are generally favoured for gene expression data because wrapper-based and embedded methods combine feature selection and the learning process to select optimal subsets of features: a scheme that typically requires the use of computationally expensive nested cross-validation and is prone to model overfitting [30]. Therefore, using a filter-based method such as Minimum Redundancy Maximum Relevance (MRMR) is a better option when working with high-dimensionality datasets such as those yielded by gene expression experiments [34].

1.7 Data resources for machine learning-based drug response predictions.

Machine learning algorithms require benchmarked input datasets for training. Furthermore, the learning algorithms require many training data instances to both achieve highly accurate predictions and generalise well to new examples [35]. However, larger training datasets require more computation to train learning algorithms [36]. In the context of using machine learning to develop drug response prediction algorithms, the development of high-throughput molecular

biological data generation technologies now allows us to examine the sequences, and expression levels of thousands of genes that may be associated with the responses of tumours to drugs [37].

Using these high-throughput technologies, various projects have measured the drug response profiles of hundreds of cancer cell lines derived from different tissues to hundreds of anticancer drugs. These projects include the Cancer Cell Line Encyclopaedia (CCLE) [7], the Genomics of Drug Sensitivity (GDSC) [10] project, the Library of Integrated Network-Based Cellular Signatures (LINCS) [38] project, the Cancer Therapeutic Response Portal (CTRP) [39], and the Cancer Target Discovery and Development (CTD^2) [40] project: all of which provide databases that are freely accessible.

Besides drug response profiles, some of these databases also contain information on molecular profiles of cancer cell lines, including the mutation profiles of the cell lines, their gene expression signatures, and their DNA methylation profiles. Since the sensitivity of tumours to anticancer drugs (i.e., their chemosensitivity) is determined by the molecular phenotypes of the tumours, the molecular profiles of cell lines can be used to identify biomarkers that may be useful for predicting which anti-cancer drugs would be most effective in the treatment of specific cancer cell lines [41].

Whereas the CCLE and GDSC databases provide molecular profiles of drug-untreated cell lines, the LINCS project database provides proteomics and transcriptional profiles of the same cell lines measured pre- and post-drug or genetic perturbation [42].

Another significant resource is the Achilles project (DepMap: The Cancer Dependency Map Project at Broad Institute), which applies genome-scale RNAi and CRISPR-Cas9 genetic perturbations to silence or knockout individual genes and, across hundreds of cancer cell lines, identifies the genes that are most strongly associated with cellular fitness.

This abundance of multi-omics data has led to the development of various models to predict drug responses and identify drug targets in different cancer cell lines. For example, Dong et al. [43] constructed an SVM model to predict how various cancer cell lines would respond to several anticancer drugs. Gene expression datasets from the CCLE database were used to train the model which achieved an accuracy of more than 80%. These studies have revealed that gene expression data is particularly useful for predicting drug responses in different cell lines [6,44]. Datasets from these various resources (e.g., the Achille cell fitness and GDSC dose response datasets), can be integrated to test, for example, how knocking out or reducing the expression of a downregulated gene in each cancer cell line might influence both the chemosensitivity and the fitness of the cell line. Therefore, there is great potential for applying machine learning approaches with these currently available datasets to transform the precision of cancer treatment.

1.8 Machine learning methods for predicting drug responses.

Identifying causal relationships between the quantifiable molecular features of cell lines and the responses of these cell lines to drugs establishes the groundwork for predicting the pharmacological responses of any given patient tumour based on the quantifiable molecular features of the tumour [45]. In this regard several large-scale studies have been done on how an array of anticancer drugs interact with proteins expressed in cancer cells [46] and these have been followed-up by additional studies applying supervised machine learning approaches to predict the chemosensitivity of cancer cells to various anticancer drugs.

SVM models have been applied in many instances to accurately predict the responses of cancer cells to drug perturbation. For example, Stetson et al., trained a SVM to identify multi-omics correlates of anti-cancer therapeutic responses [47]. Another widely applied class of machine learning models used for predicting diseases and chemosensitivity of cancer cells is the simple

logistic regression. For example, logistic regression has been applied to determine the course of treatment and prognosis for breast cancer [48]. Other methods broadly and more recently applied in supervised learning tasks include ANNs and DNNs. For instance, Burke et al. applied an ANN to predict the survival of breast cancer patients and achieved 77% accuracy for the five-year survival and 73 % accuracy for the ten-year survival of patients [49].

To predict the chemosensitivity of breast cancer cells, it is important to consider employing multiple models. Additionally, exploring machine learning interpretability with techniques like Local Interpretable Model-agnostic Explanations (LIME) [50], SHapley Additive exPlanations (SHAP) [51] and Explain Like I'm 5 (Eli5) [52] is important to explain the predictions.

This approach is known as Explainable Artificial Intelligence. The main idea is that a model should provide justification for any predictions or recommendations that it makes. These explanations, which might be meaningless artifacts within the input data, will enable end users, like doctors, to assess a model before acting on its predictions or recommendations [23]. The confidence of healthcare professionals in AI-enabled disease detection could diminish if the algorithm's predictions lack explainability [53].

1.10 Evaluating the model.

Inaccurate reporting of a model's true performance can occur. This issue is evident when researchers choose to evaluate their models using only a single performance metric. For instance, Wang et al. [54] assessed a gene selection technique for micro-array cancer classification based solely on accuracy. They employed four traditional machine learning models for this classification problem, with the best-performing model achieving an accuracy of 88.71%. However, relying solely on accuracy to evaluate this method poses a problem, as accuracy does not perform well in the case of imbalanced datasets. Conversely, the F1-score performs well in both balanced and imbalanced data, and since most cancer research data is

imbalanced, Wang et al. [54] may have produced misleading results. Therefore, it is advisable to employ multiple metrics to comprehensively evaluate a model's performance.

1.10.1 Evaluation of a classification problem

In classification problems, the aim is to make predictions based on classes (e.g., benign/malignant for a cancer biopsy or sensitivity/resistance of a tumour to a drug). The performance of a trained classifier can be scored using metrics such as (1) accuracy, (2) precision, (3) recall, (4) F1 score,

(5) ROC, and (6) AUC. Most of these metrics can be calculated from the entries of a Confusion Matrix: a 2x2 table of values indicating the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications. (Figure 3).

		Actual	
		Positive	Negative
Predicted	Positive	*TP*	*FP*
	Negative	*FN*	*TN*

Table 1: The confusion matrix of a binary classification problem. A test result is a TP when it accurately predicts the presence of a condition and that condition turns out to be present, and TN when it accurately predicts the absence of a condition, and that condition turns out to be absent. When a test result predicts a condition's absence when the condition is present, it is known as FP, and when it predicts a condition's absence when it is present, it is known as FN [55].

TP, FP, TN, and FN are the four potential outcomes of any prediction. For example, if it is predicted that a cancer cell line is sensitive to an anti-cancer drug, a TP is achieved if the prediction is correct, and a FP is noted if the prediction is incorrect. If it is predicted that the

cell line is insensitive to an anti-cancer drug when in fact the cell line is in fact insensitive to the drug, then a TN is achieved whereas if the cell line is in fact sensitive to the drug, then a FN is noted.

The recall metric tells us what proportion of the actual positives tested (i.e., TP+FN) were TPs. (equation 1).

$$Recall = \frac{TP}{TP + FN} \qquad (1)$$

The precision metric indicates the proportion of all the positives that were identified (i.e., TP+FP) were TPs (equation 2).

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

The recall and precision metrics are used together to calculate the F1 score: a metric used to gauge a classifier's effectiveness (equation 3). A value closer to 1 denotes a classifier performs well, whereas a value farther from 1 denotes a poorly performing classifier.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (3)$$

The accuracy metric is the proportion of all predictions that were correct (i.e., TP+ TN; equation 4).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (4)$$

Rather than being metrics *per se*, Receiver Operating Curve and Area Under Curve are graphs of sensitivity vs specificity that indicate how well a classifier performs (Figure 4). The sensitivity and the specificity respectively represent the genuine positivity and negative rates.

The sensitivity and specificity change as we move up the graph. The ROC includes the AUC. It reveals how effectively the classifier works. The closer to 1 the AUC, the better the classifier as an AUC of 1 denotes a flawless classifier.

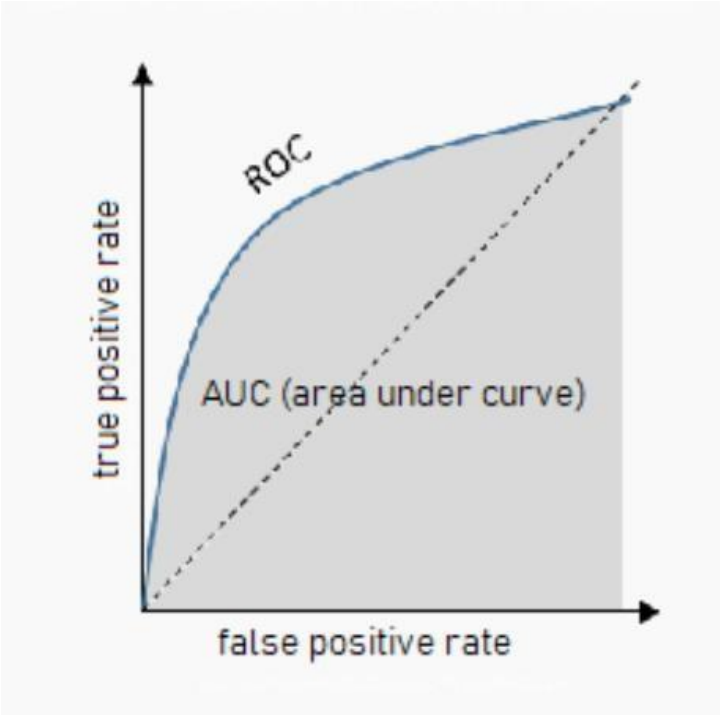

Figure 3: Illustrates the ROC curve, which is a graphical representation of the model's probability performance. The AUC value in the ROC curve indicates the level of separability achieved by the model. A higher AUC indicates a better ability of the model to differentiate between cancer cells that are sensitive to a drug and those that are resistant to the drug *[56]*.

1.10.2 Evaluation of regression problems.

In regression problems, we work with real-value data, for example predicting the concentration required for inhibition of proteins that cause cancer (such as the HER2 protein which controls cancer growth). The most used metrics for evaluating regression models are the Root Mean Square

Error (RSME), Mean Absolute Error (MAE), Coefficient of Determination (R^2) and Mean Square Error (MSE). Generally, n indicates the number of samples in a dataset, $y_i$indicates the actual values for each sample, $\bar{yi}$ indicates the predicted values of each sample and $\hat{y}_i$ indicates the mean values of the samples.

The closer the RSME value is to 0.0, the better the correlation between predicted and observed values (equation 5).

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (5)$$

The MAE value indicates the size of the errors between predicted and observed values (equation 6).

$$MAE = \frac{1}{n}\sum_{1=1}^{n}(y_i - \bar{y}_i) \qquad (6)$$

The MSE value indicates the amount of error in the statistical model. The larger MSE is, the greater the error. An MSE of 0.0 indicates a perfect model (equation 7).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \qquad (7)$$

The R^2 value tells us the percentage of the outcome's variance that may be accounted for by the predictor variables. The closer the R^2 value is to 1, the better the correlation between observed and predicted values (equation 8).

$$R^2 = 1 - \frac{\sum(y_i - \bar{y}_i)}{\sum(y_i - \hat{y}_i)} \qquad (8)$$

Different approaches have been proposed for investigating drug sensitivity using regression models. For instance, Pouryahya et al. [57] used a random forest regression model to accurately predict the drug sensitivity of pancreatic cancer cells using the IC50 values from the GDSC database, achieving an R^2 of 0.79.

1.10 .3 Conclusion

Machine learning has demonstrated success in breast cancer by predicting treatment response and providing personalized treatment recommendations. By analysing diverse patient data, including genomic profiles and clinical information, machine learning models can predict how individual patients will respond to specific treatments. This allows for the identification of patients who are likely to benefit from certain therapies, leading to more targeted and effective treatment strategies.

Machine learning models are used mostly in health research settings and minimally in clinical settings. However, with model interpretability approaches doctors can makes use of these models and speed up their decision making, this will make a more efficient personalised treatment plan or selection.

1.11 Rationale of Study

In the field of oncology, predicting drug responses in cancer cell lines will be paramount for the progression of personalised medicine. The study will leverage datasets from the GDSC and the Achilles project, focusing primarily on breast cancer. Utilising machine learning models, the research will aim to predict drug sensitivities, thereby laying the groundwork for tailored treatment strategies. The models will be trained using gene knocked-out data, to predict drug sensitivity in untested cell lines and identify potential gene targets for anti-cancer therapy. The

classification analysis will further differentiate between drug-sensitive and drug-resistant cancer cell lines, offering a nuanced understanding of drug responses. This approach will not only promise enhanced patient outcomes but also provide a deeper insight into the molecular mechanisms governing drug sensitivities in breast cancer.

1.12 Aims and Objectives

1.12.1 Aim

To employ machine learning techniques in predicting drug responses in breast cancer cell lines and to identify genes that can be targeted during anti-cancer therapy based on their gene expression and CRISPR -mediated gene knockout datasets.

1.12.2 Objectives

1. To curate and integrate datasets from the GDSC and Achilles project with a focus on breast cancer.
2. To establish machine learning models that can predict drug sensitivities using gene knocked-out data.
3. To evaluate the performance of the established models using robust metrics.
4. To delve into bioinformatics analysis, using pathway enrichment and kinase enrichment analysis, to gain deeper insights into the molecular mechanisms governing drug sensitivities.

1.13 Statement on ethics

The study protocol was approved by the University of Cape Town, Health Science Research Ethics Committee IRB00001938. The publicly available datasets were collected by the LINCS, GDSC and Achilles projects and made available through their respective databases.

2. Predicting cancer cell line drug responses and mining of essential genes that promote oncogenesis.

2.1 Introduction

The aim of the analyses performed in this chapter was to predict the sensitivity of cancer cell lines to a drug and to investigate the effects of drug treatment on the genes that promote oncogenesis when knocked out. To achieve this, a combination of machine learning techniques, including KNN imputation for missing data and the MRMR algorithm for feature selection were used.

In addition to these techniques, various machine learning models were trained to predict drug sensitivity and examine the link between various anti-cancer drugs and CRISPR-derived cell fitness scores. These models were trained using CRISPR-mediated gene knockout and drug sensitivity data from a particular group of cancer cell lines, enabling us to predict drug sensitivity for untested cell lines and identify the genes that could be targeted during anti-cancer therapy.

2.2 Materials and Methods

2.2.1 Dataset collection

Three datasets were used, including the dose-response data of the cancer cell lines from GDSC, and cell line and samples annotation information and CRISPR-derived cell fitness scores from the Cancer Dependence Map Database as profiled by the Project Achilles. To analyse the response of cancer cell lines to anti-cancer drugs, we used the GDSC, which is a comprehensive database that provides a detailed mapping between anti-cancer drug sensitivities and signalling pathways across an extensive collection of over one hundred distinct cancer cell lines. This resource offers invaluable insights into the molecular determinants of drug responses, aiding

in the understanding of how different cancers react to various therapeutic agents. The data was downloaded from the Sanger Institute website ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC2_fitted_dose_res ponse_25Feb20.xlsx and ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC1_fitted_dose_res ponse_25Feb20.xlsx. The drug responses were measured using fitted IC50 values, which show the drug concentration needed to induce a 50% reduction in cell proliferation. To obtain these datasets, the *wget* function was run on facilities provided by the UCT HPC team (http://hpc.uct.ac.za) and the files were transferred from the remote computer to a local computer using the MobaXterm application (v23.0).

To evaluate the impact of CRISPR-mediated gene knockouts on the viability of cancer cell lines, the Achilles dataset was used (provided by the Broad Institute). The Achilles dataset for CRISPR based gene perturbations includes 17386 genes that have been systematically knocked out (one gene only for each experiment), across more than one thousand cell lines cultured from 31 primary diseases. A Sample Information 22Q2 dataset included 1840 cell lines cultured from 33 primary diseases. These datasets were downloaded from https://depmap.org/portal/download/all. By integrating these datasets, the aim was to gain insights into the relationship between CRISPR derived gene dependencies (or gene essentiality) and the chemosensitivity of cancer cells: an endeavour that could ultimately contribute to the advancement of personalized treatment approaches and the improvement of patient care in breast cancer.

2.2.2 GDSC data preprocessing

For GDSC data preprocessing, the BRCA dataset for drug responses was accessed from the GDSC database, using both the GDSC1 and GDSC2 datasets. The datasets were converted from *.xlsx* format into *.csv* format. The *tidyverse* (v2.0.0) [58] package, which is embedded with other packages useful in the data wrangling process, was used to read the datasets. The two datasets were merged into one, and duplicates were removed. If a cancer cell line treated with a drug was present in both GDSC1 and GDSC2, the reading in GDSC1 was deleted and the reading in GDSC2 retained because GDSC2 has an improved drug screening procedure and assay compared to GDSC1 [59].

To enable classification, the cancer cell lines in the merged GDSC dataset were split into two groups: (1) those responsive to the drug, and (2) those resistant to the drug. This was done by discretizing the IC50 z-scores into three classes: (1) sensitive, (2) intermediate, and (3) resistant, and the intermediate class was deleted (Figure 5). To investigate the effect of different IC50 z-score thresholds on the performance of the model, the following thresholds were used: [-0.1,0.1], [-0.2,0.2], [-0.3,0.3], [-0.4,0.4], [-0.5,0.5], [-0.6,0.6], [-0.7,07], [-0.8,08], and [-0.9,09]. According to the GDSC documentation (https://www.cancerrxgene.org/help) , a lower IC50 than the threshold value indicates high sensitivity to the drug, while a high IC50 than the threshold value indicates resistance. For regression analysis, the same process was applied, but the continuous IC50 values were used.

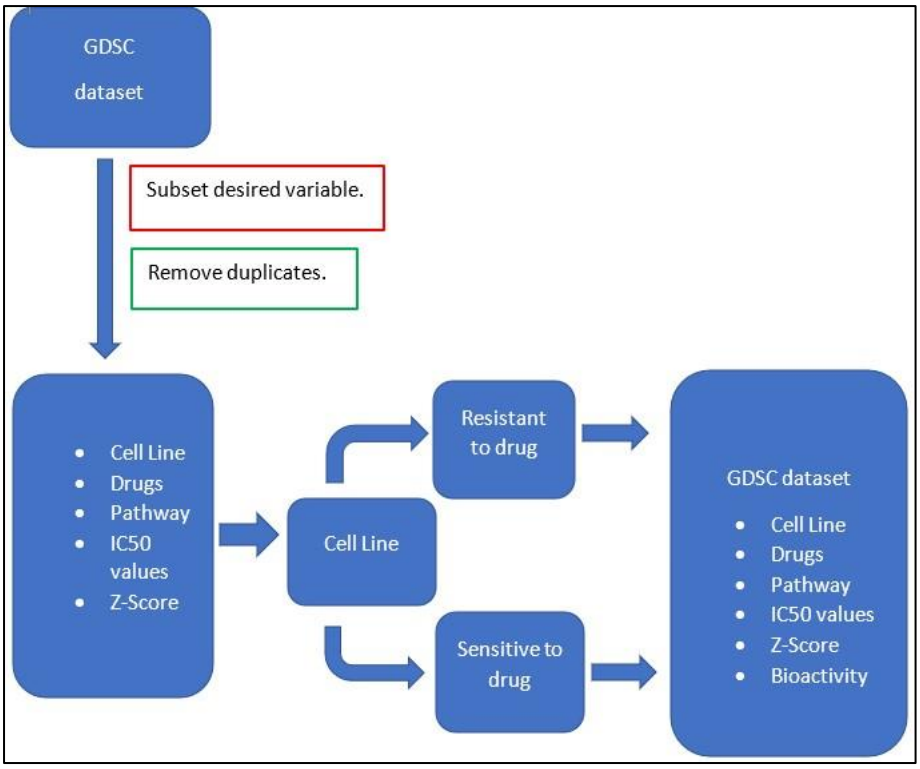

Figure 4: Data preprocessing for classification analysis. The cell line, drugs, pathways, IC50 value and the normalized IC50 values (Z-Scores) were subset from the dataset. The cell lines were separated into those sensitive to the drug and those resistant to the drug according to GDSC standards.

2.2.3 Achilles data preprocessing

To analyse the effect of gene knockdown or knockout on cell line fitness the CRISPR-derived cell fitness data and the sample information were loaded and merged to create one dataset. However, the dataset contained missing values, so rows and columns with more than 50% missing values were removed. KNN imputation was then used (implemented in the *caret* (v1.3.4) [60] and *RANN* (v2.1.3) [61] packages) to predict missing values for rows and columns that had less than 50% missing values. The KNN algorithm takes missing values in a dataset and finds the *k* closest samples based on the values of other similar instances in the dataset. This method works well when the number of missing values is small.

The genes with a correlation of over 70% were removed. This is important for several reasons: (1) it increases efficiency by reducing the computational time needed to train the model; (2) it reduces the risk of overfitting and improves the model's ability to generalize; and (3) it improves the interpretation of the model, since highly correlated genes can make it difficult to understand how each gene contributes to the outcome. Next, the gene effect dataset was merged with the GDSC drug sensitivity dataset using a common cell line column, creating the GDSC_Achilles dataset (Figure 6).

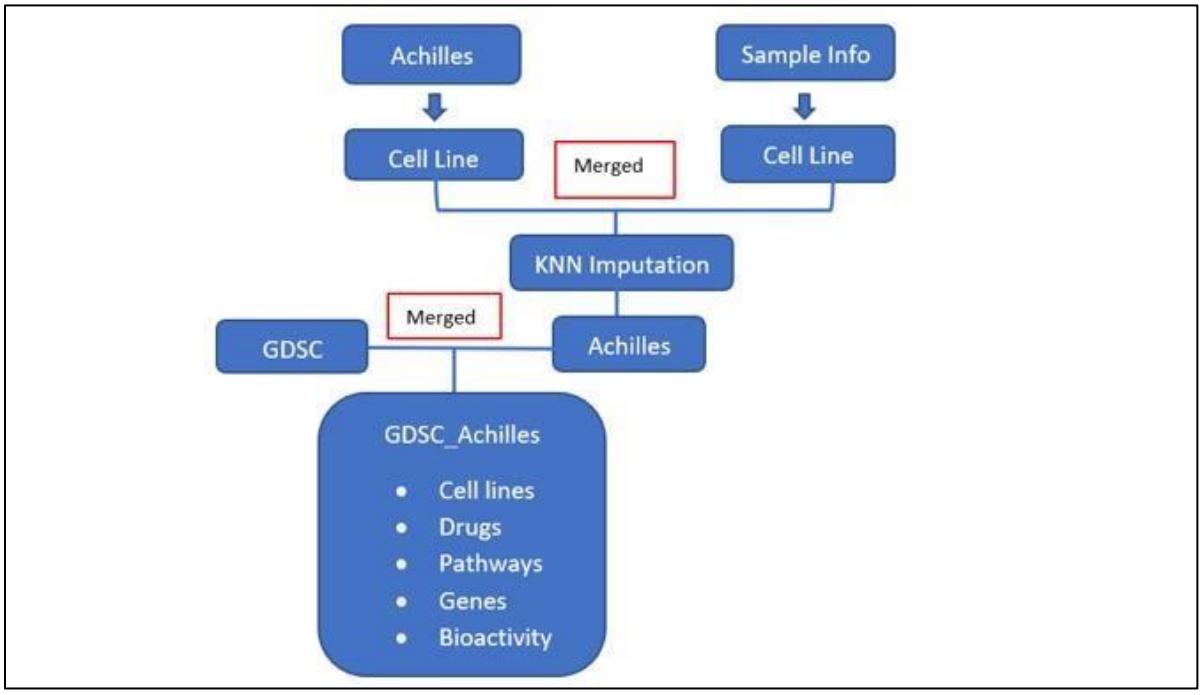

Figure 5: Data preprocessing of the Achilles dataset. The sample information and the gene effect dataset are merged using the cell lines. KNN imputation was used to handle the missing values and preserve the patterns in the data. The Achilles and GDSC dataset were merged using common columns.

2.2.4. Feature Selection

MATLAB-R2022b was employed to conduct feature selection on the merged GDSC_Achilles dataset. In the classification analysis, the *fscmrmr* function was utilized, while the *fsrmrmr* function was employed for the regression analysis. These functions employ the minimum

redundancy maximum relevancy (MRMR) [62] algorithm to determine the rankings of predictors. As a result, the algorithm generates indexes (idx) that contain the ordered indices of predictors based on their levels of significance. For classification problems, the idx identifies the most relevant predictors and arranges them according to their relevance. The top 5000 features (CRISPR-derived fitness scores) were selected.

2.2.5. Model establishment.

Classification analysis was conducted in R (v4.1.3) [63] and R (v4.2.0) [63]. Because of memory issues with R, regression analyses were conducted using python/miniconda3-py39-usr [64]. Both the R scripts and python scripts were run on the UCT HPC.

2.2.5.1 Model establishment for classification analysis

Each dataset was pre-processed by normalising the data. The data was split into 70% training and 30% testing sets using the *createDataPartition* function from *caret* (v1.3.4) [60]. The character variables in both training and testing sets were converted into factors using *as.data.frame(unclass())*. Four machine learning models, K-Nearest Neighbour (KNN), Support Vector Machine Radial (svmRadial), Gradient Boosted Machine (GBM), and Random Forest (RF), were trained on the training set using the *train* function of *caret* (v1.3.4) [60] with 10-fold cross-validation. The models were evaluated using several classification metrics (including the AUC, accuracy, precision, recall, and F1-score), and the results for each model were appended to the *results_table* data frame. Finally, the results were programmatically written to a csv file.

2.2.5.2 Model establishment for regression analysis

For the regression analysis, Pandas [65] and NumPy [66] were imported to manipulate and handle data. The *scikit-learn* [67] library was used. The regression models ElasticNet, KNeighborsRegressor, RandomForestRegressor, GradientBoostingRegressor, and SVR were

imported and regression evaluation metrics such as *mean_squared_error*, *r2_score*, and *mean_absolute_error* were imported.

The datasets were read using pandas' *read_csv()* function. The datasets were pre-processed before applying regression models by normalising the numerical columns of the datasets using the *sklearn.preprocessing.scale* function. The *pd.get_dummies* function was used to convert categorical variables to numerical ones.

The datasets were split into training and testing sets using the *np.random.rand* function, where 70% of the data was used for training and the remaining 30% was used for testing. Five different regression models were created using the sklearn library: ElasticNet, KNeighborsRegressor, RandomForestRegressor, GradientBoostingRegressor, and SVR. Each model was trained on the training set and evaluated on the testing set using Root Mean Square Error (RSME), R-square (R^2), and Mean Absolute Error (MAE) metrics. The results were appended to the *results_table_reg* data frame. Finally, the *results_table_reg* data frame was saved as a csv file using the *to_csv()* function.

2.2.6 Model interpretability

Identifying the variables (genes) that significantly influenced the model's accuracy was the area of interest. To achieve this, the *varImp* function from the *caret* (v1.3.4) [60] package was used to calculate variable importance scores using the permutation feature importance method. This method shuffles the values of each feature one at a time and measures how much this affects the model's accuracy. If shuffling a feature causes a decrease in model accuracy, it is considered an important feature. Conversely, if shuffling a feature has no effect on model accuracy, the feature is deemed "unimportant" [68]. The *varImp ()* function selects the important features.

2.2.7. Results and Discussion

2.2.7.1Classification analysis

Classification analysis was done to identify the genes that are associated with drug-sensitive cell lines or drug-resistant cell-lines. The IC50 z-scores obtained from the GDSC database were discretized to classify the drug responses of cancer cell lines as either drug-sensitive or drug resistant. To assess the models' robustness and their ability to differentiate between drug-sensitive and drug-resistant cell lines, nine datasets were created using different IC50 z-score thresholds: [0.1, 0.1], [-0.2, 0.2], [-0.3, 0.3], [-0.4, 0.4], [-0.5, 0.5], [-0.6, 0.6], [-0.7, 0.7], [-0.8, 0.8], and [-0.9, 0.9]. Four classification models were trained (KNN, svmRadial, GBM, and RF), and their performance was evaluated using four metrics (Tables 2-10) (Figures 6-14).

In this analysis, the focus was on the F1-score, which provides a balanced measure of performance by combining precision and recall. A higher F1-score indicates better overall performance in correctly identifying sensitive or resistant cancer cell lines with a low false positivity rate. All the models exhibited a high F1-score at the [-0.9, 0.9] threshold, with the random forest model achieving the best performance (F1-score of 0.86), followed by the GBM (0.84), svmRadial (0.82), and KNN (0.83). This trend was consistent across different scenarios. The high F1-score at the IC50 threshold of [-0.9, 0.9] suggests that the models performed well in identifying cancer cells that were either highly sensitive or highly resistant under extreme conditions, particularly the random forest model. As the IC50 thresholds decreased, there was a noticeable decline in the models' performance.

Table 2: Performance of the classification models at the IC50 threshold [-0.1,0.1].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.663573	0.70155	0.726908	0.714004
SVM	0.651972	0.667797	0.791165	0.724265
GBM	0.707657	0.730337	0.783133	0.755814
RF	0.761021	0.778626	0.819277	0.798434

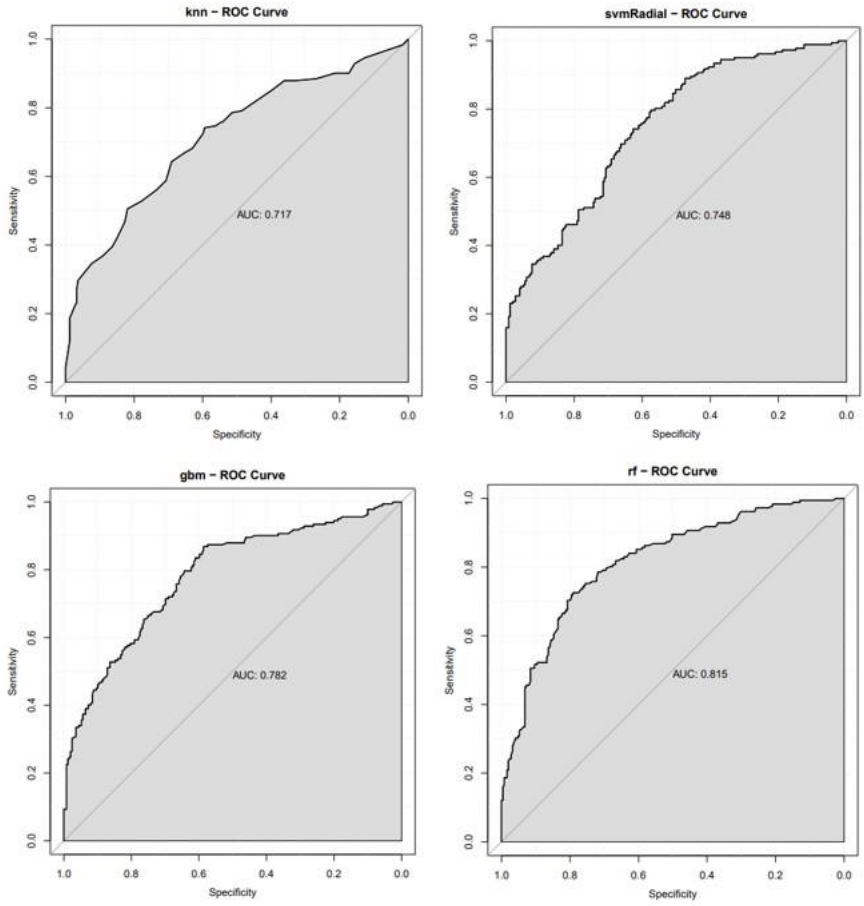

Figure 6: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.1, 0.1]. This shows sensitivity and specificity. The area under the curve of the KNN (0.717), svmRadial (0.748), GBM (0.782) and the RF (0.815) all show a moderate predictive power.

Table 3: Performance of the classification models at the IC50 threshold [-0.2,0.2]

Models	Accuracy	Precision	Recall	F1-score
KNN	0.6618	0.69112	0.752101	0.720322
SVM	0.671533	0.678201	0.823529	0.743833
GBM	0.715328	0.72	0.831933	0.77193
RF	0.705596	0.720755	0.802521	0.759443

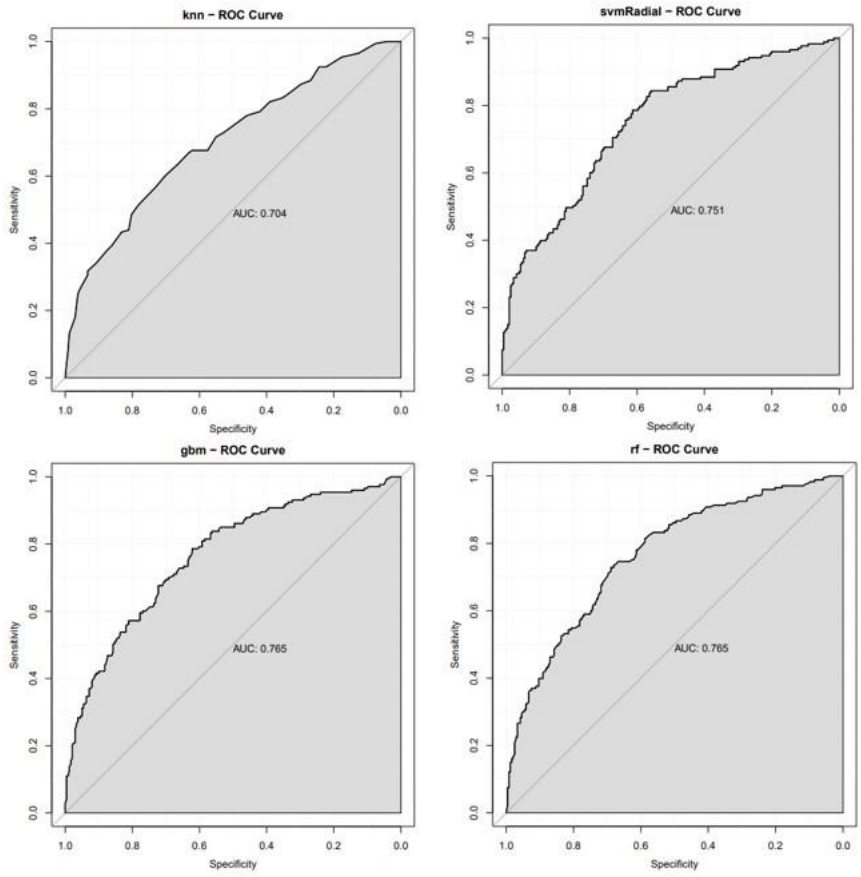

Figure 7: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.2, 0.2]. This shows sensitivity and specificity. The area under the curve of the KNN (0.704), svmRadial (0.751), GBM (0.765) and the RF (0.765) all show a moderate predictive power.

Table 4: Performance of the classification models at the IC50 threshold [-0.3,0.3].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.676166	0.695652	0.785714	0.785714
SVM	0.663212	0.666667	0.839286	0.839286
GBM	0.696891	0.716599	0.790179	0.790179
RF	0.738342	0.741176	0.741176	0.84375

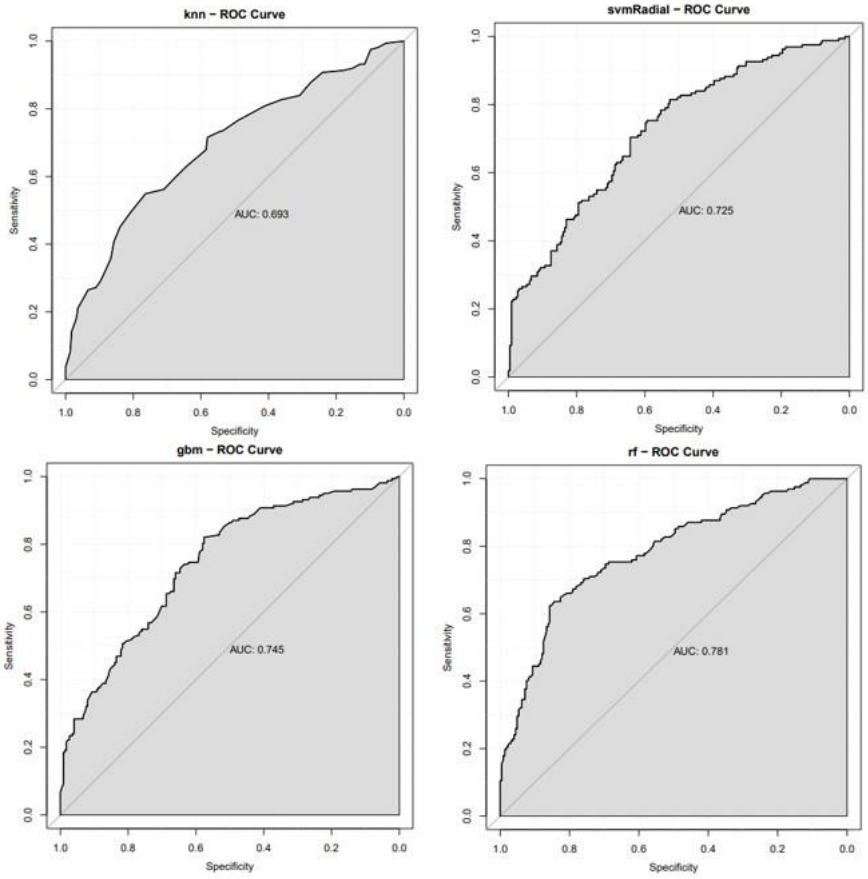

Figure 8: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.3, 0.3]. This shows sensitivity and specificity. The area under the curve of the KNN (0.693), svmRadial (0.725), GBM (0.745) and the RF (0.781) all show a moderate predictive power.

Table 5: Performance of the classification models at the IC50 threshold [-0.4,0.4].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.670061	0.701068	0.716364	0.708633
SVM	0.625255	0.650165	0.716364	0.681661
GBM	0.688391	0.701987	0.770909	0.734835
RF	0.702648	0.717172	0.774545	0.744755

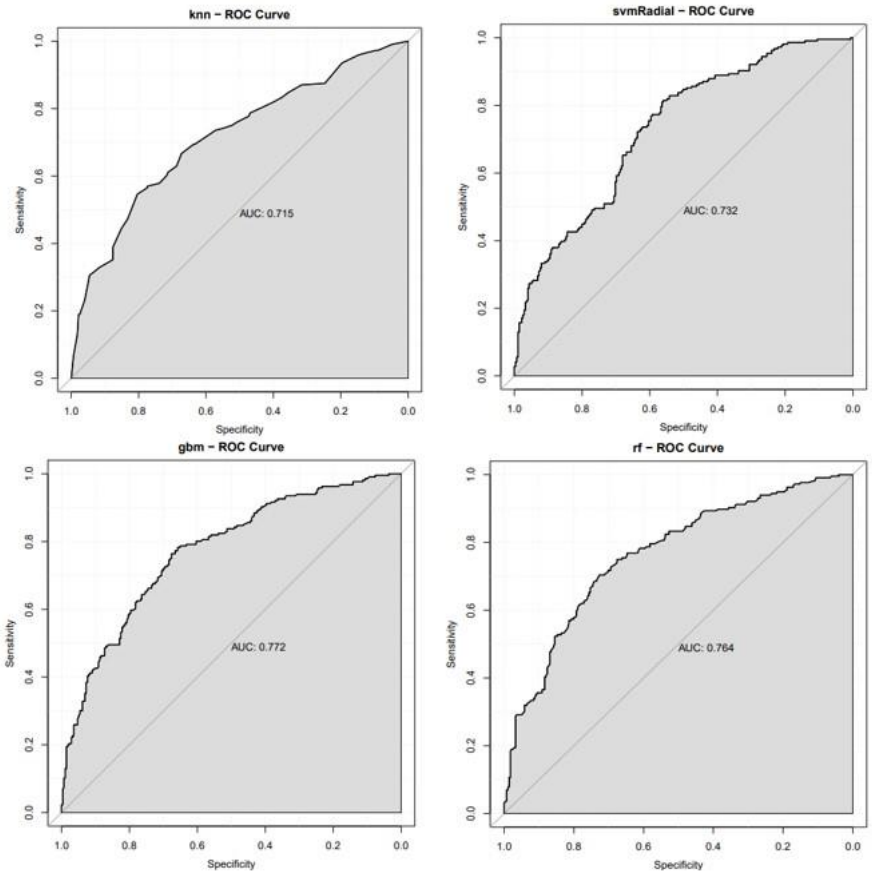

Figure 9: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.4, 0.4]. This shows sensitivity and specificity. The area under the curve of the KNN (0.715), svmRadial (0.732), GBM (0.772) and the RF (0.764) all show a moderate predictive power.

Table 6: Performance of the classification models at IC50 threshold [-0.5,0.5].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.662252	0.687285	0.763359	0.723327
SVM	0.649007	0.648415	0.858779	0.738916
GBM	0.706402	0.729537	0.782443	0.755064
RF	0.735099	0.761029	0.790076	0.775281

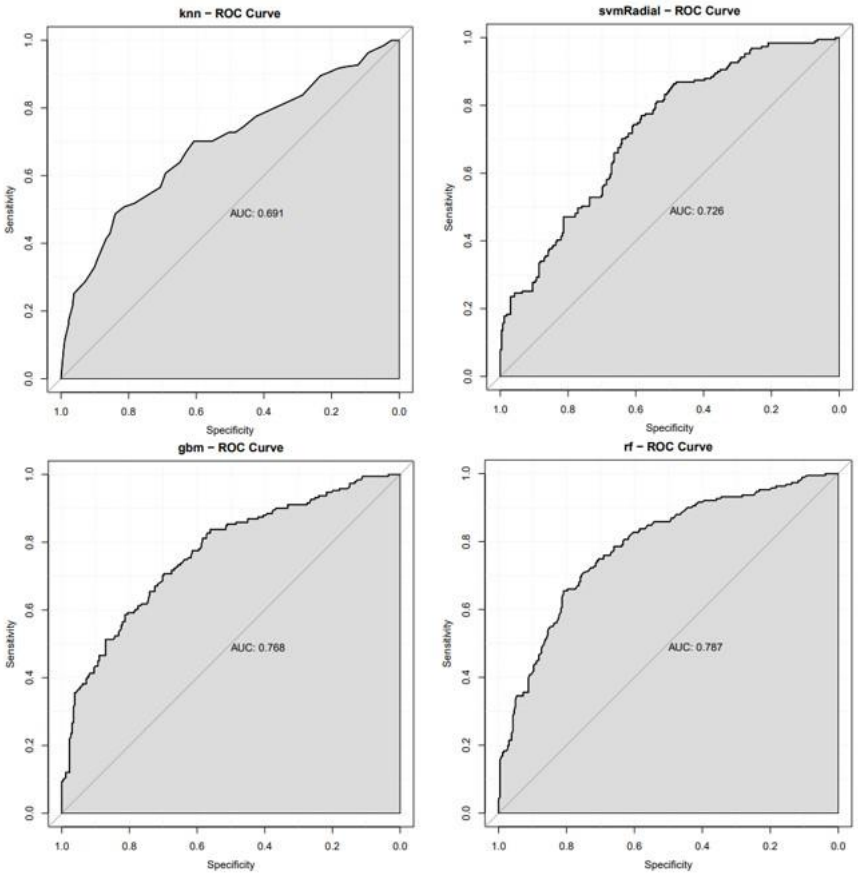

Figure 10: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.5, 0.5]. This shows sensitivity and specificity. The area under the curve of the KNN (0.691), svmRadial (0.726), GBM (0.768) and the RF (0.787) all show a moderate predictive power.

Table 7: Performance of the classification models at IC50 threshold of [-0.6,0.6].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.651442	0.709163	0.712	0.710579
SVM	0.629808	0.681818	0.72	0.700389
GBM	0.680288	0.703833	0.808	0.752328
RF	0.694712	0.73384	0.772	0.752437

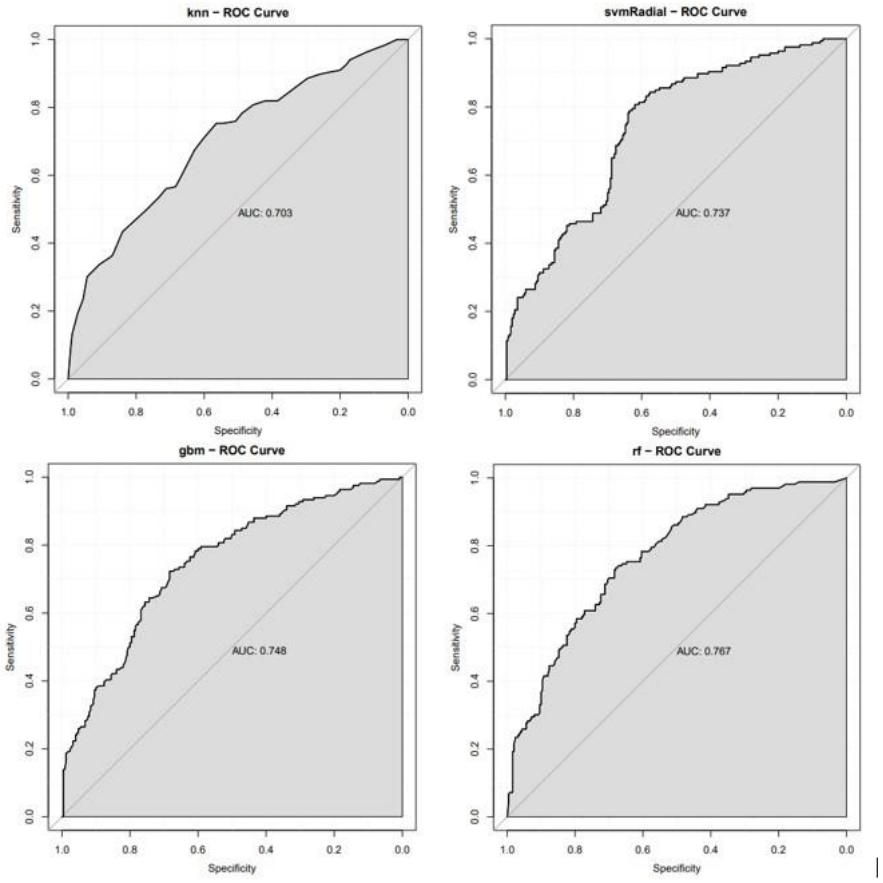

Figure 11: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.6, 0.6]. This shows sensitivity and specificity. The area under the curve of the KNN (0.703), svmRadial (0.737), GBM (0.748) and RF (0.767) all show a moderate predictive power.

Table 8: Performance of the classification models at IC50 threshold [-0.7,0.7].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.707124	0.724638	0.851064	0.782779
SVM	0.701847	0.716312	0.859574	0.781431
GBM	0.751979	0.758242	0.880851	0.814961
RF	0.759894	0.770677	0.87234	0.818363

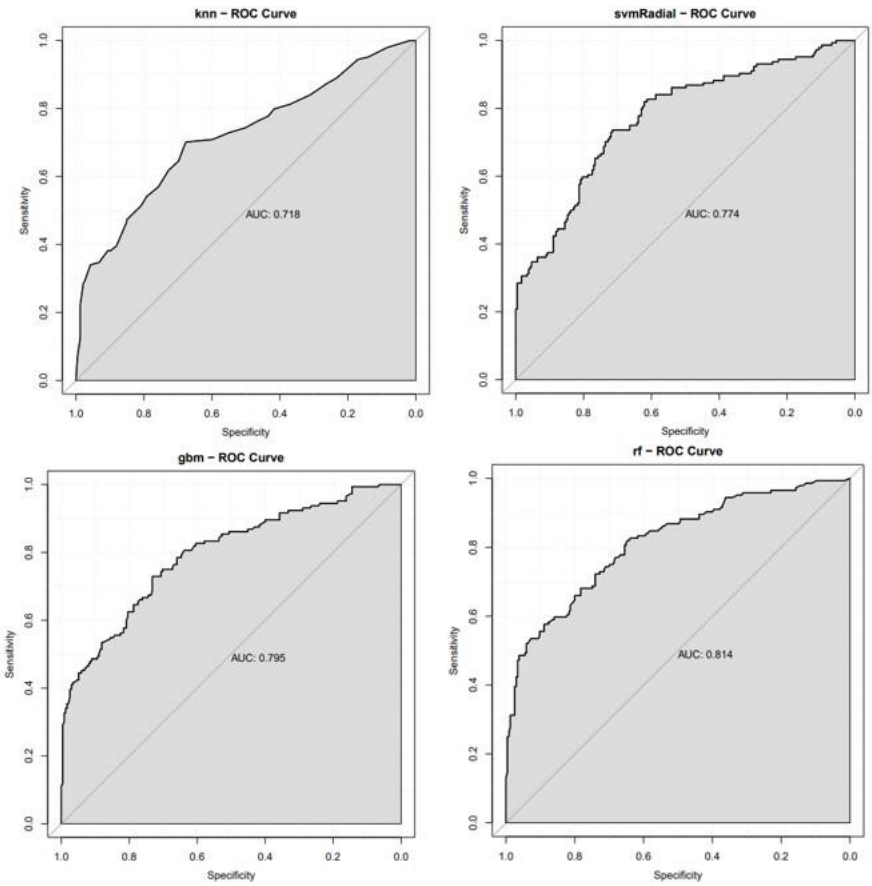

Figure 12: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.7, 0.7]. This shows sensitivity and specificity. The area under the curve of the KNN (0.718), svmRadial (0.774), GBM (0.795) and the RF (0.814) all show a moderate predictive power.

Table 9: Performance of the classification models at IC50 threshold [-0.8,0.8].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.732938	0.782805	0.804651	0.793578
SVM	0.724036	0.769912	0.809302	0.789116
GBM	0.756677	0.792952	0.837209	0.81448
RF	0.750742	0.786026	0.837209	0.810811

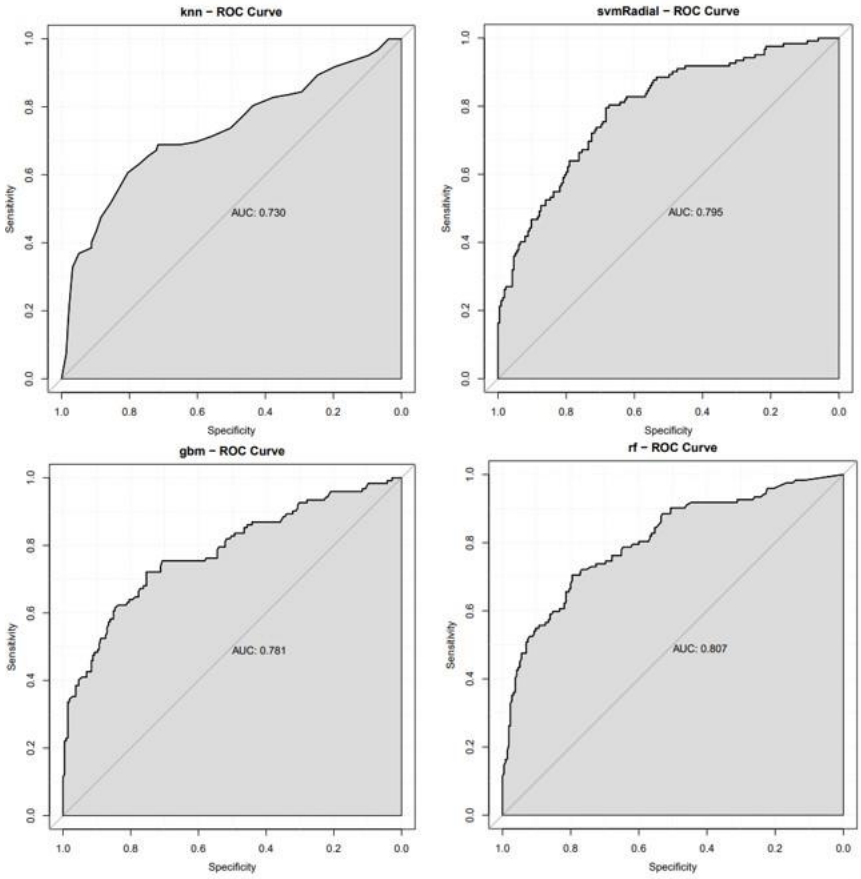

Figure 13: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.8, 0.8]. This shows sensitivity and specificity. The area under the curve of the KNN (0.730), svmRadial (0.795), GBM (0.781) and the RF (0.807) all show a moderate predictive power.

Table 10: Performance of the classification models at IC50 threshold of [-0.9,0.9].

Models	Accuracy	Precision	Recall	F1-score
KNN	0.759322	0.778802	0.880208	0.826406
SVM	0.755932	0.77027	0.890625	0.826087
GBM	0.779661	0.806763	0.869792	0.837093
RF	0.80678	0.816901	0.90625	0.859259

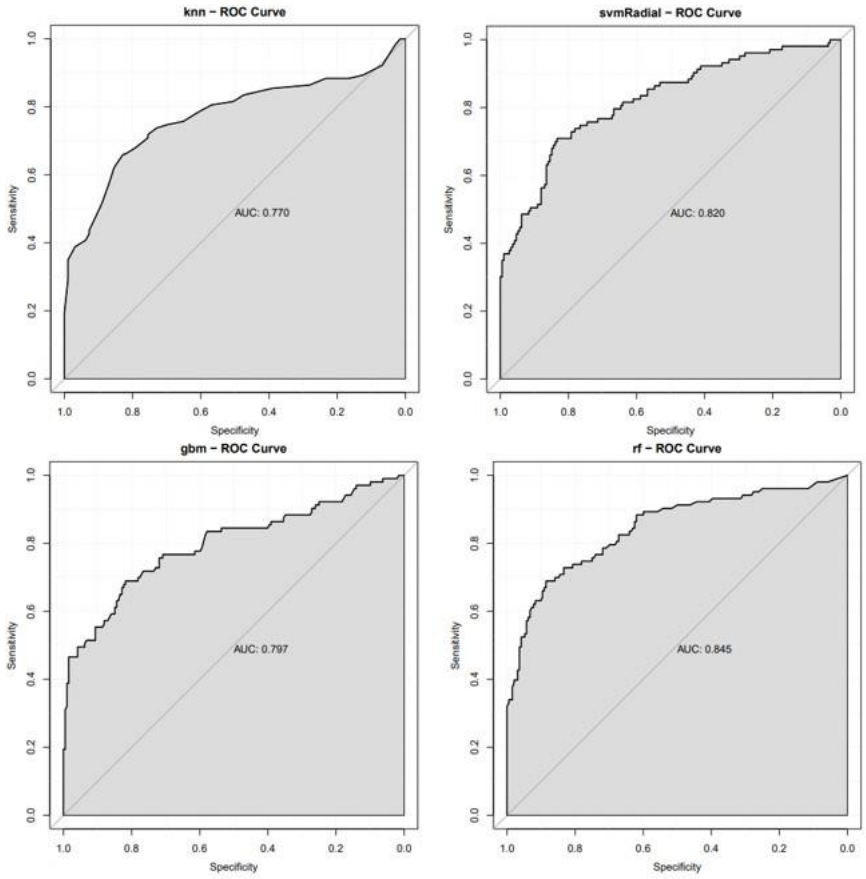

Figure 14: Receiver Operating Curves for predicting chemo-sensitivity of breast cancer cells at IC50 threshold of [-0.9, 0.9]. This shows sensitivity and specificity. The area under the curve of the KNN (0.770), svmRadial (0.820), GBM (0.797) and the RF (0.845) all show a moderate predictive power.

2.2.7.2 Model interpretation

An attempt was made to identify the most significant predictors for drug response in the sensitive class of the binary classification problem. The dataset encompassed 5000 input variables. The permutation feature importance model was employed to extract these predictors, whereby the model's performance was assessed by shuffling each predictor variable (i.e., genes). The scoring metric utilised was the mean decrease accuracy. The permutation feature method determines the difference between the original accuracy achieved by the model and the accuracy attained when the variables are randomly shuffled.

The top 300 genes were identified, and a ranked list of genes generated based on the mean decrease accuracy in descending order. (Table 10) (Supplementary Table 1). Further analysis focused on the top ranked genes.

TMEM126B, with a mean decrease accuracy of 0.698851, has been previously correlated with hypoxia-related pathways, which are associated with cancer [69,70]. RPL26 had the second highest mean decrease accuracy (0.683669), and the upregulation of this protein has been associated with increased cancer cell proliferation, and decreased cell apoptosis [71].

The results of this analysis may provide a steppingstone towards identifying the biological processes that these genes impact and the contribution of perturbances in these processes towards oncogenesis. The use of model interpretability saved a lot of time in identifying specific genes that influence the responses of cancer cells to anti-cancer drugs. This provides the opportunity to prioritise those genes that have a high degree of predictive power instead of exhaustively interrogating every gene in the expression.

Table 11: Top 25 genes ranked by mean decrease accuracy.

Genes	Scores
TMEM126B	0.698851
RPL26	0.683669
EXD3	0.681662
OLAH	0.678776
RPUSD3	0.677741
XPNPEP2	0.675405
CXorf38	0.673233
TARBP1	0.672867
MTERF3	0.670017
SAC301	0.668597
MS4A1	0.667992
GNB4	0.667809
ABCA4	0.666545
C4orf46	0.66648
STIP1	0.666353
SYN1	0.665921
POM121L12	0.664685
ELK1	0.664501
TAS2R30	0.663741
PTPN12	0.663026
SP9	0.661808
DOCK7	0.661707
SNIP1	0.6609
HNRNPH3	0.660635
RBM19	0.659444

2.2.7.3 Regression analysis

The aim of this analysis was to understand how the IC50 value (i.e., the dependent variable) changes as the genes (i.e., the independent variables) change. Regression analysis was also performed using the CRISPR-derived fitness scores but with the actual continuous IC50 scores rather than discretised versions of the scores. Four regression models were trained and evaluated using three performance metrics (Tables 12-20) (Figures 15-23).

Among the four models that were trained, the Elastic Net model showed the poorest performance in explaining the variability of cancer drug response as indicated by R^2 values consistently close to zero across the different thresholds (between -0.01399 and -0.00058. Although the KNN model yielded slightly higher R^2 than the Elastic Net across the different thresholds (R^2 values between 0.60353, and 0.0046922), it still had a low predictive power. The GBM and RF models achieved better performance across the different thresholds (R^2 values 0.689381 and 0.529491 for GBM and between 0.6555536 and 0.46291 for RF). This suggests that linear models like elastic net and nearest neighbour models like KNN have difficulty in capturing the complex relationships that likely exist between cancer drug responses and potential predictors of these responses.

When considering the RSME prediction accuracy metric, which measures the average difference of the values predicted by the model and the actual values, the GBM generally performed best across the different thresholds RSME between 0.664143 and 0.580671), with RF second best (RSME between 0.70958 and, 0.591839) and Elastic Net (RSME between 1.029431 and

0.970639)and KNN worst (RSME between 1.027488 and 0.937907).

Similarly with the MAE, which measures the average absolute difference between the predicted and actual values. The Elastic Net had high values across the different thresholds (MAE between
0.753038 and 0.7809017), followed by the KNN (MAE between 0.735661 and 0.807091). The GBM had moderate performance (MAE between 0.490535 and 0.506249) and the RF (MAE between 0.463283 and 0.5480002)

Overall, the GBM and the RF had the best predictive power. They were able to capture the complex relationship between predictors and the cancer drug response and make predictions that are closer to the true values of the drug response.

Table 12: Performance of the regression models at IC50 threshold of [-0.1,0.1].

Models	R_2	RSME	MAE
Elastic Net	-0.00601	1.04652	0.799116
KNN	0.030243	1.027488	0.799841
GBM	0.657258	0.610842	0.483337
RF	0.613844	0.648376	0.507908

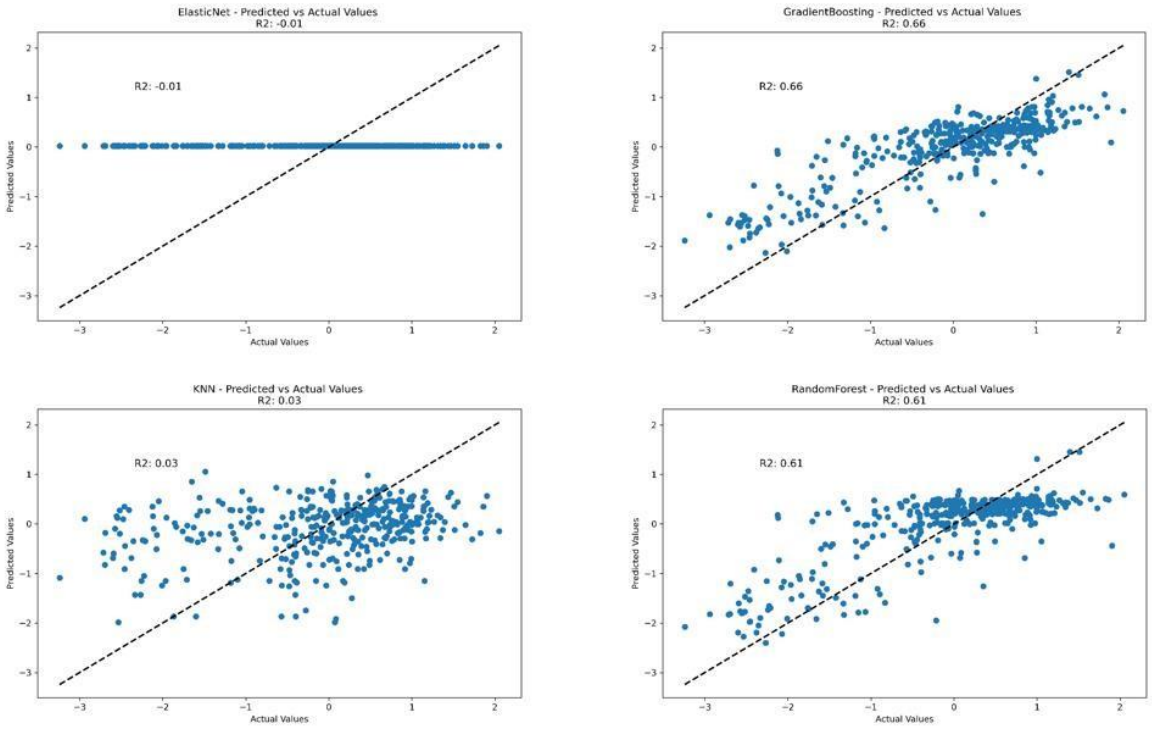

Figure 15: Comparison of regression models at IC50 threshold [-0.1,0.1] based on their R^2. This scatterplot displays the R2 values of the different models. The R2 values for each model are as follows: Elastic Net (-0.00499), KNN (0.055779), GBM (0.529591) and RF (0.46291).

Table 13: Performance of the regression models at IC50 threshold of [-0.2,0.2].

Models	R_2	RSME	MAE
Elastic Net	-0.01399	1.015428	0.77464
KNN	0.041551	0.987225	0.787714
GBM	0.668413	0.580671	0.460819
RF	0.655536	0.591839	0.471672

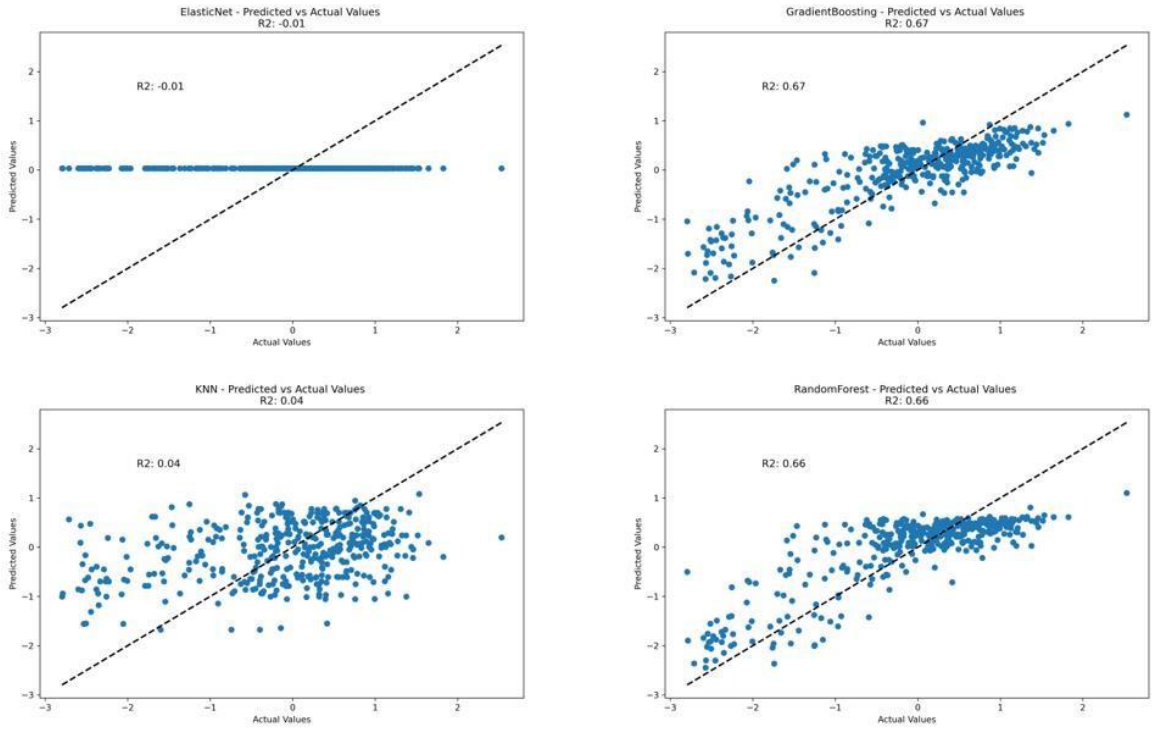

Figure 16: Comparison of regression models at IC50 threshold [-0.2,0.2] based on their R^2. This scatterplot displays the R2 values of the different models. The R2 values for each model are as follows: Elastic Net (-0.01399), KNN (0.041551), GBM (0.668413) and RF (0.655536).

Table 14: Performance of the regression models at IC50 threshold of [-0.3,0.3].

Models	R_2	RSME	MAE
Elastic Net	-0.01279	0.973729	0.763764
KNN	0.060353	0.937907	0.751046
GBM	0.604191	0.608724	0.474368
RF	0.577656	0.628797	0.495592

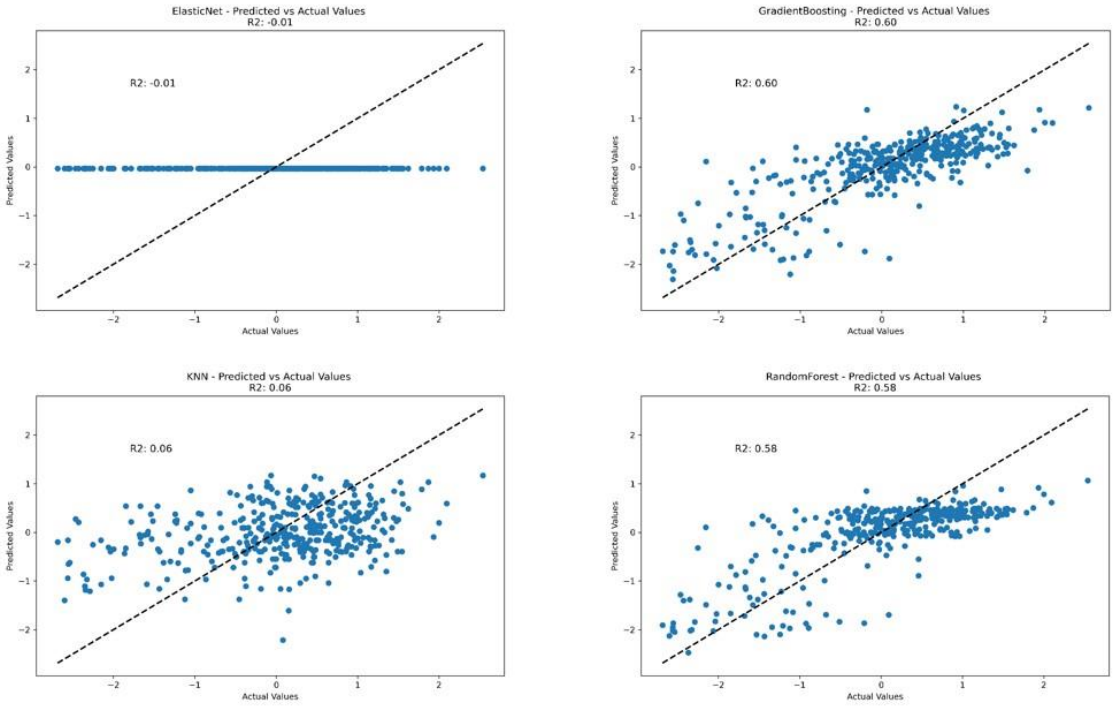

Figure 17: Comparison of regression models at IC50 threshold [-0.3,0.3] based on their R^2. This scatterplot displays the R^2 values of values of the different models. The R2 values for each model are as follows: Elastic Net (-0.01279), KNN (0.060353), GBM (0.60419) and RF (0.577656).

Table 15: Performance of the regression models at IC50 threshold of [-0.4,0.4].

Models	R_2	RSME	MAE
Elastic Net	-0.00058	1.051155	0.807475
KNN	0.040349	1.029431	0.807091
GBM	0.689381	0.585673	0.458537
RF	0.65574	0.616572	0.48435

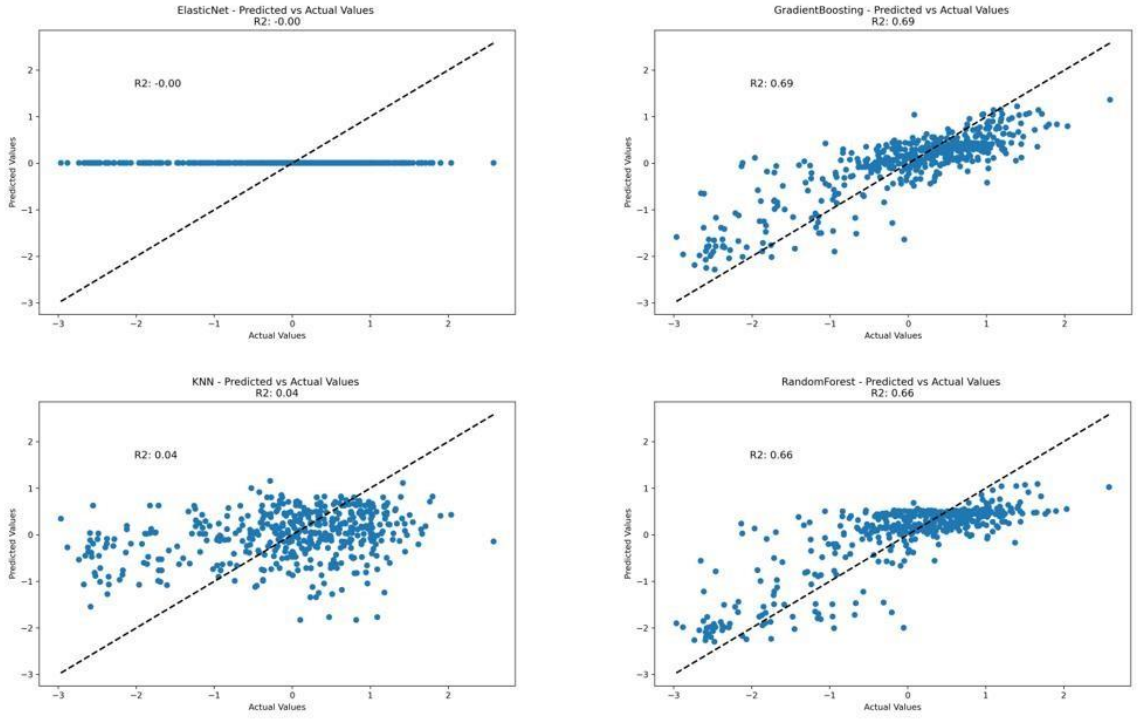

Figure 18: Comparison of regression models at IC50 threshold [-0.4,0.4] based on their R^2. This scatterplot displays the R^2 values of the different models. The R2 values for each model are as follows: Elastic Net (-0.00058), KNN (0.040349), GBM (0.689381) and RF (0.65574).

Table 16: Performance of the regression models at IC50 threshold of [-0.5,0.5].

Models	R_2	RSME	MAE
Elastic Net		1.034139	0.809017
KNN	0.0046922	1.009556	0.779364
GBM	0.642809	0.61804	0.490535
RF	0.599725	0.654253	0.518187

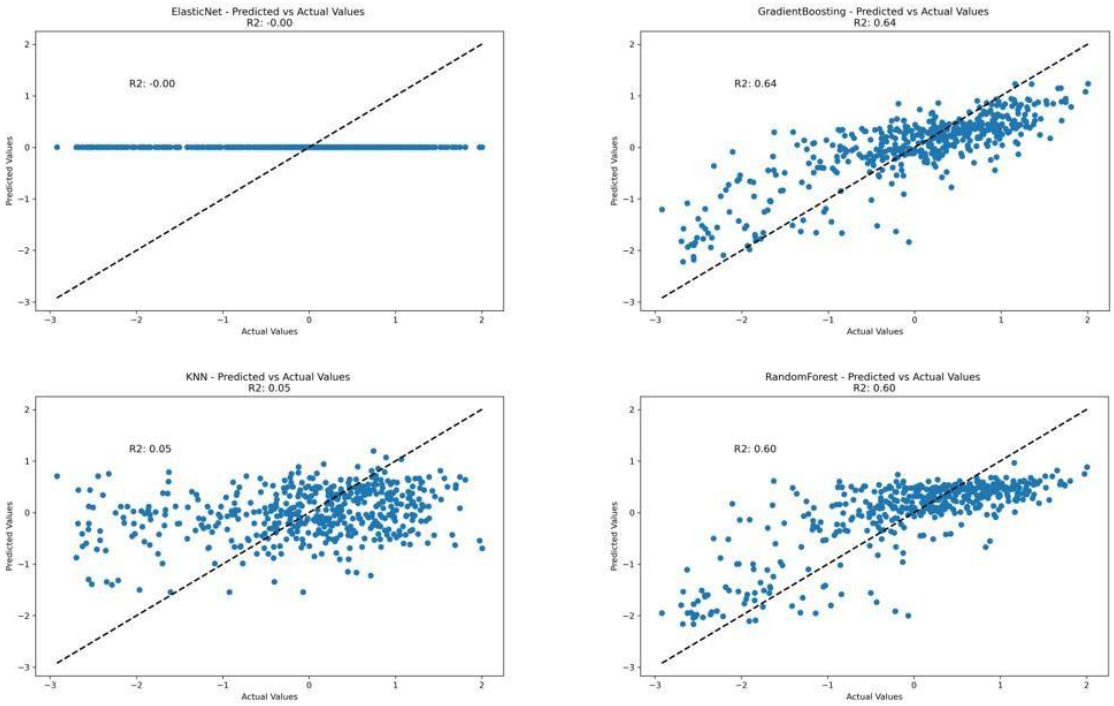

Figure 19: Comparison of regression models at IC50 threshold [-0.5,0.5] based on their R^2. This scatterplot displays the R^2 values of the different models. The R^2 values for each model are as follows: Elastic Net (-0.000058), KNN (0.046922), GBM (0.642809) and RF (0.599725).

Table 17: Performance of the regression models at IC50 threshold of [-0.6,0.6].

Models	R_2	RSME	MAE
Elastic Net	-0.00059	1.002694	0.781587
KNN	0.083782	0.959486	0.769012
GBM	0.624308	0.614406	0.481898
RF	0.572304	0.655552	0.520309

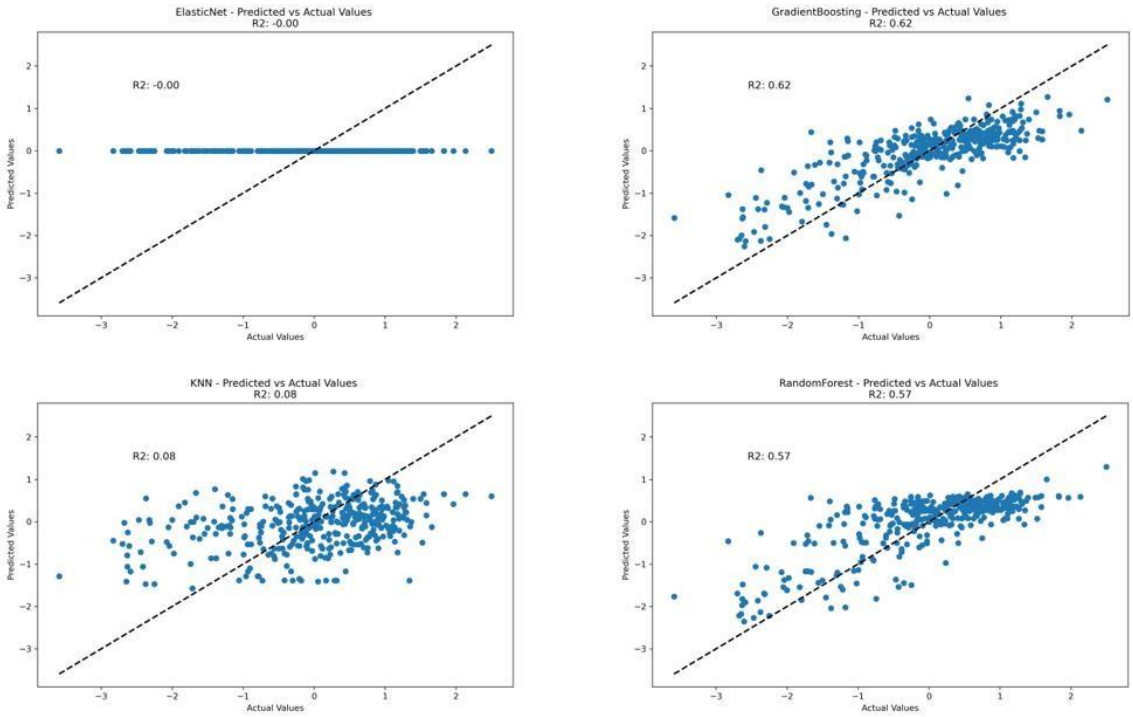

Figure 20: Comparison of regression models at IC50 threshold [-0.6,0.6] based on their R^2. This scatterplot displays the R^2 values of the different models. The R^2 values for each model are as follows: Elastic Net (-0.00059), KNN (0.083782), GBM (0.624308) and RF (0.572304).

Table 18: Performance of the regression models at IC50 threshold of [-0.7,0.7].

Models	R_2	RSME	MAE
Elastic Net	-0.00225	0.974824	0.753038
KNN	0.060909	0.94361	0.735661
GBM	0.61635	0.603123	0.463283
RF	0.564546	0.642554	0.463283

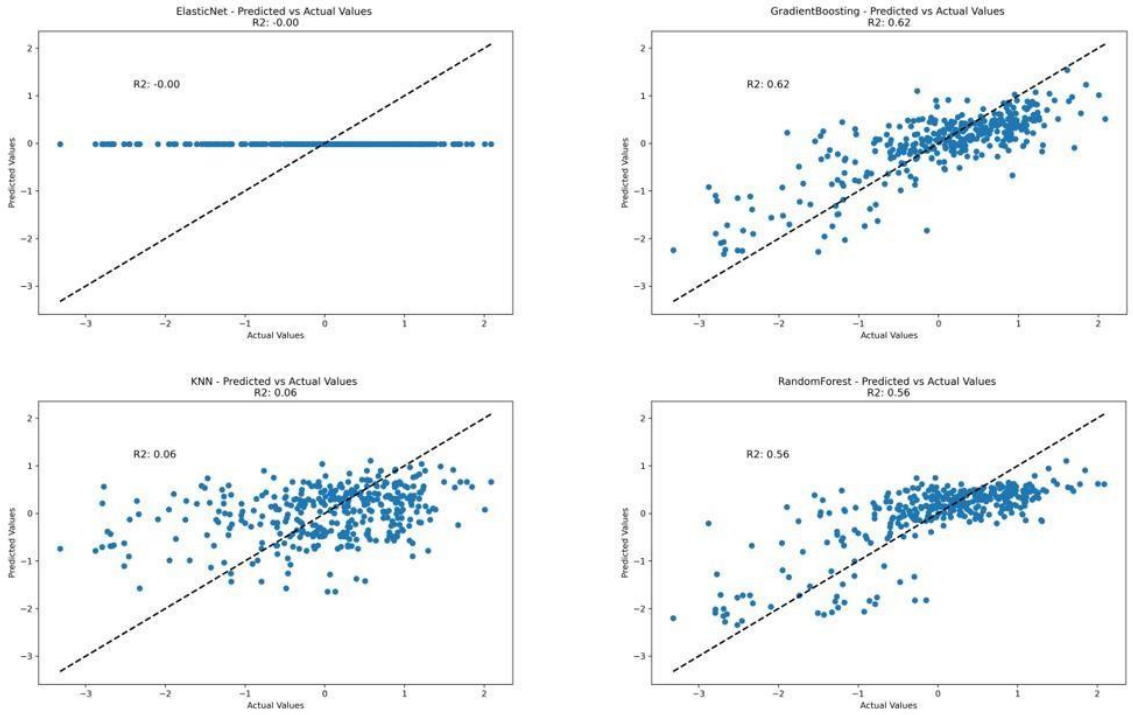

Figure 21: Comparison of regression models at IC50 threshold [-0.7,0.7] based on their R^2. This scatterplot displays the R^2 values of the different models. The R2 values for each model are as follows: Elastic Net (-0.00225), KNN (0.060909), GBM (0.61635) and RF (0.564546).

Table 19: Performance of the regression models at IC50 thresholds of [-0.8,0.8]

Models	R_2	RSME	MAE
Elastic Net		1.000309	0.781539
KNN	0.04188	0.979129	0.767758
GBM	0.607896	0.62637	0.486976
RF	0.554558	0.667615	0.528951

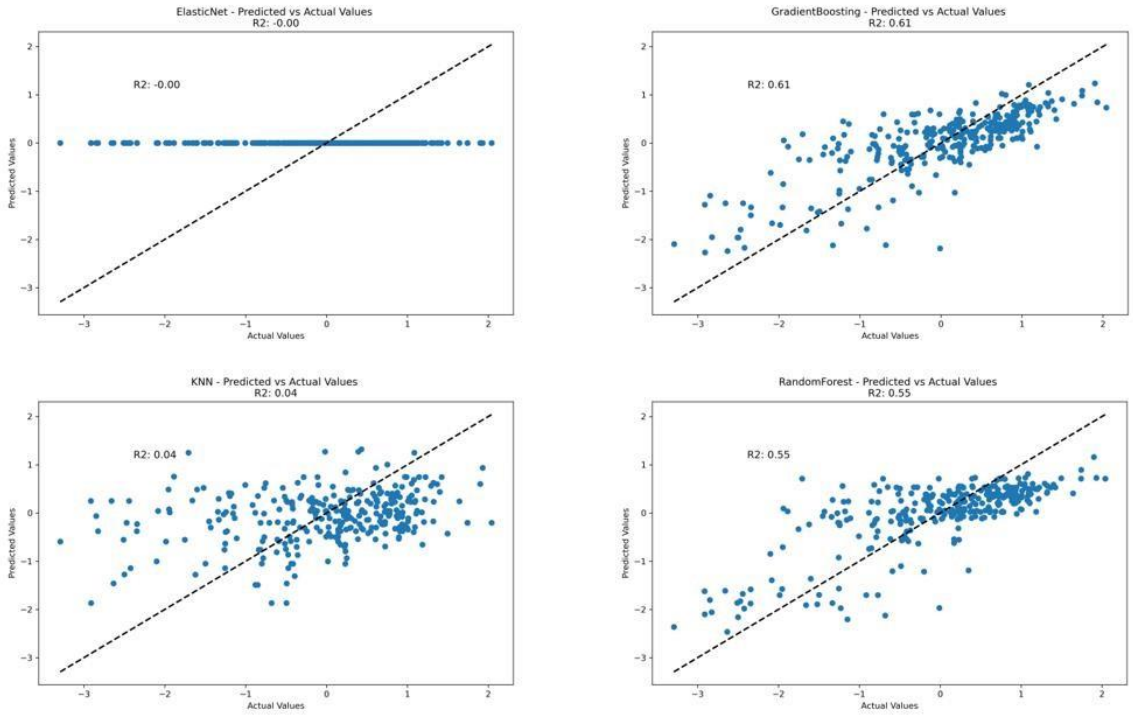

Figure 22: Comparison of regression models at IC50 threshold [-0.8,0.8] based on their R^2. This scatterplot displays the R^2 values of the different models. The R2 values for each model are as follows: Elastic Net (-0.000018), KNN (0.04188), GBM (0.607896) and RF (0.554558).

Table 20: Performance of the regression models at IC50 threshold of [-0.9,0.9].

Models	R_2	RSME	MAE
Elastic Net	-0.00499	0.970639	0.762016
KNN	0.055779	0.940838	0.770507
GBM	0.529491	0.664143	0.506249
RF	0.46291	0.70958	0.548002

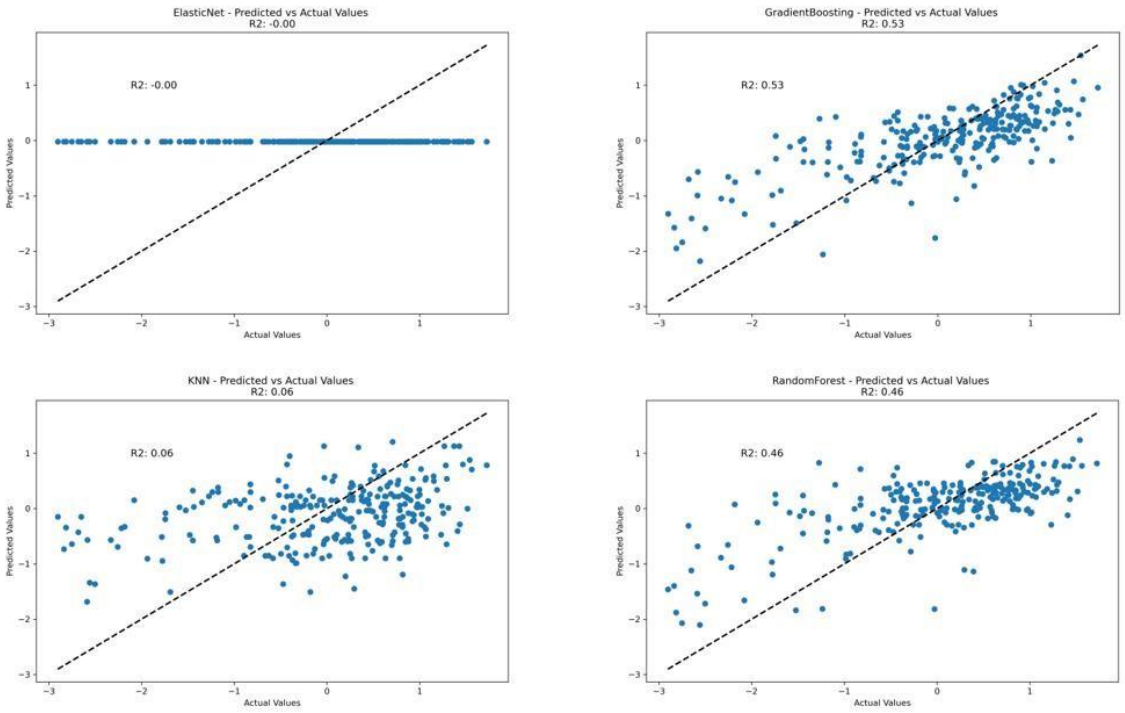

Figure 23: Comparison of regression models at IC50 threshold [-0.9,0.9] based on their R2. This scatterplot displays the R2 values of the different models. The R2 values for each model are as follows: Elastic Net (-0.00499), KNN (0.055779), GBM (0.529591) and RF (0.46291).

The integration of these methods enabled accurate predictions of drug sensitivity in cancer cell lines and could yield improved understanding of how, in cancer cell lines at least, oncogenes impact drug sensitivity. These findings will provide valuable insights into the underlying mechanisms of drug activity and furnish potential targets for future breast cancer therapies, chapter 3 discusses this further.

3. Mining pathways and kinases enriched in essential genes linked to cancer cells across different classes of anti-cancer drugs.

3.1 Introduction

The aim of this Chapter was to explore the relevance of genes identified in Chapter 2 as potential targets for breast cancer treatments using: (1) pathway enrichment analysis (PEA) to identify the pathways associated with oncogenesis, (2) identifying enriched kinases within these pathways (kinases are well established as eminently "druggable" targets) and (3) perform drug sensitivity analysis of the identified pathways and kinases. PEA was designed to detect biological pathways that have a significant association with the risk of developing a disease [72] using a gene collection of interest. This can be done by analysing the overrepresentation of genes or gene-classes within pre-defined pathways. PEA is useful both in interpreting the mechanistic underpinnings of tumorigenesis and identifying potential drug targets. Identifying significantly enriched pathways crucial in breast cancer formation involved utilising the extensive KEGG database [73–75].

On the other hand, kinase enrichment analysis (KEA) focuses on finding kinase signalling events that might be responsible for observed gene expression patterns. Through phosphorylation mechanisms, which alter protein activity, stability, and interaction networks, kinases are essential for controlling cellular functions. While dysregulation of kinases is associated with oncogenesis [76,77], it is also associated with neurological diseases [78] and metabolic diseases [79]. Kinases are therefore among the proteins most frequently targeted by drugs (second only to G protein coupled receptors). To identify the kinase activity and signalling pathways that may be causing observed changes in gene expression within breast cancer cells, the enrichment of kinase-substrate interactions within the gene collection identified in section 2.2.6 of Chapter 2 was analysed. Performing KEA is important because it

can assist in identifying drug targets which could be potentially targeted by one or more of the broad arrays of already known kinase inhibitors.

This was achieved by identifying which specific kinases were enriched within the input gene list.

3.2 Materials and Methods

3.2.1 Pathway enrichment analysis

Overrepresentation analysis was conducted on the set of genes identified in section 2.2.6 using the KEGG database. The list of genes was imported into R using the *tidyverse* (v2.0.0) [58] package and processed by removing any genes from the list that did not have a contribution. Enrichment analysis was performed using the *enricher()* function from *clusterProfiler* (v4.6.2) [80,81], where a 300 gene list (Table 10 in chapter 2) (Supplementary table 1) was the inputs. Then, a gene list consisting of multiple gene symbols was created.

The gene list included various genes such as “TMEM126B”, “RPL26”, “EXD3”, and 297 others. The *org.Hs.eg.db* from *BiocManager* (v1.30.21) [82] package was used to convert HUGO gene symbols to Entrez Gene IDs. The *mapIds()* function was used to map the gene symbols from the gene list to their corresponding Entrez Gene IDs using the “org.Hs.eg.db” database. After obtaining the mapping results, the *enrichKEGG ()* function from *clusterProfiler* (v4.6.2) [80,81] was used to perform KEGG pathway enrichment analysis. The *gene* parameter was set to the obtained Entrez Gene IDs, the *organism* parameter was set to *hsa* (indicating that the genes were human in origin), and the *pvalueCutoff* parameter was set to 0.05.

The enrichment analysis results were stored in the "ans.kegg" object, which was then converted to a data frame named "tab.kegg". Various operations were performed on "ans.kegg" object and

"tab.kegg" data frame to extract and format the results. Specifically, the "enriched" data frame was created by separating the ratio values, converting them to numeric format, and calculating the enrichment ratio ("k.K"). Finally, the top 20 enriched pathways were visualized using a bar plot created with *ggplot2* (v3.4.2) [83].

3.2.2 Kinase enrichment analysis

The input gene list included those identified by PEA (see section 3.2.1). The *org.Hs.eg.db* from *BiocManager* (v1.30.21) [82] was loaded to enable the conversion of Entrez Gene IDs to HUGO gene symbols, the preferred input of eXpression-2 Kinases.

The eXpression-2 Kinases (X2K) enrichment analysis online tool (https://maayanlab.cloud/X2K/) was used to connect lists of proteins and genes with the kinases that phosphorylate them [84]. This program uses Fisher's exact test to calculate the probability of kinase enrichment, based on the distribution of kinase-substrate proportions in the background kinase substrate database, compared to the protein kinases associated with a given input list of proteins.

3.3 Results and Discussion

3.3.1 Pathway Enrichment Analysis

Performing PEA was essential to gain insights in the biological pathways that are associated with the risk of developing breast cancer. This analysis revealed several statistically significant enriched pathways (i.e., with p-values < 0.05), including those involved in oxidative phosphorylation, thermogenesis, amyotrophic lateral sclerosis, huntington disease, prion

disease, ribosome related pathways, chemical carcinogenesis, parkinson disease, non-alcoholic fatty liver disease, fatty acid biosynthesis and diabetic cardiomyopathy (Figure 24).

While oxidative phosphorylation is a vital metabolic process for energy production in normal cells, in cancer cells it plays a crucial role in tumour progression and metastasis [85,86]. The upregulation of oxidative phosphorylation in cancer cells generates reactive oxygen species that can damage DNA and contribute to the genomic instability of these cells [87]. Thus, the enrichment of oxidative phosphorylation in this analysis suggests that this pathway may be an attractive target for developing new anti-breast cancer therapies.

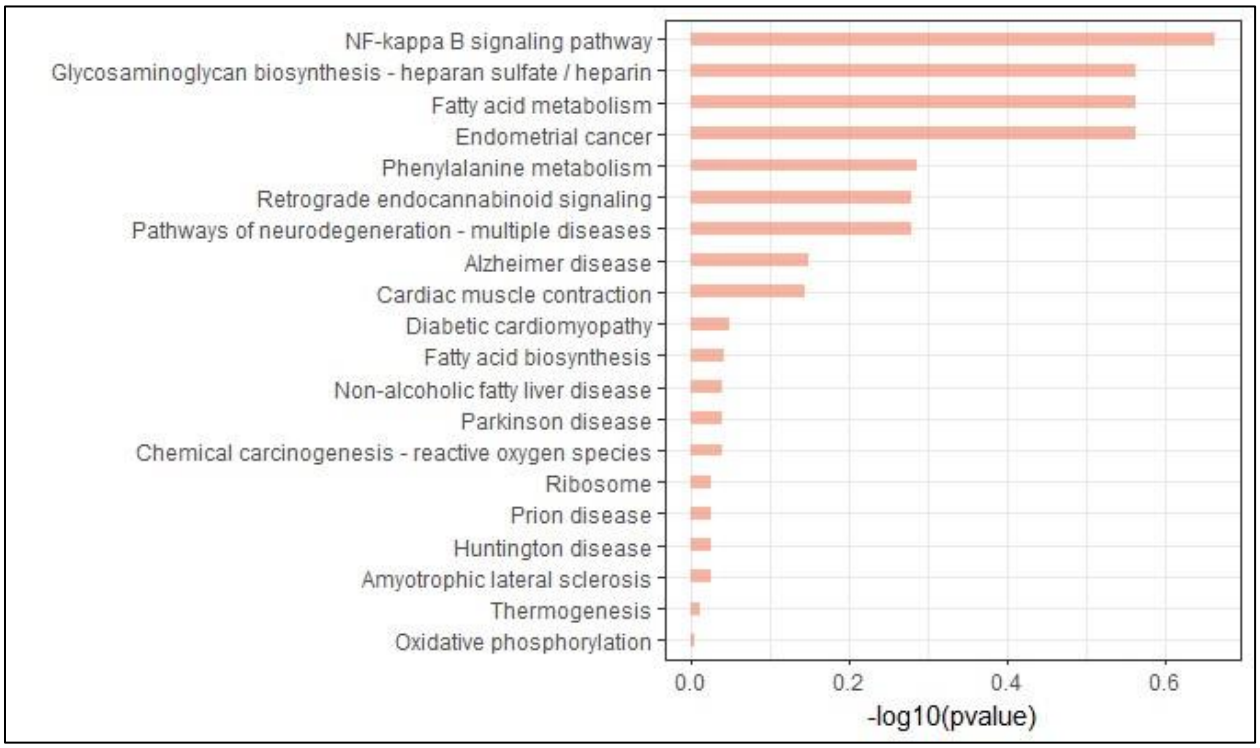

Figure 24: Top 20 significantly enriched pathways arranged in descending order by their adjusted p-values.

Fatty acids are important energy sources and structural components of cells in most species, including humans. However, high levels of fatty acids have been associated with increased risk

of developing cancer, including breast cancer [88,89]. The excessive presence of fatty acids through fatty acid biosynthesis is needed to sustain the uncontrolled growth and increased survival of cancer cells. Therefore, restricting the availability of fatty acids in cancer cells can serve as a therapeutic approach.

The PEA also identified chemical carcinogenesis (reactive oxygen species pathways). Chemical carcinogenesis refers to the process through which exposure to certain chemicals, such as tobacco smoke and alcohol, among others, can potentially lead to the development of cancer. Chemical carcinogenesis frequently involves the generation of ROS with cancer cells often exhibiting higher levels of ROS compared to normal cells [90]. ROS can cause DNA damage, leading to an increase in mutagenesis and genomic instability which, in turn, can promote the occurrence of mutations that favour tumour growth and resistance to therapy. It is therefore important to minimize exposure to chemical carcinogens and adopt a healthy lifestyle to reduce the risk of cancer development.

In conclusion, the PEA analysis illuminates the molecular underpinnings of breast cancer and identifies novel therapeutic targets, paving the way for the application of machine learning techniques to personalise treatments, thereby enhancing their precision and effectiveness.

3.3.2 Kinase Enrichment Analysis

Breast cancer is a complex disease characterized by the disruption of multiple signalling pathways, including those involving kinases. In the exploration of pathways identified by the PEA in section 3.2.2, attention was given to the specific lists of genes involved, with a focus on kinases that might be targetable by anti-breast cancer drugs. The top 20 significantly enriched kinases within the pathways yielded by the PEA identified (all with associated p-values < 0.05 (Figure 25) were identified.

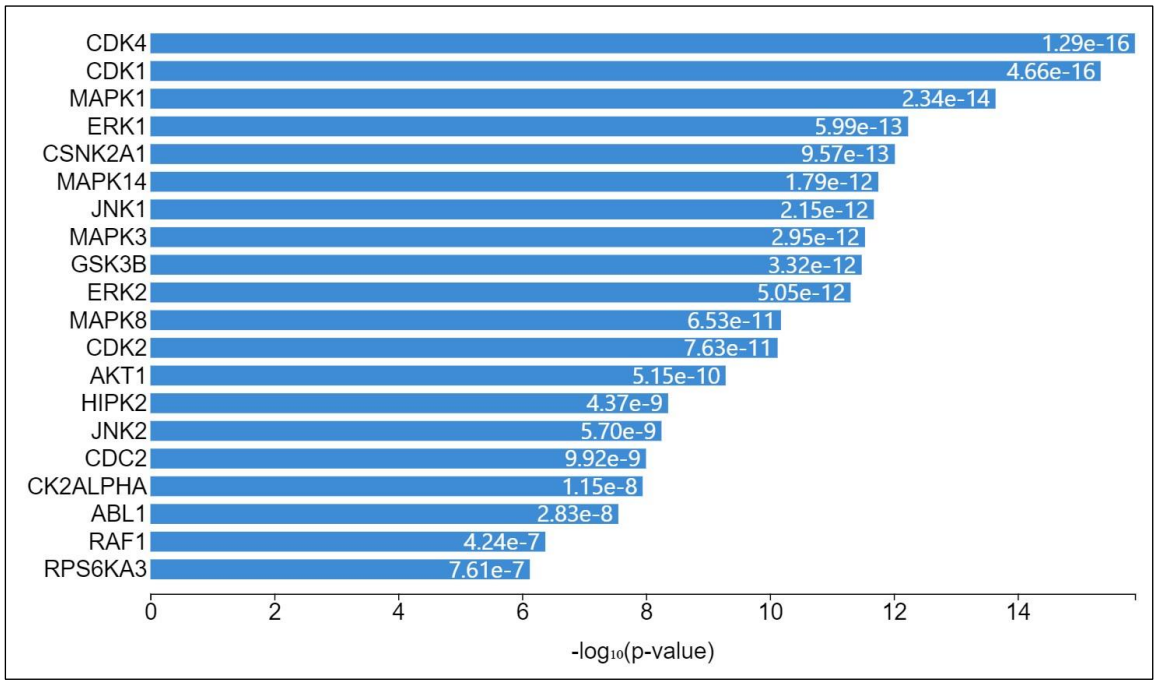

Figure 25: The top 20 significantly enriched kinases from the input gene set ranked by their p-values.

The two kinases that were most enriched in breast cancers were CDK4 and CDK1. These kinases belong to the cyclin-dependent kinase family and have long been recognized as key regulators of the cell cycle [91,92]. Their enrichment in breast cancer further supports the critical role of cell cycle dysregulation in tumour development. CDK4 has been identified as one of several potential targets in treating chemotherapy resistant TNBC [86].

The fifth most enriched kinase in breast cancers was CSNK2A1; also known as casein kinase 2 alpha. CSNK2A1 has been implicated in various cellular processes, such as cell proliferation and survival, and the inhibition of CSNK2A1 has been proved to decrease the proliferation and invasiveness of breast cancers [93].

Many of the top 10 most enriched kinases in breast cancers are members of the mitogen-activated protein kinase (MAPK) family, including ERK1, ERK2, MAPK1, MAPK14, and

MAPK3. These play important roles in intracellular signalling pathways and have been associated with cell growth, migration, and invasion [94,95].

Another notable kinase that is enriched in breast cancers is CK2ALPHA (17th on the list): a serine/threonine kinase, known to play a role in cell proliferation, differentiation, and tumorigenesis [96,97].

Other kinases that were identified as being enriched in association with breast cancers in our analysis, have diverse functions (DNA repair, apoptosis, cell survival, and protein synthesis) with many such as HIP2K [98] and GSK3B [99] having known associations with cancer.

The enrichment of these various kinases in breast cancer cells highlights both their potential involvement in promoting the progression of breast cancer, and their potential as targets for therapeutic interventions. While promising it is important to note that this analysis represents only an initial step in understanding the role of these kinases in breast cancer. Further studies are required to both validate the functional significance of each of these kinases and unravel the underlying molecular mechanisms through which they contribute to breast cancer development and/or pathogenesis. This is in light of the fact that there are only 72 FDA approved kinase inhibitors that target different protein kinases [100].

4. Investigating gene expression changes in breast cancer cell lines after drug perturbation.

4.1 Introduction

Breast cancer remains a significant global health concern, and the quest for effective treatment strategies continues to be at the forefront of cancer research. Understanding the treatment responses of different breast cancer tumours is crucial for improving therapeutic outcomes and tailoring treatment regimens to individual patients. One promising avenue of investigation is the study of gene expression changes in breast cancer cells following therapy. By analysing alterations in gene expression profiles, one can uncover molecular mechanisms underlying treatment responses and resistance, paving the way for personalised medicine approaches.

In this chapter, the focus on investigating gene expression changes in breast cancer cell lines and explaining the molecular factors contributing to treatment response and treatment resistance in controlled experimental settings. To accomplish this, the GDSC_LINCS dataset, representing posttreatment outcomes and the GDSC_CCLE, representing pre-treatment states are used. Both are comprehensive resources covering diverse breast cancer cell lines and their response to various therapeutic interventions. Specifically, the analyses performed centred around observing six hours after drug treatment, changes in the expression of the important genes identified in section 2.2.6 (Table 11) (Supplementary Table 1).

4.2. Materials and Methods

4.2.1 Data Collection

To acquire the necessary datasets for this analysis, the LINCS datasets were obtained from the Gene Expression Omnibus (GEO) the GEO Accession viewer (nih.gov) using the GSE101406 accession ID. The LINCS L1000 Level 4 gene expression dataset the LINCS gene information and LINCS drug information datasets were downloaded. The LINCS datasets offer detailed gene expression profiles, encompassing data on specific genes and drugs. This rich dataset facilitated exploration into how gene expression shifted in relation to therapeutic interventions. Additionally, the GDSC data was obtained from the Sanger Institute website through the following FTP links:

ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC2_fitted_dose_res

ponse_25Feb20.xlsx and

ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/current_release/GDSC1_fitted_dose_res ponse_25Feb20.xlsx. The GDSC dataset offers insights into drug responses across various cancer cell lines. Gene expression data from the CCLE and sample information data were obtained from the Broad Institute website through the following link: https://depmap.org/portal/download/all/ .

In this chapter, the aim was to delve into the connection between gene expression variations and drug sensitivity/resistance in breast cancer cell lines. Using such diverse datasets facilitated the linking of gene expression patterns with the chemosensitivity of cancer cells. The goal of

integrating and analysing these datasets, was to illuminate the complex relationships between chemosensitivity and gene expression changes upon drug-induced perturbations.

4.2.2 Data preprocessing

4.2.2.1 Collapsing LINCS replicates.

The LINCS dataset has several repeated measurements of drug perturbations, collapsing this dataset simplifies the data making it easier to understand, analyse and interpret. Each cell line had triplicate post-treatment mRNA readings for each drug, which were collapsed into a single signature using the method shown in Figure 26, which is the standard method for converting Level 4 LINCS data with triplicates into single-signature Level 5 data. These single-signature data were saved as csv files.

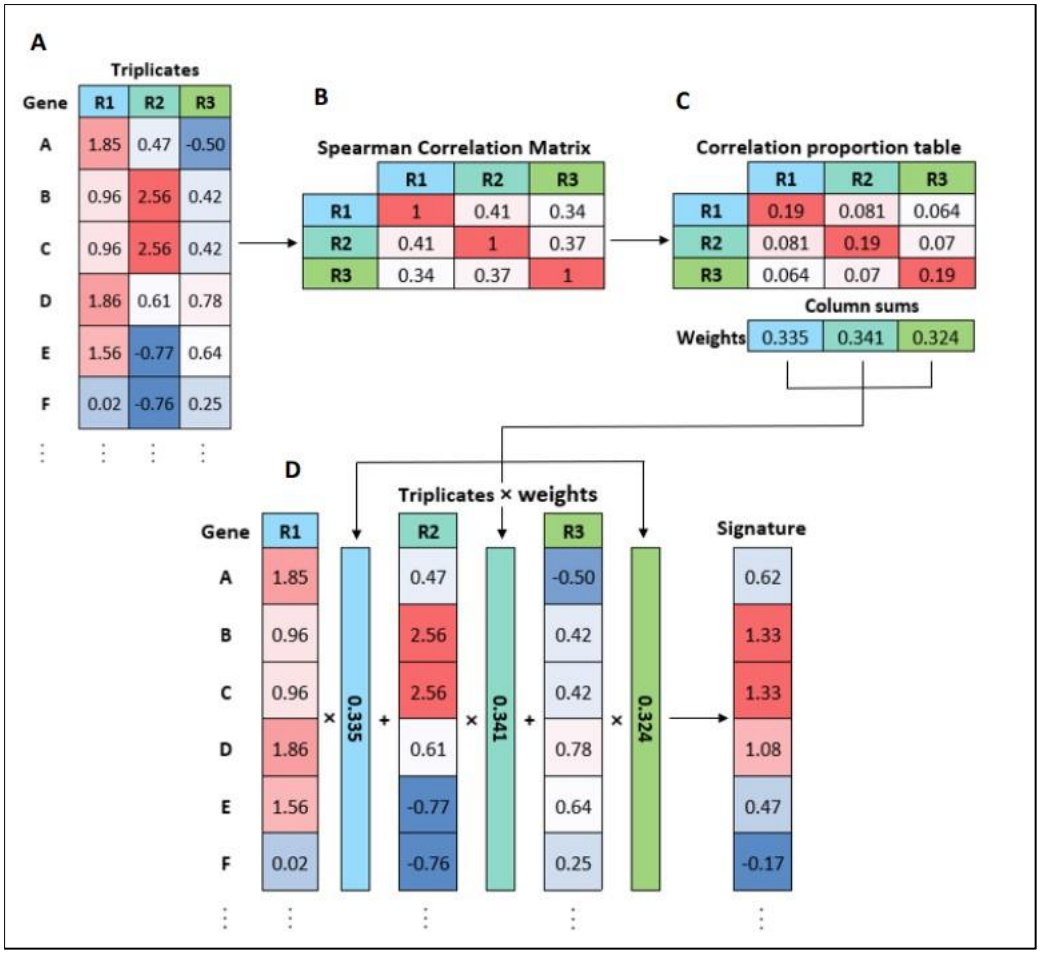

Figure 26: Method for collapsing triplicates of post-treatment data for a given cell line (X), treated with a given drug (Y). (A) The data from is adjusted to a Z-score. (B) A spearman correlation matrix was created to show how each set of data correlates with the others. (C) Each correlation from the matrix was divided by the sum of all values to get a table showing proportions of correlations. (D) The signature is obtained from the sum of the replicates and their weights.

4.2.2.2 GDSC_LINCS data preprocessing

To understand the effects of the drugs on the expression of genes, and to identify the drug responsive genes the GDSC and LINCS datasets were acquired. The *tidyverse* (v2.0.0) [58] package was loaded for data manipulation. The MCF7 L1000 level 5 collapsed signature dataset, which includes breast cancer cell lines was imported using the *read_csv()* function.

The GDSC datasets, GDSC1 and GDSC2, were imported and merged as described previously in section 2.2.2. The dataset was filtered to include only the drug response profiles of breast cancer cell lines based on the GDSC cell line annotations. The focus was on the GDSC dataset which had an IC50 threshold of [-0.9 to 0.9], because it produced the best performance during the machine learning analysis. Lastly, the GDSC and LINCS dataset were merged (Figure 27).

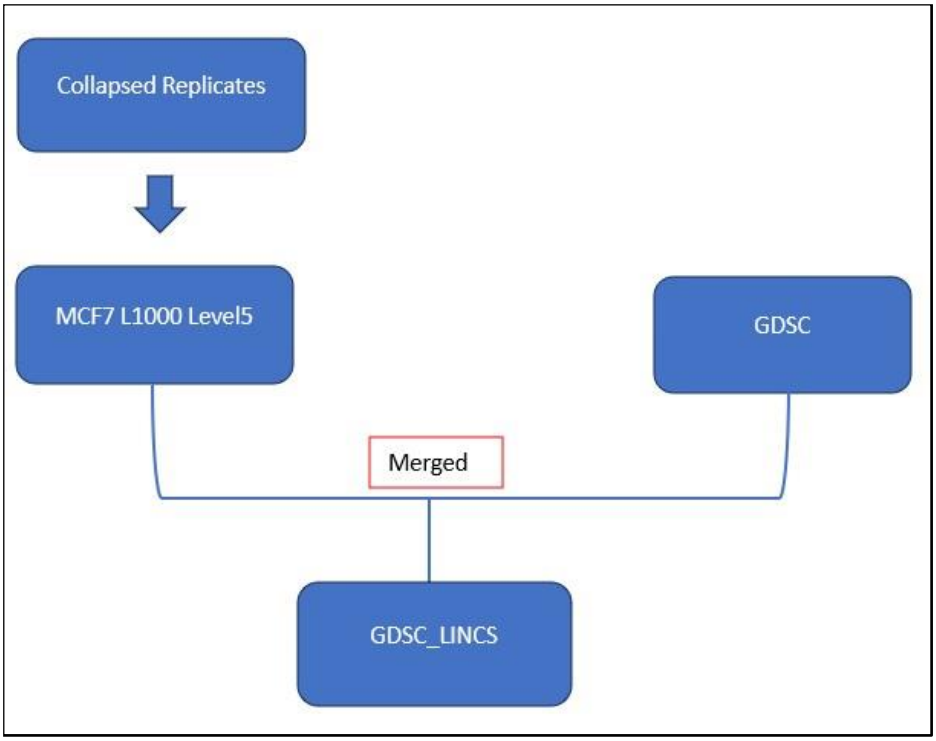

Figure 27: Data preprocessing of GDSC_LINCS. The sample information and the gene expression data were merged using cell lines. The LINCS and the GDSC datasets were merged using common columns.

4.2.2.3 GDSC_CCLE data preprocessing

The CCLE dataset and the GDSC datasets were merged. The focus here was on the GDSC dataset which had an IC50 threshold of [-0.9 to 0.9], because this dataset yielded the best performance during the machine learning analysis (Figure 28).

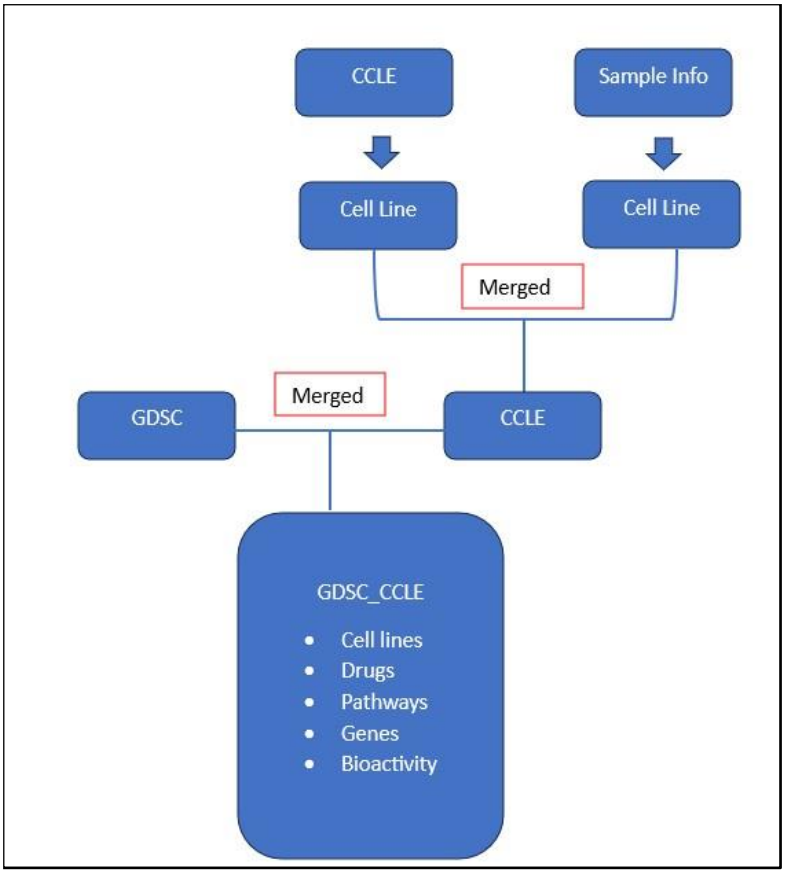

Figure 28: Data preprocessing of CCLE dataset. The sample information and the gene expression data were merged using cell lines. The CCLE and the GDSC datasets were merged using common columns.

4.2.2.4 Gene expression change data preprocessing.

The *tidyverse* (v2.0.0) [58] package was employed for data manipulation and analysis. The merged GDSC_LINCS dataset, representing post-treatment, and GDSC_CCLE dataset, representing pretreatment, were obtained by importing the respective csv files. Data preprocessing included filtering the datasets to include only entries with "RESISTANT" values in the BIOACTIVITY column, facilitating the analysis of gene expression changes in drug-resistant cell lines. Subsequently, the datasets were filtered to include entries with "SENSITIVE" values in the BIOACTIVITY column, enabling the analysis of gene expression changes in drug-sensitive cell lines. Furthermore, the data was filtered to include only the cell

lines, pathways and drugs that are present in both the GDSC_CCLE and GDSC_LINCS datasets (Table 21). The data was normalised.

Table 21: This table represents 31 cell lines, the 3 drugs that treat the cell lines and the pathways targeted before and after treatment.

Cell Lines		Pathways	Drugs
AU656	HDQ-P1	Chromatin other	I-BET-763
BT-20	JIMT-1	Genomic Integrity	KU-55933
CAL-120	MCF7	PI3K/MTOR signalling	OSI-027
CAL-51	MDA-MB-157		
CAMA-1	MDA-MB-157-VII		
COLO-824	MDA-MB-231		
HCC1143	MDA-MB-415		
HCC1187	MFM-223		
HCC1395	UACC812		
HCC1419	UACC893		
HCC1428	ZR-75-30		
HCC1937			
HCC1954			
HCC202			
HCC2218			
HCC38			
HCC70			

4.2.2.5 Gene expression change analysis.

A list of the top genes identified in section 2.2.6 was generated, the availability of these genes was checked in the GDSC_LINCS and GDSC_CCLE dataset. These genes were extracted by comparing the gene names with the columns of the GDSC_LINCS and GDSC_CCLE datasets. Corresponding columns were extracted from both datasets, and the numerical data was normalized using the *scale()* function.

To compare the expression levels of the selected genes between pre-treatment and post-treatment datasets, the paired samples Wilcoxon signed rank test was conducted for both sensitive cell lines and resistant cell lines. The Wilcoxon signed rank test is a suitable statistical test for comparing means between two groups when the data is not normally distributed, making it appropriate for assessing the significance of gene expression changes. Statistical significance was determined using the significance threshold of $p<0.05$. Genes with p-values below the significance threshold were considered to have significant expression differences between pre- and post-treatment. Mean differences were used to describe changes in expression levels. A positive mean difference indicates an increase in gene expression levels in response to the treatment, while a negative mean difference indicates a decrease in gene expression levels.

4.3 Results and Discussion

A total of 144 genes (Supplementary Table 2) of the 300 genes identified in section 2.2.6 were found in both GDSC_LINCS and GDSC_CCLE dataset and used for the analysis.

4.3.1 Gene expression changes in sensitive cell lines

Out of the 144 genes (Supplementary Table 2) found in both the GDSC_CCLE and GDSC_LINCS, only 1 gene -FIG4-showed statistical significance with a p-value of 0.03 in the difference in expression between the pre and post treatment datasets. (Table 22) (Figure 30).

While, to my knowledge, there are no studies on the contributions of OLAH, EXD3, MTERF, MRPS30, FIG4 or TGM5 to either cancer development or the treatment of cancers, the other five identified genes have all been previously implicated in the development of various cancers:

SAC3D1: is abnormally expressed in multiple types of cancers and it is associated with gastric cancer progression [101] and over expression is detected in hepatocellular cancer [102].

STIP: is overexpressed in breast cancer tissue and the knockdown of STIP1 results in the decrease of cell proliferation inducing apoptosis [103]. In this analysis the mean difference is positive, indicating an increase in expression in sensitive cell lines.

TARBP1: is overexpressed in hepatocellular carcinoma [104] and non-small-cell-lung cancer [105]. The positive mean difference indicates an increase in the expression of TARBP1 in breast cancer cells after treatment. This could potentially mean that TARBP1 could be prognostic marker for breast cancer.

RBM19: is commonly expressed in the intestinal epithelium and is critical for intestinal morphogenesis [106]. However, RBM19 is overexpressed in patients with hepatocellular carcinoma [107]. In this analysis a negative expression value was observed.

HYALI: a decrease in expression is associated with the migration and invasiveness of colorectal cancer, and overexpression is associated with the suppression of colorectal cancer [108].

4.3.2 Gene expression changes in resistant cell lines

Out of the 144 genes (Supplementary Table 2) found in both the GDSC_CCLE and GDSC_LINCS none of the genes showed statistical significance (Table 22) (Figure 31).

TMEM126B: belongs to the trans membrane protein family where some members have been linked to oncogenesis. For instance, TMEM158 upregulation has been observed in triple negative breast cancer [109,110]. However, based on the findings, it appears there are no studies directly linking TMEM126B with cancer. However, TMEM126B has been associated with pathways involved in hypoxia. Regions characterized by high inflammation, such as rapidly proliferating cancer cells, commonly exhibit low oxygen levels [69,70].

TARBP1: is overexpressed in hepatocellular carcinoma [104] and non-small-cell-lung cancer [105]. A positive expression value is observed in this analysis.

SNY1: is implicated in neural function within the nervous system. In gliomas, which are tumours of the central nervous system, SYN1 expression is reduced due to the upregulation of RE-1 silencing transcription factor (REST) and REST corepressor 1 (RCOR1), and it contributes to the maintenance of cancer stem-like phenotypes that promote the development of gliomas [111].

PTPN12: is a tumour suppressor that is downregulated in triple negative breast cancer [112]. In this analysis, the negative mean difference of PTPN12 indicates low expression levels.

HNRNPH3: is overexpressed in meningioma, benign tumours of the central nervous system which can transform to malignancy [113]. There has been no information about the expression of HNRNPH3 on malignant tumours.

EIF3D: is overexpressed in gallbladder [114], colorectal [115] and breast cancer [116]. It has been linked to an advanced tumour stage and unfavourable prognosis. This analysis reveals that, the mean expression is negative in resistant cell lines.

In this analysis there are instances where the mean expression value of the genes is negative, but the cell lines are still resistant to the drugs. This can be attributed to several factors:

(1) Insufficient drug concentrations in the cancer cell lines, could cause incomplete suppression of the gene expression.

(2) The drugs inability to effectively target the specific signalling pathways responsible for the cancers' aggressive nature, leading to the incomplete reduction in the gene expression.

(3) Alternative pathways are activated, causing the rapid proliferation of the cell lines.

This supports the argument presented by Lippert et al [117], that drug resistance occurs whether it is intrinsic resistance (resistance without the need for mutations) or acquired resistance (resistance caused by mutations) and this resistance is influenced by multiple factors.

Table 22: Detailed gene expression differences for top 20 genes, between pre-treatment and post-treatment for sensitive cell lines

Gene	Mean Difference	p-value
TMEM126B	0.393930648	0.41433123
EXD3	-0.075263627	0.94993218
OLAH	0.209618742	0.52975227
XPNPEP2	0.436372400	0.65668081
TARBP1	0.34364094	0.66026766
MTERF3	0.197665453	0.75355161
SAC3D1	0.010475705	0.94993218
ABCA4	0.216124667	0.20917435
STIP1	0.337878908	0.14867855
SYN1	-0.340327942	0.85043627
PTPN12	-0.059609700	0.90006131
HNRNPH3	0.359805342	0.14867855
RBM19	-0.438893611	0.31505520
EIF3D	-0.269321441	0.90006131
NRDC	0.207876427	0.57198701
TGM5	-0.412318423	0.45103590
HYAL1	-0.474000552	0.20917435
FIG4	-0.419018074	0.02796919
MRPS30	0.261364086	0.41433123
ANXA5	0.113945288	0.53005668

Table 23: Detailed gene expression differences for top 20 genes, between pre-treatment and post-treatment for resistant cell lines.

Gene	Mean Difference	p-value
TMEM126B	-0.221434526	0.43042361
EXD3	-0.100615753	0.55311782
OLAH	-0.107093138	0.65001467
XPNPEP2	-0.027695099	0.14053883
TARBP1	0.083265557	0.90557727
MTERF3	0.114704630	0.0.76414961
SAC3D1	-0.149037390	0.25539360
ABCA4	0.106313031	0.64013540
STIP1	-0.095563016	0.86152000
SYN1	-0.056953205	0.75348411
PTPN12	-0.211916467	0.14475373
HNRNPH3	-0.228287077	0.35339789
RBM19	0.138240929	0.62031183
EIF3D	-0.238357762	0.21166964
NRDC	-0.053441606	0.83964317
TGM5	--0.056529508	0.0,90556583
HYAL1	-0.092624808	0.87249949
FIG4	-0.172379059	0.39075840
MRPS30	0.003654691	0.93881970
ANXA5	-0.231625218	0.41428437

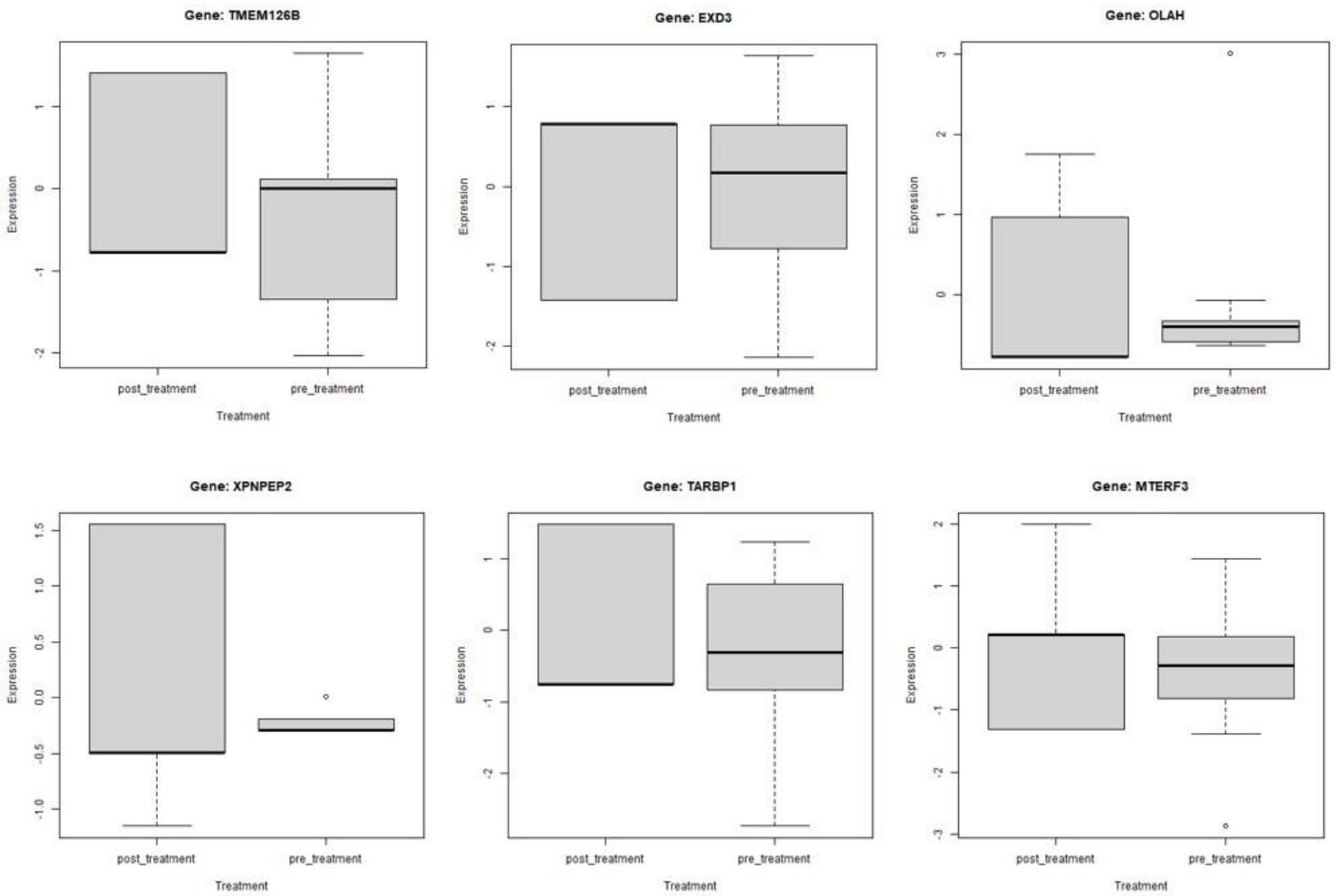

Figure 29: Gene expression change for the first 6 genes in sensitive cell lines. A positive mean difference indicates an increase in expression levels, while a negative indicates a decrease in expression levels. These boxplots illustrate the distribution of expression values in pre-treatment and posttreatment, highlighting the differences between the two groups for genes after 6 hours of treatment.

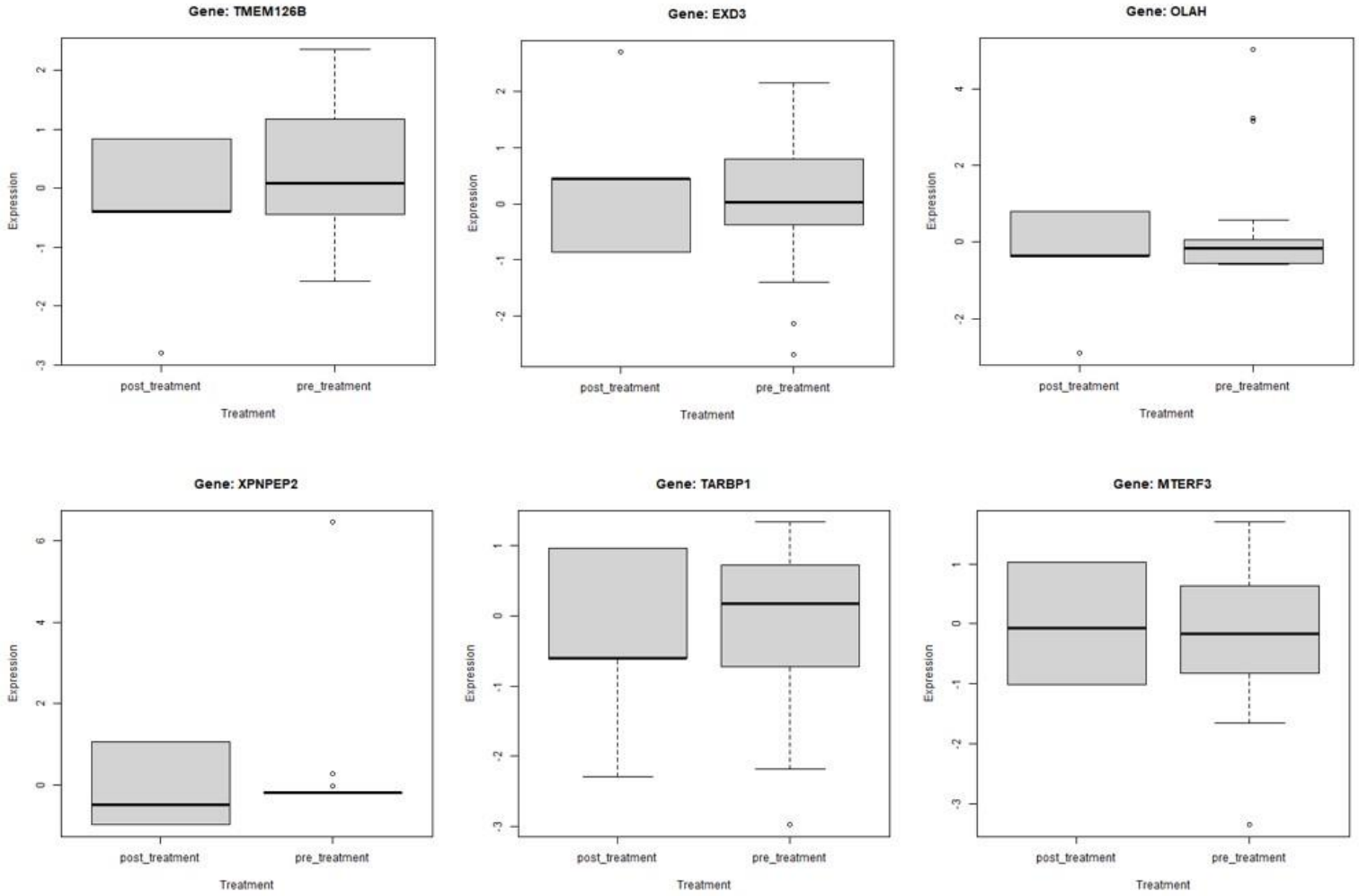

Figure 30: Gene expression change for the first 6 genes in resistant cell lines. A positive mean difference indicates an increase in gene expression, while a negative mean difference indicates a decrease in gene expression. These boxplots illustrate the distribution of expression values in pre-treatment and post-treatment, highlighting the difference between the two groups for genes after 6 hours of treatment.

5. Future work and Conclusion

This research has provided insight into the relationships between gene essentiality and the chemosensitivity of cancer cells. There are, however, several limitations that need to be considered for further investigation. In this research the focus was on using gene expression and CRISPR derived fitness score datasets, but future work could also include additional omics data such as methylation, mutations, and copy number variations.

Another limitation of this study is the size of the datasets that was used to train the machine learning models. Increasing the size of the datasets could have increased the performance and the models. Also, hyperparameter optimization could have been performed to see if the models will overfit, or their performance will improve or stay the same. A future step would be to explore more advanced machine learning models such as deep neural networks: especially since the elastic net and nearest neighbour methods did not perform well for regression analysis. An important final step would be to perform experimental validation in a wet lab to confirm the findings of the kinase enrichment analysis.

To understand gene expression changes pre-treatment and post-treatment, a future step would be to increase the sample size, and to extend the analysis to encompass extended time intervals from 6 hours to 12 hours and 24 hours after treatment. Even though there is a change in the expression of the genes after 6 hours, the changes are not statistically significant. Significant gene expression changes could be observed if the treatment duration is extended.

In conclusion this research has contributed to the advancement of drug response prediction and gene expression analysis, thereby providing insights into personalised breast cancer treatment. By using machine learning to interrogate thoughtfully gathered gene expression data, it should be possible in the future to obtain a deep understanding of drug responses at the molecular level

in individual patients and to reactively use this understanding to maximise the patients' health outcomes.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 2021;71:209–49. https://doi.org/10.3322/caac.21660.

[2] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer 2015;136:E359–86. https://doi.org/10.1002/ijc.29210.

[3] Navin NE, Hicks J. Tracing the tumor lineage. Mol Oncol 2010;4:267–83. https://doi.org/10.1016/j.molonc.2010.04.010.

[4] Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 2013;108:479–85. https://doi.org/10.1038/bjc.2012.581.

[5] McLeod HL. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. Science 2013;339:1563–6. https://doi.org/10.1126/science.1234139.

[6] Parca L, Pepe G, Pietrosanto M, Galvan G, Galli L, Palmeri A, et al. Modeling cancer drug response through drug-specific informative genes. Sci Rep 2019;9:15222. https://doi.org/10.1038/s41598019-50720-0.

[7] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7. https://doi.org/10.1038/nature11003.

[8] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016;166:740–54. https://doi.org/10.1016/j.cell.2016.06.017.

[9] Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570–5. https://doi.org/10.1038/nature11005.

[10] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research 2012;41:D955–61. https://doi.org/10.1093/nar/gks1111.

[11] Howell A, Anderson AS, Clarke RB, Duffy SW, Evans DG, Garcia-Closas M, et al. Risk determination and prevention of breast cancer. Breast Cancer Res 2014;16:446. https://doi.org/10.1186/s13058-014-0446-2.

[12] Ortega MA, Fraile-Martínez O, Asúnsolo Á, Buján J, García-Honduvilla N, Coca S. Signal Transduction Pathways in Breast Cancer: The Important Role of PI3K/Akt/mTOR. Journal of Oncology 2020;2020:1–11. https://doi.org/10.1155/2020/9258396.

[13] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell 2011;144:646–74. https://doi.org/10.1016/j.cell.2011.02.013.

[14] Melchor L, Benítez J. The complex genetic landscape of familial breast cancer. Hum Genet 2013;132:845–63. https://doi.org/10.1007/s00439-013-1299-y.

[15] Diamond TM, Sutphen R, Tabano M, Fiorica J. Inherited susceptibility to breast and ovarian cancer. Current Opinion in Obstetrics and Gynecology 1998;10:3.

[16] International network of cancer genome projects. Nature 2010;464:993–8. https://doi.org/10.1038/nature08987.

[17] Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes Dev 2011;25:534–55. https://doi.org/10.1101/gad.2017311.

[18] Creighton CJ. Making use of cancer genomic databases. Curr Protoc Mol Biol 2018;121:19.14.1-19.14.13. https://doi.org/10.1002/cpmb.49.

[19] Zhang X, Acencio ML, Lemke N. Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review. Front Physiol 2016;7. https://doi.org/10.3389/fphys.2016.00075.

[20] Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. Journal of Internal Medicine 2018;284:603–19. https://doi.org/10.1111/joim.12822.

[21] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol n.d.;9:14. https://doi.org/10.1167/tvst.9.2.14.

[22] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69:S36–40. https://doi.org/10.1016/j.metabol.2017.01.011.

[23] Ahmad MA, Eckert C, Teredesai A, McKelvey G. Interpretable Machine Learning in Healthcare 2018:7.

[24] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94–8. https://doi.org/10.7861/futurehosp.6-2-94.

[25] Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Medical Research Methodology 2019;19:64. https://doi.org/10.1186/s12874-019-0681-4.

[26] Weintraub WS, Fahed AC, Rumsfeld JS. Translational Medicine in the Era of Big Data and Machine Learning. Circ Res 2018;123:1202–4. https://doi.org/10.1161/CIRCRESAHA.118.313944.

[27] Automated machine learning_ Review of the state-of-the-art and opportunities for healthcare | Elsevier Enhanced Reader n.d. https://doi.org/10.1016/j.artmed.2020.101822.

[28] Song Q, Ni J, Wang G. A Fast Clustering-Based Feature Subset Selection Algorithm for HighDimensional Data. IEEE Transactions on Knowledge and Data Engineering 2013;25:1–14. https://doi.org/10.1109/TKDE.2011.181.

[29] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70–9. https://doi.org/10.1016/j.neucom.2017.11.077.

[30] Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics 2017;18:9. https://doi.org/10.1186/s12859-016-1423-9.

[31] Chen R-C, Dewi C, Huang S-W, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data 2020;7:52. https://doi.org/10.1186/s40537-02000327-4.

[32] Zhao Z, Anand R, Wang M. Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform 2019.

[33] Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 2003;27:45–63. https://doi.org/10.1016/s0933-3657(02)00086-6.

[34] De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics 2013;29:2365–8. https://doi.org/10.1093/bioinformatics/btt383.

[35] Khan MU, Choi JP, Shin H, Kim M. Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, p. 5148–51. https://doi.org/10.1109/IEMBS.2008.4650373.

[36] Ramezan CA, Warner TA, Maxwell AE, Price BS. Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. Remote Sensing 2021;13:368. https://doi.org/10.3390/rs13030368.

[37] Clarke PA, te Poele R, Wooster R, Workman P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential22Abbreviations: ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; Cy, Cyanine; DLCL, diffuse large cell lymphoma; dNTPs, deoxyribonucleotides; ESTs, expressed sequence tags; mRNA, messenger RNA; NHL, non-Hodgkin's lymphoma; ORF, open reading frame; RT-PCR, reverse transcription-polymerase chain reaction; and 17AAG, 17-allylamino,17-demethoxygeldanamycin. Biochemical Pharmacology 2001;62:1311–36. https://doi.org/10.1016/S0006-2952(01)00785-7.

[38] Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse largescale cellular perturbation response data. Nucleic Acids Research 2018;46:D558–66. https://doi.org/10.1093/nar/gkx1063.

[39] Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 2016;12:109–16. https://doi.org/10.1038/nchembio.1986.

[40] Aksoy BA, Dančík V, Smith K, Mazerik JN, Ji Z, Gross B, et al. CTD2 Dashboard: a searchable web interface to connect validated results from the Cancer Target Discovery and Development Network. Database 2017;2017:bax054. https://doi.org/10.1093/database/bax054.

[41] Zhang H, Chen Y, Li F. Predicting Anticancer Drug Response With Deep Learning Constrained by Signaling Pathways. Frontiers in Bioinformatics 2021;1.

[42] Musa A, Tripathi S, Kandhavelu M, Dehmer M, Emmert-Streib F. Harnessing the biological complexity of Big Data from LINCS gene expression signatures. PLoS ONE 2018;13:e0201937. https://doi.org/10.1371/journal.pone.0201937.

[43] Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer 2015;15:489. https://doi.org/10.1186/s12885-015-1492-6.

[44] Huang S, Hu P, Lakowski TM. Predicting breast cancer drug response using a multiple-layer cell line drug response network model. BMC Cancer 2021;21:648. https://doi.org/10.1186/s12885-021-08359-6.

[45] Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics 2016;32:2891–5. https://doi.org/10.1093/bioinformatics/btw344.

[46] Cui W, Aouidate A, Wang S, Yu Q, Li Y, Yuan S. Discovering Anti-Cancer Drugs via Computational Methods. Frontiers in Pharmacology 2020;11.

[47] Stetson LC, Pearl T, Chen Y, Barnholtz-Sloan JS. Computational identification of multi-omic correlates of anticancer therapeutic response. BMC Genomics 2014;15:S2. https://doi.org/10.1186/14712164-15-S7-S2.

[48] Sultana J, Khader Jilani A, .. Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers. IJET 2018;7:22. https://doi.org/10.14419/ijet.v7i4.20.22115.

[49] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79:857–62. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y.

[50] Ribeiro MT, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018;32. https://doi.org/10.1609/aaai.v32i1.11491.

[51] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc.; 2017.

[52] ELI5 2023.

[53] Dave D, Naik H, Singhal S, Patel P. Explainable AI meets Healthcare: A Study on Heart Disease Dataset 2020.

[54] Wang X, Gotoh O. A Robust Gene Selection Method for Microarray-based Cancer Classification. Cancer Inform 2010;9:15–30.

[55] Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining 2021;14:13. https://doi.org/10.1186/s13040-021-00244-z.

[56] (12) Understanding the ROC & AUC | LinkedIn n.d. https://www.linkedin.com/pulse/understanding-roc-auc-gautam-k/ (accessed May 19, 2023).

[57] Pouryahya M, Oh JH, Mathews JC, Belkhatir Z, Moosmüller C, Deasy JO, et al. Pan-Cancer Prediction of Cell-Line Drug Sensitivity Using Network-Based Methods. International Journal of Molecular Sciences 2022;23:1074. https://doi.org/10.3390/ijms23031074.

[58] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software 2019;4:1686. https://doi.org/10.21105/joss.01686.

[59] Lenhof K, Eckhart L, Gerstner N, Kehl T, Lenhof H-P. Simultaneous regression and classification for drug sensitivity prediction using an advanced random forest method. Sci Rep 2022;12:13458. https://doi.org/10.1038/s41598-022-17609-x.

[60] Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 2008;28:1–26. https://doi.org/10.18637/jss.v028.i05.

[61] Sunil Arya, David Mount , Samuel E. Kemp, Gregory Jefferis. RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric 2019.

[62] Ding C, Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data n.d.

[63] R Core team. R: A language and Environment for Statistical Computing 2022.

[64] Anaconda. Anaconda Software Distribution 2016.

[65] pandas documentation — pandas 2.1.1 documentation n.d. https://pandas.pydata.org/docs/ (accessed October 10, 2023).

[66] NumPy Documentation n.d. https://numpy.org/doc/ (accessed October 10, 2023).

[67] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

[68] Molnar C. 8.5 Permutation Feature Importance | Interpretable Machine Learning. n.d.

[69] Mucaj V, Shay JES, Simon MC. Effects of hypoxia and HIFs on cancer metabolism. Int J Hematol 2012;95:464–70. https://doi.org/10.1007/s12185-012-1070-5.

[70] Fuhrmann DC, Olesch C, Kurrle N, Schnütgen F, Zukunft S, Fleming I, et al. Chronic Hypoxia Enhances β-Oxidation-Dependent Electron Transport via Electron Transferring Flavoproteins. Cells 2019;8:172. https://doi.org/10.3390/cells8020172.

[71] Li C, Ge M, Yin Y, Luo M, Chen D. Silencing expression of ribosomal protein L26 and L29 by RNA interfering inhibits proliferation of human pancreatic cancer PANC-1 cells. Mol Cell Biochem 2012;370:127–39. https://doi.org/10.1007/s11010-012-1404-x.

[72] Liu H, Yuan M, Mitra R, Zhou X, Long M, Lei W, et al. CTpathway: a CrossTalk-based pathway enrichment analysis method for cancer research. Genome Medicine 2022;14:118. https://doi.org/10.1186/s13073-022-01119-6.

[73] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30.

[74] Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci 2019;28:1947–51. https://doi.org/10.1002/pro.3715.

[75] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomybased analysis of pathways and genomes. Nucleic Acids Res 2022;51:D587–92. https://doi.org/10.1093/nar/gkac963.

[76] Hunter T, Cooper JA. Protein-Tyrosine Kinases. Annual Review of Biochemistry 1985;54:897– 930. https://doi.org/10.1146/annurev.bi.54.070185.004341.

[77] Cicenas J, Zalyte E, Bairoch A, Gaudet P. Kinases and Cancer. Cancers 2018;10:63. https://doi.org/10.3390/cancers10030063.

[78] Chong ZZ, Shang YC, Wang S, Maiese K. A Critical Kinase Cascade in Neurological Disorders: PI 3-K, Akt, and mTOR. Future Neurol 2012;7:733–48.

[79] Tabit CE, Shenouda SM, Holbrook M, Fetterman JL, Kiani S, Frame AA, et al. Protein Kinase Cβ Contributes to Impaired Endothelial Insulin Signaling in Humans With Diabetes Mellitus.Circulation 2013;127:86–95. https://doi.org/10.1161/CIRCULATIONAHA.112.127514.

[80] clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters | OMICS: A Journal of Integrative Biology n.d. https://www.liebertpub.com/doi/10.1089/omi.2011.0118 (accessed August 25, 2023).

[81] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141.

[82] Martin Morgan, Marcel Ramos. BiocManager: Access the Bioconductor Project Package Repository 2023.

[83] Hadley Wickam. ggplot2: Elegant Graphics for Data Analysis 2016.

[84] Chen EY, Xu H, Gordonov S, Lim MP, Perkins MH, Ma'ayan A. Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. Bioinformatics 2012;28:105–11. https://doi.org/10.1093/bioinformatics/btr625.

[85] Nayak AP, Kapur A, Barroilhet L, Patankar MS. Oxidative Phosphorylation: A Target for Novel Therapeutic Strategies Against Ovarian Cancer. Cancers (Basel) 2018;10:337. https://doi.org/10.3390/cancers10090337.

[86] Evans KW, Yuca E, Scott SS, Zhao M, Paez Arango N, Cruz Pico CX, et al. Oxidative Phosphorylation Is a Metabolic Vulnerability in Chemotherapy-Resistant Triple-Negative Breast Cancer. Cancer Res 2021;81:5572–81. https://doi.org/10.1158/0008-5472.CAN-20-3242.

[87] Poetsch AR. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. Comput Struct Biotechnol J 2020;18:207–19. https://doi.org/10.1016/j.csbj.2019.12.013.

[88] Monaco ME. Fatty acid metabolism in breast cancer subtypes. Oncotarget 2017;8:29487–500. https://doi.org/10.18632/oncotarget.15494.

[89] Yoon H, Lee S. Fatty Acid Metabolism in Ovarian Cancer: Therapeutic Implications. International Journal of Molecular Sciences 2022;23:2170. https://doi.org/10.3390/ijms23042170.

[90] Kashyap D, Tuli HS, Sak K, Garg VK, Goel N, Punia S, et al. Role of Reactive Oxygen Species in Cancer Progression. Curr Pharmacol Rep 2019;5:79–86. https://doi.org/10.1007/s40495-019-00171-y.

[91] Malumbres M, Barbacid M. Mammalian cyclin-dependent kinases. Trends in Biochemical Sciences 2005;30:630–41. https://doi.org/10.1016/j.tibs.2005.09.005.

[92] Malumbres M. Cyclin-dependent kinases. Genome Biology 2014;15:122. https://doi.org/10.1186/gb4184.

[93] Bae JS, Park S-H, Jamiyandorj U, Kim KM, Noh SJ, Kim JR, et al. CK2α/CSNK2A1 Phosphorylates SIRT6 and Is Involved in the Progression of Breast Carcinoma and Predicts Shorter Survival of Diagnosed Patients. The American Journal of Pathology 2016;186:3297315. https://doi.org/10.1016/j.ajpath.2016.08.007.

[94] Qi M, Elion EA. MAP kinase pathways. Journal of Cell Science 2005;118:3569–72. https://doi.org/10.1242/jcs.02470.

[95] Morrison DK. MAP Kinase Pathways. Cold Spring Harb Perspect Biol 2012;4:a011254. https://doi.org/10.1101/cshperspect.a011254.

[96] Fu L, Zhang L. Physiological functions of CKIP-1: From molecular mechanisms to therapy implications. Ageing Research Reviews 2019;53:100908. https://doi.org/10.1016/j.arr.2019.05.002.

[97] Wu D, Yin Y-Q, Li Y, Zhang L, Jiang Y-H, Wang Z. CK2α causes stemness and chemotherapy resistance in liver cancer through the Hedgehog signaling pathway. Hepatobiliary & Pancreatic Diseases International 2023;22:383–91. https://doi.org/10.1016/j.hbpd.2021.09.003.

[98] Kuwano Y, Nishida K, Akaike Y, Kurokawa K, Nishikawa T, Masuda K, et al. HomeodomainInteracting Protein Kinase-2: A Critical Regulator of the DNA Damage Response and the Epigenome. Int J Mol Sci 2016;17:1638. https://doi.org/10.3390/ijms17101638.

[99] Domoto T, Uehara M, Bolidong D, Minamoto T. Glycogen Synthase Kinase 3β in Cancer Biology and Treatment. Cells 2020;9:1388. https://doi.org/10.3390/cells9061388.

[100] Roskoski R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update. Pharmacological Research 2023;187:106552. https://doi.org/10.1016/j.phrs.2022.106552.

[101] Liu A-G, Zhong J-C, Chen G, He R-Q, He Y-Q, Ma J, et al. Upregulated expression of SAC3D1 is associated with progression in gastric cancer. Int J Oncol 2020;57:122–38. https://doi.org/10.3892/ijo.2020.5048.

[102] Han M-E, Kim J-Y, Kim GH, Park SY, Kim YH, Oh S-O. SAC3D1: a novel prognostic marker in hepatocellular carcinoma. Sci Rep 2018;8:15608. https://doi.org/10.1038/s41598-018-34129-9.

[103] Lin L, Wen J, Lin B, Xia E, Zheng C, Ye L, et al. Stress-induced phosphoprotein 1 facilitates breast cancer cell progression and indicates poor prognosis for breast cancer patients. Human Cell 2021;34:901–17. https://doi.org/10.1007/s13577-021-00507-1.

[104] Ye J, Wang J, Tan L, Yang S, Xu L, Wu X, et al. Expression of protein TARBP1 in human hepatocellular carcinoma and its prognostic significance. Int J Clin Exp Pathol 2015;8:9089–96.

[105] Ye J, Wang J, Zhang N, Liu Y, Tan L, Xu L. Expression of TARBP1 protein in human non-smallcell lung cancer and its prognostic significance. Oncology Letters 2018;15:7182–90. https://doi.org/10.3892/ol.2018.8202.

[106] Lorenzen JA, Bonacci BB, Palmer RE, Wells C, Zhang J, Haber DA, et al. Rbm19 is a nucleolar protein expressed in crypt/progenitor cells of the intestinal epithelium. Gene Expression Patterns 2005;6:45–56. https://doi.org/10.1016/j.modgep.2005.05.001.

[107] Wang L, Zhang Z, Li Y, Wan Y, Xing B. Integrated bioinformatic analysis of RNA binding proteins in hepatocellular carcinoma. Aging 2020;13:2480–505. https://doi.org/10.18632/aging.202281.

[108] [108] Jin Z, Zhang G, Liu Y, He Y, Yang C, Du Y, et al. The suppressive role of HYAL1 and HYAL2 in the metastasis of colorectal cancer. Journal of Gastroenterology and Hepatology 2019;34:1766–76. https://doi.org/10.1111/jgh.14660.

[109] Player A, Abraham N, Burrell K, Bengone IO, Harris A, Nunez L, et al. Identification of candidate genes associated with triple negative breast cancer. Genes Cancer 2017;8:659–72. https://doi.org/10.18632/genesandcancer.147.

[110] Huang J, Liu W, Zhang D, Lin B, Li B. TMEM158 expression is negatively regulated by AR signaling and associated with favorite survival outcomes in prostate cancers. Frontiers in Oncology 2022;12.

[111] Yucebas M, Susluer SY, Caglar HO, Balci T, Sigva ZOD, Akalin T, et al. Expression profiling of RE1-silencing transcription factor (REST), REST corepressor 1 (RCOR1), and Synapsin 1 (SYN1) genes in human gliomas n.d.

[112] Xunyi Y, Zhentao Y, Dandan J, Funian L. Clinicopathological significance of PTPN12 expression in human breast cancer. Braz J Med Biol Res 2012;45:1334–40. https://doi.org/10.1590/S0100879X2012007500163.

[113] Hwang M, Han M-H, Park H-H, Choi H, Lee K-Y, Lee YJ, et al. LGR5 and Downstream Intracellular Signaling Proteins Play Critical Roles in the Cell Proliferation of Neuroblastoma, Meningioma and Pituitary Adenoma. Exp Neurobiol 2019;28:628–41. https://doi.org/10.5607/en.2019.28.5.628.

[114] Zhang F, Xiang S, Cao Y, Li M, Ma Q, Liang H, et al. EIF3D promotes gallbladder cancer development by stabilizing GRK2 kinase and activating PI3K-AKT signaling pathway. Cell Death Dis 2017;8:e2868. https://doi.org/10.1038/cddis.2017.263.

[115] Li C, Lu K, Yang C, Du W, Liang Z. EIF3D promotes resistance to 5-fluorouracil in colorectal cancer through upregulating RUVBL1. J Clin Lab Anal 2023;37:e24825. https://doi.org/10.1002/jcla.24825.

[116] Fan Y, Guo Y. Knockdown of eIF3D inhibits breast cancer cell proliferation and invasion through suppressing the Wnt/β-catenin signaling pathway. Int J Clin Exp Pathol 2015;8:10420–7.

[117] Lippert T, Ruoff H-J, Volm M. Intrinsic and Acquired Drug Resistance in Malignant Tumors. Arzneimittelforschung 2011;58:261–4. https://doi.org/10.1055/s-0031-1296504.

Appendix A

Supplementary Table 1: Genes ranked according to the mean decrease accuracy using the permutation feature method.

Genes	Scores	Genes	Scores
ZNF608	0.659196	MRPL28	0.645773
EIF3D	0.658207	GFI1B	0.645581
NRDC	0.657712	ADAM28	0.645444
TGM5	0.656741	ALYREF	0.645187
TROAP	0.655779	SAV1	0.645068
CIITA	0.652636	PAXBP1	0.644985
HYAL1	0.652453	MRM2	0.644711
FIG4	0.652059	ARHGEF2	0.644582
SRFBP1	0.651628	ARAF	0.644491
MRPS30	0.651601	PSMA2	0.644097
ZNF607	0.651124	SAE1	0.643941
ANXA5	0.651005	POLR2F	0.643574
MRPL19	0.650886	TSPAN3	0.643153
PDLIM2	0.650694	EDA2R	0.64308
KBTBD3	0.65019	ODF2L	0.642878
POC5	0.650126	NLRCA	0.642521
MRPL38	0.649521	RPS11	0.642365
MRPS34	0.649338	HSD17B8	0.641962
SMARCAD	0.64811	SPRR1B	0.641018
RETSAT	0.648046	TCEAL7	0.641018
MRPL15	0.647743	RACK1	0.640844
ELAVL1	0.647679	BMT2	0.64078
GADD45G	0.647184	ADAM23	0.640716
MEDAG	0.646012	TPRG1	0.639946
KRT82	0.645828	PDE6H	0.6399

Genes	Scores	Genes	Scores
PARS2	0.639662	EML3	0.635118
VARS2	0.639653	PTPRR	0.634403
RPL30	0.639497	EIF4EBP2	0.634064
KCNC2	0.639488	ZSCAN20	0.633853
ENPP2	0.639314	ACAP2	0.633642
EXTL2	0.639241	MATN2	0.633496
DNAJC5G	0.639057	NABP2	0.633377
U2AF1L4	0.638874	L3MBTL2	0.633349
PCDHB13	0.638801	IRF8	0.633276
CCDC120	0.638663	COX7A2	0.632919
ZNF672	0.638553	HLF	0.632323
NDP	0.638462	MAGEE2	0.632149
ASNSD1	0.638352	RPS9	0.632048
OCLN	0.638242	MOSMO	0.631755
ZSCAN10	0.637527	PER3	0.631673
NAPEPLD	0.637509	MUS81	0.631654
SNX31	0.636904	KLF8	0.631627
XAGE5	0.636886	ABCF3	0.631608
TMEM69	0.636877	YBEY	0.630885
TMEM241	0.636739	MPEG1	0.63071
MRPL37	0.636629	VIP	0.63061
ZZEF1	0.636391	SMARCA5	0.630509
PRIM1	0.636171	CCDC153	0.630234
DAPL1	0.635988	XAGE3	0.629638
TSPOAP1	0.635915	ATAD5	0.629226
RBM25	0.635557	PMPCA	0.62907
ZNF7	0.635319	MYEF2	0.62897
RBP4	0.635273	DCLRE1B	0.628768
SYPL1	0.635145	OSBPL10	0.628695
KIAA0319L	0.635136	COPA	0.628631

Genes	Scores	Genes	Scores
CNNM1	0.628576	NPW	0.623994
AUH	0.628521	MIF	0.623637
COX6C	0.628502	NDUFB7	0.623591
ZSWIM4	0.6282	ADGRV1	0.623124
MRPS28	0.627623	PDC	0.623069
MITD1	0.627366	ADRA1A	0.622977
MRPL53	0.627137	SRRM4	0.62284
SEZ6L2	0.626908	IER3IP1	0.622822
KDM4C	0.626816	GLOD5	0.622712
ACSL4	0.626697	FADS6	0.622684
CDCA7L	0.626624	CTTNBP2NL	0.622373
CHD1L	0.626477	ZNF181	0.622162
ATP23	0.626294	SHISA4	0.621979
IGSF1	0.626248	TEX37	0.62196
OSER1	0.626028	TTI1	0.62175
COPB1	0.625974	PCDHGA6	0.621649
CPT1A	0.625882	EFCAB9	0.621631
KRT80	0.625552	SLC8A3	0.621557
MYOZ2	0.625433	DCAF8L1	0.621392
MRPL55	0.625131	HDX	0.621273
ZNF28	0.624938	TET3	0.621163
ABCA12	0.624526	CWF19L2	0.621044
INTS7	0.624526	MRPL54	0.621017
FCGRT	0.624453	ANKRD13B	0.620888
RPF2	0.624361	GADD45G	0.620751
ATL2	0.624242	KRT85	0.620668
WASF1	0.624223	P3H1	0.620632
EXTL3	0.624022	CAP2	0.6205

Genes	Scores	Genes	Scores
AQP2	0.620494	UQCRQ	0.617572
COX10	0.620476	AUNIP	0.617517
NRXN1	0.620375	AURKAIP1	0.617425
STK39	0.62021	CNTNAP5	0.617379
SHOC2	0.620155	ZDHHC16	0.617379
ANKRD12	0.619835	HKDC1	0.61693
ATP9A	0.619725	NDUFA3	0.616903
LCTL	0.619688	ANKRD31	0.616903
CAPZA3	0.619624	ZNF124	0.616811
C16orf87	0.619587	MRPL52	0.616582
ISL2	0.619514	TLR9	0.616545
FGF8	0.619386	BTK	0.616445
SLCO2A1	0.619303	CTNNBIP1	0.6166353
LRP12	0.61912	HCAR3	0.615913
LMOD3	0.619019	MRGPRG	0.61584
PARD6B	0.618818	PRR14L	0.615776
SYNGR1	0.618754	METTL24	0.615776
CNNM2	0.618671	AOC2	0.615602
PSMB1	0.618625	CCT8	0.615528
EIF4EBP1	0.618341	NDUFB1	0.615244
TP63	0.618286	ZNF33B	0.615244
NCBP1	0.618002	RBM15	0.615171
HOXB3	0.617984	SH3BGRL3	0.614832
GPM6B	0.617975	NSG2	0.614795
SLC22A23	0.617929	CLUL1	0.614786

Appendix B

Supplementary Table 2: The 144 genes used for analysing gene expression changes found in both GDSC_CCLE and GDSC_LINCS.

TMEM126B	PDLIM2	NDP	COX6C
EXD3	MRPS34	ASNSD1	MRPS28
OLAH	RETSAT	RBP4	KDM4C
XPNPEP2	MRPL15	SYPL1	ACSL4
TARBP1	GADD45GIP1	KIAA0319L	CHD1L
MTERF3	MRPL28	EML3	IGSF1
SAC3D1	ADAM28	EIF4EBP2	OSER1
ABCA4	SAV1	MATN2	COPB1
STIP1	ARHGEF2	NABP2	CPT1A
SYN1	ARAF	IRF8	MYOZ2
PTPN12	POLR2F	COX7A2	ABCA12
HNRNPH3	TSPAN3	HLF	FCGRT
RBM19	EDA2R	PER3	WASF1
EIF3D	HSD17B8	MUS81	EXTL3
NRDC	RACK1	KLF8	MIF
TGM5	ADAM23	ABCF3	NDUFB7
HYAL1	KCNC2	VIP	ADRA1A
FIG4	ENPP2	MYEF2	IER3IP1
MRPS34	EXTL2	COPA	GADD45G
ANXA5	PCDHB13	CNNM1	KRT85

MRPL19	ZNF672	AUH	P3H1
CAP2	PRDX3	SH3BGRL3	PTBP3
AQP2	EIF4EBP1	CLUL1	DMBT1
COZ10	TP63	ANKH	TRAPPC2L
STK39	GPM6B	TNFSF14	NDUFA5
SHOC2	GPM6B	SDS	FPGS
ANKRD12	UQCRQ	PTDSS1	ETNK1
ATP9A	AUNIP	NELFE	TLE6
FGF8	AURKAIP1	SLC25A1	GPR137B
SLCO2A1	HKDC1	TFB2M	RPL8
LRP12	NDUFA3	VANGL1	ICAM5
PARD6B	MRPL52	MPPED2	CLTCL1
SYNGR1	CTNNBIP1	TNIK	HGS
CNNM2	NDUFB1	MORC2	CTF1
PSMB1	ZNF33B	GTPBP8	COX4I1
GPR88	POU6F1	MYH15	PTPRR